R package practicalSigni Supplements Statistical Significance

Abstract

The literature on p-hacking primarily highlights its prevalence and the need to pay attention to the practical significance of each regressor. A new R package, 'practicalSigni,' reports 13 indexes for p regressors called m1 to m13. The following methods are missing from the literature. (m6) Generalized partial correlation coefficients (GPCC). (m7) extending Psychologists' "Effect Sizes" to two or more regressors. (m8, m9) Two kernel regression partial derivatives from np and NNS packages. (m10) The NNS.boost function of NNS. (m11) re-imagines the regression as a cooperative game (Shapley Value) of forecasting. Two random forest feature importance measures (m12, m13) use out-of-bin (OOB) calculations. Since regression is important for all quantitative sciences, my package offers simple few line commands that summarize the ranks by newer nonparametric methods (m6 to m13).

1 P-hacking and Practical Significance

A p-value less than 0.05 often identifies statistically significant scientific explanations of phenomena. P-hacking occurs when researchers game the estimation to achieve the 0.05 threshold and sell practically insignificant explanations as significant. The p-hacking problem is recently discussed in a special issue of the *American Statistician*, having 43 papers. Vinod (2022) reports a p-hacking example in econometrics by Khan et al. (2020).

Let us consider a usual ordinary least squares (OLS) linear regression,

$$y_t = \alpha_0 + \Sigma_{i=1}^{p} \alpha_i x_{it} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \tag{1}$$

where $t = 1, \ldots T$. Assume that research seeks to reveal the relative importance of the p regressors in explaining y. The first measure **(m1)** of relative importance is a set of

absolute values ($|\hat{\alpha}_i|, i = 0, 1, \ldots p$). Since ($|\hat{\alpha}_i|$) are sensitive to units of measurement, their ranking need not indicate relative importance.

The linear functional form in (1) is too restrictive and subject to specification errors. Hence we use kernel regression based on the kernel density algorithm here. The empirical cumulative distribution function (ecdf) provides $F(x)$. The density $f(x) = (dF/dx)$ is the derivative of the ecdf. One definition of a numerical derivative in calculus is the limit of the central difference. Accordingly, the density is:

$$f(x) = \lim_{h \to 0} \left(\frac{1}{h}\right) \left[F\left(x + \frac{h}{2}\right) - F\left(x - \frac{h}{2}\right)\right]. \tag{2}$$

In the 1950s Rosenblatt suggested replacing the central difference in (**??**) with a kernel weight function. His kernel weights are positive and must integrate (add up) to unity. A normal kernel yields convenient weights w_t $K(w_t) \sim N(0, \sigma^2)$. The standard algorithms for density estimation use $w_t = \frac{(x_t - x)}{h}$, where x is the point at which the density is evaluated and x_t are nearby observed data points defined on an expanded grid of points. See Vinod (2021) for grid expansion details used in R.

Let X denotes a matrix of all regressors in (1). Let $f(y, X)$ denote the joint density of (y, X) variables and $f(X)$ denote the joint density of regressors, where we use a generic notation $f(.)$ for densities. The left-hand side of (1) is a conditional mean $E(y|X) = f(y, X)/f(X)$. Kernel regression applies kernel algorithm to the two densities involved in $E(y|X)$. The resulting regression equation (in simplified notation) is

$$y_t = R(X)_t + \epsilon_t, \quad R(x) = \frac{\sum_{t=1}^{T} y_t K(w_t)}{\sum_{t=1}^{T} K(w_t)}. \tag{3}$$

Since $R(X)$ in (3) does not contain any regression parameters α_i, it is called nonparametric regression. The conditional expectation function $E(y|X) = R(X)$ based on the kernel regression algorithm traces a freehand curve very close to the data values of y. The main appeal of kernel regression is its superior fit compared to the OLS.

Our second commonly reported measure **(m2)** is the value of t-statistic which is not

sensitive to units. Its absolute value ($|t_i|$) is relevant when the sign of the coefficient does not matter. For example, if the researcher wishes to show that xi increases y, and the coefficient is negative with $t_i < 0$, then the unexpected sign indicates that what the researcher wishes to show is rejected by the data. Reporting all p-values is equivalent to t-stats, except that p-values hide "wrong" signs of coefficient estimates.

Our third measure **(m3)** overcomes the units problem of ($|\hat{\alpha}_i|$) by re-scaling all variables to have zero means and unit standard deviations. Re-scaled coefficients used to be called 'beta coefficients' or ($\hat{\beta}_i$). If the sign of the coefficient does not matter, (m3) becomes the absolute value ($|\hat{\beta}_i|$) as a measure of the importance of xi.

The commonly used fourth method **(m4)** for assessing regressor importance is the size of Pearson correlation coefficient, $r_{y,xi} = \sqrt{R^2}$, where R^2 is the coefficient of determination in regressing y on xi. The sign of the square root $r_{y,xi}$ is that of the covariance $cov(y, xi)$. If the sign matters in the problem at hand, a "wrong" sign of $r_{y,xi}$ casts doubt on the researcher's claim. The first limitation of $r_{y,xi}$ is that it ignores nonlinear dependence. When one uses nonlinear nonparametric kernel regression (3), the magnitude of the R^2 depends on the regression direction (regress y on xi or regress of xi on y).

Vinod (2014) defines the square roots of the two R^2 values as elements of the asymmetric matrix R^* of generalized correlation coefficients. Kernel regression of y on xi from the square root of the R^2 denoted by $r^*_{y|xi} \neq r^*_{xi|y}$ overcomes the linearity assumption and is called **(m5)** here. The R package **generalCorr** offers a convenient function `depMeas(y,x)` for computing (m5). The second limitation of Pearson correlation ($r_{y,xi}$) is that the coefficient is between only two variables (y, xi), ignoring all others. The method (m5) overcomes linearity but continues to measure only two variables at a time.

Now we turn to the multivariate extension of ($r_{xi,xj}$). The textbook formula for partial correlation coefficient between (X_1, X_2) after removing the effect of (X_3) is:

$$r_{12;3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r^2_{13})}\sqrt{(1 - r^2_{23})}}. \tag{4}$$

The numerator $(r_{12} - r_{13}r_{23})$ has the *linear* correlation coefficient between X_1 and X_2 after subtracting the *linear* effect of X_3 on both X_1 and X_2 defined in terms of regression residuals. The denominator does a normalization to obtain a scale-free correlation coefficient. Vinod (2021) generalizes a multivariate version of (4) by replacing various correlations with the square roots of R^2 from appropriate nonlinear kernel regressions.

In the context of our notation, the traditional partial correlation coefficient $r_{ij|o}$ is between two variables, xi, and xj, after removing the *linear effect* of all other variables denoted by xo. However, it is necessary to generalize $r_{ij|o}$ further by avoiding the assumed linearity leading to our method **(m6)**. Vinod (2022) explains a generalized partial correlation coefficient (GPCC) $(y, xi|xo)$ of the indicated variables. It removes the *nonlinear effect* of all other regressors. A single-line R code for computing GPCCs is available on the web.

Section 2 details our seventh method **(m7)**, extending psychologists' effect sizes (ES). It is necessary to generalize ES to accommodate the usual regression problem with $p \geq 2$ regressors, where regressors in a kernel regression need not be dummy (0, 1) variables.

Our eighth and ninth methods of computing the practical significance use partial derivatives $(\partial y/\partial xi)$ from fitted nonparametric kernel regressions using standardized data. There are many differences between algorithms for the numerical computation of these partial derivatives in two R packages. Our **(m8)** method uses the **np** package, and **(m9)** uses the **NNS** package. While np uses a chosen bandwidth for each regressor, NNS uses a dynamic bandwidth for each regressor via iterated conditional means.

Now we introduce tools from machine learning literature for overcoming the p-hacking problem. These four methods **(m10 to m13)** are computer intensive. They involve repeated iterative computations using randomization and specialized cross-validation. The underlying algorithms are explained in the literature cited by authors of various R packages used here. An intuitive explanation of all machine learning algorithms used in this paper is that they use the brute power of computers to optimize an objective function (e.g., good fit and or good out-of-sample forecast) under constraints. They all use trial and error with

randomization and cross-validation.

Our tenth method measures the "importance" of a "feature" or regressor. Thus, **(m10)** refers to the results based on the R package **NNS** function `NNS.boost(.)`. It is an ensemble method using nonlinear regression based on partial moment quadrant means (NNS.reg). It stores for each random iteration a regressor list and the corresponding sum of squared prediction errors (SSPE). NNS.boost measures the importance of each regressor by the number of times (frequency) it appears among the set achieving the lowest 20% of SSPE values.

Recall that the p-hacking problem refers to the p-values of regression coefficients in (1) and (3). The method **(m11)** re-imagines the defining regression as a game of predicting y from a set of p regressors (features) along the columns of X. In the literature on cooperative game theory, Shapley Value fairly divides the payoff among p players. See Branzei et al. (2010). We obtain a new measure of practical significance by applying a machine learning R package **ShapleyValue**. Our (m11) reports the standardized relative importance of each regressor while avoiding collinearity.

The R package **randomForest** implements the famous random forest algorithm by Breiman (2002). It also reports two measures of the "importance" of a regressor based on out-of-bin (OOB) calculations in two columns. The first column reports the percent decrease in mean squared error (MSE) upon omitting each regressor. The second column reports the mean decrease in accuracy upon omitting each regressor. The importance values in two columns are called **(m12)** and **(m13)** here.

A review of all thirteen methods suggests that each represents a unique viewpoint for the importance rankings of p regressors. In every real problem, the researcher needs to select a subset of relevant viewpoints from our list. The algorithms for methods (m1 to m4) are well known, but methods (m5 to m13) are not mentioned in the p-hacking literature, and some are new. More importantly, my R package practicalSigni allows a side-by-side comparison.

The plan for the remaining paper is as follows. We describe a new generalization of effect size in Section 2. Section 3 compares (m1 to m13) rankings side-by-side on 'mtcars'

data.

2 Generalization of 'Effect Size' from Psychology

The "effect size" (ES) measurement from the Biometrics and Psychology literature can be conveniently motivated by an example in Rosenthal and Rubin (1982). They describe a medical treatment that reduced the death rate from 66% to 44%. The p-value on the death reduction exceeded 0.05, implying that the reduction was statistically insignificant. Yet the treatment saved a great many lives. See Steiger (2004), among others, for additional examples. The currently available ES measurement solves the problem with p-values subject to three limitations listed later after we establish the necessary notation. We shall see that our method (m7) generalizes the computation of ES to the general regression problem in (1).

We denote the original effect size quantifying the effect of (xi) on the outcome variable (y) as $ES_{psy}(xi)$, where the subscript (psy) indicates the psychological origin of the concept. Our extension $ES_{psyx}(xi)$ considers a set of $p \geq 2$ treatment variables (xi), $(i = 1, 2, \ldots, p)$, where the additional subscript 'x' of ES refers to extensions.

We also assume a set of control variables (xc), such as age, sex, location, etc., that influence the outcome but are outside the focus of the main research problem. In controlled experiments (xc) are placebo controls. Social scientists often employ the experimental terms 'treatments' and 'controls' though their (xi) and (xc) data are passively observed.

Following the Psychology and Biometrics literature, let us initially assume that a treatment variable (xi) is binary. That is, $xi = 1$ if the subject is treated and $xi = 0$ when the treatment is absent. If T is the sample size, let $Ti < T$ denote the subset of xi representing subjects who are treated, or those with $xi = 1$. Denote the mean treatment effect by Mxi. This is the average outcome for the subset of y items (subjects) who were treated by xi. Denote the corresponding variance by Vxi.

We assume a jointly created data generating process (DGP) having y, all xi, and control

variables xc. When passively observed DGP refers to only one set of subjects, the number of treated subjects Ti is often the same as that of control subjects Tc. Hence we generally have $Ti = Tc$, but not always. Let Mxc denote the average outcome y for the Tc control subjects. Denote the corresponding variance by Vxc.

The "effect size" in psychology is a one-sided t-statistic on the difference between two means.

$$ES_{psy}(xi) = \frac{Mxi - Mxc}{SE_i}, \tag{5}$$

where the denominator is a pooled standard error,

$$SE_i = \sqrt{(\frac{Vxi}{Ti-1} + \frac{Vxc}{Tc-1})}. \tag{6}$$

This paper extends effect size $ES_{psy}(xi)$ from psychology to more general situations having three extensions leading to our $ES_{psyx}(xi)$.

(i) $ES_{psyx}(xi)$ allows xi with $i = 1, 2, \ldots, p, (p \geq 2)$, based on regressing y on xi, and suitably defines the mean treatment effect for each xi.

(ii) $ES_{psyx}(xi)$ extends the regression to be a nonlinear kernel regression.

(iii) Our regressors xi need not be categorical variables.

Remark 1: Zero Variance Adjustment. The definition of the t-statistic of (5) or $ES_{psy}(xi)$ uses a standard textbook result on the difference between two sample means. One assumes that both random samples have sizes satisfying $(Ti \geq 30, Tc \geq 30)$, and are independent of each other. The population variances of the two samples are not assumed to be equal. The statistic (5) is not well-defined unless $SE_i \neq 0$. The central limit theorem justifies the claim that $ES_{psy}(xi)$ is an approximate scale-free Student's t statistic, unless it is degenerate.

If the xi regressor variables are converted to binary (0, 1) variables by splitting the data

at the median, it can make the variances zero, implying degenerate cases. One modifies ES ratios with zero denominators by simply setting $SE_i = 1$. The degenerate ES equals $Mxi - Mxc$, or the difference between two means, subject to measurement units, not unit-free t-statistics.

Remark 2: Sign Adjustment. If the Pearson correlation, $r(y, xi) > 0$, is positive and if y refers to something desirable (e.g., profits or longevity), one wants the treatment effect means (Mxi) to exceed the (placebo) control mean (Mxc). Then, the positive effect size $ES_{psyx}(xi)$ is desirable, and negative $ES_{psyx}(xi)$ suggests the treatment is a failure and may be harmful. The opposite is true for the negative correlation $r(y, xi) < 0$. If, on the other hand, y refers to an undesirable outcome (e.g., loss or death), any successful treatment should have $ES_{psyx}(xi) < 0$. In all cases, the sign of the effect size matters. Thus the t-statistic $ES_{psyx}(xi)$ is one-sided; their ranking depends on the problem at hand. It is clear that ordering xi by effect size measured as the absolute values $|ES_{psyx}(xi)|$ is generally inappropriate. By contrast, in OLS regressions, the t-stats of regression coefficients are often two-sided, and their ordering by absolute values can be appropriate.

Now we extend $ES_{psyx}(xi)$ to be a nonlinear kernel regression using the following steps.

1. It is well known that the OLS regression framework can handle the analysis of variance methods by letting the regressors be categorical variables. If regressors are continuous, we must convert such xi regressors into binary (0, 1) dummy variables called Bxi. Our algorithm splits the data at the median.

2. Count the number Ti of Bxi values that equal unity.

3. Kernel regressing y on Bxi identifies the Ti (fitted outcome values of y) associated with $Bxi = 1$.

4. Compute their mean as Mxi and variance as Vxi for treatments.

5. Kernel regress y on all xc control regressors ($xj, j \neq i$) and find fitted values of outcomes associated with observations having $Bxi = 1$ to make sure that their count

is Tc.

6. Find the mean and variance of these fitted values, denoted as Mxc and Vxc, for controls. If genuine control variable(s) are absent in the problem at hand, we assume that the control variable is a column of T ones, $xc = \iota$. Recall the intercept of the OLS regression model. Since each fitted value $\hat{y} = \bar{y}$, the mean of y, we have Mxc= ($\bar{y}$). Since the variance is degenerate, $Vxc = 0$, we must ignore the denominator in (5), making it a difference between two means but not a t-statistic.

7. Insert means and variances in (5) to compute the t-statistic $ES_{psy}(xi)$ for the difference between the two means. Of course, in degenerate cases (zero variance), we have a simple difference between two means in the numerator.

These steps are implemented in R code available on the web. When the t-stat defining $ES_{psyx}(xi)$ in (5) is degenerate due to a zero denominator, one must use only the numerator, the difference between two means ($Mxi - Mxc$). The $ES_{psyx}(xi)$ with a degenerate denominator is no longer a t-statistic.

Effect size measurement based on relative difference while ignoring relevant magnitudes (actual temperatures) is criticized by Pogrow (2019) with a humorous example of choosing a Greenland vacation by comparing it to Antarctica. He suggests the importance of including relevant measurements, not just relative values. Some researchers suggest confidence intervals around estimated effect sizes. A bootstrap confidence interval around our (m7) algorithm for $ES_{psyx}(xi)$ is easy to compute.

The following section illustrates a real-world application based on fuel economy data for 32 cars called 'mtcars' always available in R.

3 Fuel Economy Regressor Orderings for m1 to m13

This section illustrates the ordering of p=3 regressors by their "practical significance" based on thirteen viewpoints. The 'mtcars' data in R is from the 1974 US magazine called *Motor*

Trend. It has 32 observations for 11 variables. We choose miles per gallon (mpg) as the dependent variable y and the regressors are: the number of cylinders (cyl), horsepower (hp) and weight (wt). We let a rank value 1 represent the most important regressor, while rank value p is the least important among p regressors. The index value signs are adjusted so that they are consistent with our ranking convention.

Table 1 reports the importance index values of three regressors based on linear and or bivariate methods (m1) to (m5). The signs have been adjusted so that a larger importance index suggest greater practical importance in explaining the dependent variable y (mpg). Table 2 reports the corresponding rank ordering of regressors. The column for m2 reports traditional t-stat ranking where wt has the largest t-stat implying wt is statistically most significant regressor. The average rank in the last column also suggests wt as the most important, cyl as second most important, and hp as the least important (consistent with t-stat values)

Table 3 reports the importance index values of three regressors based on more modern nonlinear, nonparametric and or multivariate (comprehensive) methods (m6) to (m13). As before, the signs have been adjusted so that a larger importance index suggest greater practical importance in explaining the dependent variable y (mpg). Table 4 reports the corresponding rank ordering of regressors. The average rank in the last column also suggests wt as the most important, cyl as second most important, and hp as the least important (similar to the t-stat magnitudes).

Table 1: Importance indexes of (cyl, hp, wt) for methods (m1 = OLS), to (m5 =generalized correlation coefficient) after sign adjustment (larger the better)

	$\hat{\alpha}_i$	t-stat	$\hat{\beta}_i$	$r_{y,xi}$	depMeas(y,xi)
	m1	m2	m3	m4	m5
cyl	0.942	1.709	0.279	0.852	0.943
hp	0.018	1.519	0.205	0.776	0.938
wt	3.167	4.276	0.514	0.868	0.917

This paper argues that researchers can avoid inadvertent p-hacking if they supplement p-values by evaluating practical significance. We focus particular attention on readily

Table 2: Ranking by linear and or bivariate criteria m1 to m5 and average rank for the five methods in the last column

	m1	m2	m3	m4	m5	avrank
cyl	2	2	2	2	1	1.80
hp	3	3	3	3	2	2.80
wt	1	1	1	1	3	1.40

Table 3: Sign adjusted importance values of regressors by multivariate nonlinear methods (larger number means greater practical importance in explaining y)

	GPCC	ES_{psyx}	np	NNS	boost	Shapley	forest1	forest2
	m6	m7	m8	m9	m10	m11	m12	m13
cyl	0.0019	4.8739	5.1359	0.4291	0.3333	0.3362	16.75	285.1
hp	0.3886	4.5630	5.8848	0.2562	0.3333	0.2608	16.09	331.3
wt	0.4812	4.3038	5.7036	0.2633	0.3333	0.4030	15.43	368.9

measured newer methods (m6 to m13) that are more comprehensive than OLS in the sense that they allow for multivariate nonlinear nonparametric relations.

Each method (m1) to (m13) highlights different aspects and considering several methods plus graphics will mitigate the p-hacking problem.

The p-hackers focus exclusively on the ordering of regressors by their statistical significance (or our method m2 using the t-stats). Other comprehensive viewpoints (m6 to m13) reveal alternative measures of the practical importance of a regressor worthy of consideration.

4 Final Remarks

Recent literature focuses on the need to avoid the p-hacking problem, Wasserstein et al. (2019), arising from excess reliance on regression coefficient p-values (t-stat) while ignoring "practical significance." Since linearity and normality are strong assumptions, we argue for nonlinear nonparametric kernel regressions to supplement the OLS common in most p-hacking examples. While many papers demonstrate the existence of the p-hacking problem, very few provide algorithms (software) to solve the problem. Vinod (2022) proposes two methods (m6) using generalized partial correlation coefficients (GPCC) and a partial

Table 4: Ranking by nonlinear and or multivariate criteria m6 to m13 and average rank for the eight methods in the last column

	m6	m7	m8	m9	m10	m11	m12	m13	avrank
cyl	3	1	3	1	2	2	1	3	2
hp	2	2	1	3	2	3	2	2	2.10
wt	1	3	2	2	2	1	3	1	1.90

derivative (m8) using the 'np' package to solve the p-hacking problem.

Some Psychometrics and Biometrics researchers have long ago proposed an intuitive method of measuring practical significance. They consider "effect sizes" $ES_{psy}(xi)$ instead of t-stats. This paper describes a new method (m7 and software algorithm steps) for what we denote as $ES_{psyx}(xi)$. The additional subscript 'x' suggests an extension by admitting $p \geq 2$, possibly continuous regressors, and nonlinear kernel regressions.

Among new methods in this paper, (m9) is analogous to (m8) using the 'NNS' package. This paper provides algorithms for implementing four new measures of practical significance (m10 to m13) using machine learning algorithms. The machine learning methods can be sensitive to random seeds. The ultimate choice must depend on the scientific context.

This paper reports a real-world example of fuel economy regression. The average rank computed over eight methods is a good summary of various viewpoints deserving serious consideration. We also compute the average rank (avrank) over the eight methods.

References

Andersen, T. J. and Roggi, O. (2012), "Strategic Risk Management and Corporate Value Creation," Paper presented at the Strategic Management Society 32nd Annual International Conference. SMS 2012.,Prague, Czech Republic, URL `https://research-api.cbs.dk/ws/portalfiles/portal/58853215/Torben_Andersen.pdf`.

Anderson, T. J. (2008), "The performance relationship of effective risk management: Exploring the firm-specific investment rationale," *Long range planning*, 41 (2), 155–176.

Branzei, R., Branzei, O., Gök, S., and Tijs, S. (2010), "Cooperative interval games: a survey," *Central European Journal of Operations Research*, 387, 397–411.

Breiman, L. (2002), *Manual On Setting Up, Using, And Understanding Random Forests V3.1*, URL `https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf`.

Brodeur, A., Cook, N., and Heyes, A. (2020), "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110(11), 3634–3660.

Goodman, S. N. (2019), "Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics," *The American Statistician*, 73, 26–30.

Hirschauer, N., Gruner, S., Mushoff, O., Becker, C., and Jantsch, A. (2021), "Inference using non-random samples? Stop right there!" *Significance*, 18, 20–24, URL `https://doi.org/10.1111/1740-9713.01568`.

Khan, A., Khan, M. I., and Bhutto, N. A. (2020), "Reassessing the Impact of Risk Management Capabilities on Firm Value: A Stakeholders Perspective," *Sukkur IBA Journal of Management and Business*, 6(2), 81–98.

Pogrow, S. (2019), "How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings," *The American Statistician*, 73, 223–234.

Rosenthal, R. and Rubin, D. B. (1982), "A simple, general purpose display of magnitude of experimental effect," *Journal of educational psychology*, 74(2), 166–169.

Steiger, J. H. (2004), "Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis," *Psychological Methods*, 9(2), 164–182.

Vinod, H. D. (2014), "Matrix Algebra Topics in Statistics and Economics Using R," in

"Handbook of Statistics: Computational Statistics with R," , eds. Rao, M. B. and Rao, C. R., New York: North-Holland, Elsevier Science, vol. 34, chap. 4, pp. 143–176.

— (2021), "Generalized, Partial and Canonical Correlation Coefficients," *Computational Economics*, 59(1), 1–28.

— (2022), "Kernel regression coefficients for practical significance," *Journal of Risk and Financial Management*, 15, 1–13, URL `https://doi.org/10.3390/jrfm15010032`.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019), "Moving to a World Beyond p less than 0.05," *The American Statistician*, 73, 1–19.