Letter to the editor: Discussion of proposed t-statistic in "ppcor: An R Package for a fast calculation to semi-partial correlation coefficients," CSAM 2015; 22:665–674

Anthony Britto[1,a]

[a]Chair of Energy Economics, Karlsruhe Institute of Technology, Germany

1. Letter

Dear Prof. Seongjoo Song,

This letter concerns the article *ppcor: An R Package for a fast calculation to semi-partial correlation coefficients* by Prof. Seongho Kim, which appears in the November 2015 issue of *Communications for Statistical Applications and Methods* Kim (2015). I believe that there is a discrepancy which merits a detailed discussion between the t-statistic of the semi-partial correlation coefficient proposed by Prof. Kim, and the one proposed by Cohen *et al.* (2002).

Equation (2.8) in Prof. Kim's article lists the t-statistics of the partial and semi-partial correlation coefficients respectively as

$$t_{ij|S} = r_{ij|S} \sqrt{\frac{n-2-g}{1-r^2_{ij|S}}}, \tag{1.1}$$

$$t_{i(j|S)} = r_{i(j|S)} \sqrt{\frac{n-2-g}{1-r^2_{i(j|S)}}}, \tag{1.2}$$

where $r_{ij|S}$ (resp. $r_{i(j|S)}$) is the partial (resp. semi-partial) correlation coefficient of the random variables x_i and x_j with g covariates. Each of these variables are in the vector of random variables $X = (x_1, x_2, \ldots, x_n)^T$; X_S subsequently denotes the random sub-vector of X that results from deleting x_i and x_j.

An immediate consequence of the above formulae, which are identical as functions of $r_{ij|S}$ and $r_{i(j|S)}$, is that since $r_{ij|S} \neq r_{i(j|S)}$ in general, the respective t-statistics will also not agree: $t_{ij|S} \neq t_{i(j|S)}$ (We have $r_{ij|S} = r_{i(j|S)}$ only in the trivial case where each of the g covariates has zero correlation with x_i; cf. equations (2.1) and (2.2) in Kim (2015)). However, Cohen *et al.* (2002) argue that since the partial and semi-partial correlation coefficients are different scalings of the same statistical

[1] Chair of Energy Economics, Karlsruhe Institute of Technology, Hertzstr. 16 Building 06.33, Karlsruhe 76187, Germany.
E-mail: anthony.britto@kit.edu

phenomenon, they must yield an identical t-statistic. In fact, they claim further that the t-statistic of the partial and semi-partial correlation must also equal the one yielded by the β_i resulting from a multiple linear regression of x_i on x_j together with the other g covariates.

Their argument is as follows. Consider the case of a random variable y regressed on two predictors x_1 and x_2, where all variables have been standardised; the three effect sizes in the previous paragraph for the predictor x_1 for instance are then given by

$$\beta_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}, \tag{1.3}$$

$$r_{y1|2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{y2}^2}\sqrt{1 - r_{12}^2}}, \tag{1.4}$$

$$r_{y(1|2)} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}}. \tag{1.5}$$

Since these effects differ only with regard to their denominators, "none can equal zero unless the others are also zero, so it is also not surprising that they must yield the same t_i value t-statistic for the statistical significance of their departure from zero" (Cohen *et al.*, 2002).

The respective formulae for the t-statistics of these quantities are again found in Cohen *et al.* (2002) and elsewhere in the literature, for instance in the review article by Aloe and Thompson (2013). I reproduce here the formulae for the partial and semi-partial correlations (the formula for the t-statistic of the β_i is presented in the appendix), reverting now to the setup and notation of Prof. Kim above:

$$t_{ij|S} = r_{ij|S}\sqrt{\frac{\mathrm{df}}{1 - r_{ij|S}^2}}, \tag{1.6}$$

$$t_{i(j|S)} = r_{i(j|S)}\sqrt{\frac{\mathrm{df}}{1 - R_{i|jS}^2}}, \tag{1.7}$$

where df is the degrees of freedom, and $R_{i|jS}^2$ the coefficient of determination of the linear model where x_i is regressed on x_j and other covariates,

$$\tilde{x}_i = \beta_j\tilde{x}_j + \sum_{x_k \in X_g} \beta_k\tilde{x}_k \tag{1.8}$$

for $X_g \subset X$ the set of g covariates. The tilde indicates that the variables have been standardised; since this means that the intercept vanishes, we have

$$\mathrm{df} = n - g - 2. \tag{1.9}$$

Comparing equations (1.6) and (1.7) with equations (1.1) and (1.2), we see that the sole difference to Prof. Kim's article is in equation (1.2), namely, that $r_{i(j|S)}^2$ would need to be replaced with $R_{i|jS}^2$.

Table A.1: Description of variables in Duncan's dataset

`income`	Percentage of occupational incumbents in the 1950 US Census who earned \$3,500 or more per year (about \$36,000 in 2017 US dollars).
`education`	Percentage of occupational incumbents in 1950 who Ire high school graduates.
`prestige`	Percentage of respondents in a social survey who rated the occupation as "good" or better in prestige.

A trivial consequence of this, but one worth noting, is that such a change would automatically simplify equation (2.9) in Prof. Kim's article: since $t_{i,j|S} = t_{i(j|S)}$, the partial and semi-partial correlation coefficients would then have identical p-values:

$$p_{i,j|S} = 2\Phi\left(-\left|t_{i,j|S}\right|, \mathrm{df}\right) = 2\Phi\left(-\left|t_{i(j|S)}\right|, \mathrm{df}\right) = p_{i(j|S)}\,, \tag{1.10}$$

where $\Phi(\cdot)$ is the cumulative density function of a Student's t distribution with degrees of freedom df as in equation (1.9) above.

I present an example calculation within the framework of Cohen *et al.* (2002) in the appendix, demonstrating that their formulae do in fact work as claimed. I extend my gratitude to Prof. Kim for his `R` package, and look forward to a productive discussion on this discrepancy.

Sincerely,
Anthony Britto

Appendix: Example calculation: Duncan's occupational prestige data

I demonstrate here that the formulae of Cohen *et al.* (2002) work as claimed; I employ a standard, freely-available data set, *Duncan's Occupational Prestige Data*, available in the `carData` package in `R` Fox *et al* (2020). The data quantifies the prestige and other characteristics of 45 U.S. occupations in 1950; descriptions of the three columns used in the analysis are found in Table A.1.

The correlation matrix of the three variables is computed as follows:

	`income`	`education`	`prestige`
`income`	1.0000	0.7245	0.8378
`education`	0.7245	1.0000	0.8519
`prestige`	0.8378	0.8519	1.0000

From this matrix, it is possible to directly compute the β's and the partial and semi-partial correlations using equations (1.3) through (1.5) above, where I set `income` to be the y variable and `education` and `prestige` to be x_1 and x_2 respectively (see table below).

The t-statistics of the partial and semi-partial correlations can then be computed from equations (1.6) and (1.7); for equation (1.7), we need the coefficient of determination of the model where y is regressed on x_1 and x_2. Using the method of least squares, this is found to be $R^2_y = 0.7023$ (df $= 45 - 2 = 43$). As claimed by Cohen *et al.* (2002), the t-statistics of the partial and semi-partial

correlation coefficients agree; for the variable `prestige` for instance, we see that

$$t_{y2|1} = r_{y2|1}\sqrt{\frac{\text{df}}{1-r^2_{y2|1}}} = 0.6111\sqrt{\frac{43}{1-0.6111^2}} = 5.0625$$
$$\approx 0.4212\sqrt{\frac{43}{1-0.7023}} = r_{y(2|1)}\sqrt{\frac{\text{df}}{1-R^2_y}} = t_{y(2|1)}. \tag{A.1}$$

This should be contrasted with what occurs if the Prof. Kim's expression, equation (1.2), is employed:

$$t_{y(2|1)} = r_{i(j|S)}\sqrt{\frac{n-2-g}{1-r^2_{i(j|S)}}} = 0.4212\sqrt{\frac{43}{1-0.4212}} = 3.6304. \tag{A.2}$$

Finally, for the t-statistic of the β_i we have the usual formula

$$t(\beta_i) = \frac{\beta_i}{\text{s.e.}(\beta_i)}, \tag{A.3}$$

where s.e.(β_i) is the standard error of the estimate Cohen *et al.* (2002),

$$\text{s.e.}(\beta_i) = \sqrt{\frac{1-R^2_y}{\text{df}(1-R^2_{i|J})}}. \tag{A.4}$$

Here, R^2_y is as above, and $R^2_{i|J}$ is the coefficient of determination of the model where the ith predictor is regressed on all other predictors. In the case of a model with just two predictors, $R^2_{i|j} = r^2_{ij}$, and the standard errors of the coefficients are identical: 0.1589 in this case. Hence, from equations A.3, and then 1.10, I finally arrive at the following coefficient table, which neatly summarises that these effect sizes exist on different scales, but nevertheless share their statistical significance.

	β_i	$r_{yi\|j}$	$r_{y(i\|j)}$	t–stat.	p–value
`education`	0.0393	0.0377	0.0206	0.2473	0.8058
`prestige`	0.8043	0.6111	0.4212	5.0625	0.0000

References

Aloe A and Thompson C (2013). The synthesis of partial effect sizes, *Journal of the Society for Social Work and Research*, **4**, University of Chicago Press.

Cohen J, Cohen P, West S, and Aiken L (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.)*(pp64–101), Lawrence Erlbaum Associates, Inc.

Fox J, Weisberg S, and Price B (2020). carData: Companion to applied regression data sets, R package version 3.0-4.

Kim SH (2015). ppcor: An R Package for a fast calculation to semi-partial correlation coefficients, *Communications for Statistical Applications and Methods*, **22**, 665-674.

Vallat R (2018). Pingouin: Statistics in Python, *Journal of Open Source Software*, **3**.