Behav Res (2013) 45:880–895
DOI 10.3758/s13428-012-0289-7

SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients

Bruce Weaver · Karl L. Wuensch

Published online: 24 January 2013
© Psychonomic Society, Inc. 2013

Abstract Several procedures that use summary data to test hypotheses about Pearson correlations and ordinary least squares regression coefficients have been described in various books and articles. To our knowledge, however, no single resource describes all of the most common tests. Furthermore, many of these tests have not yet been implemented in popular statistical software packages such as SPSS and SAS. In this article, we describe all of the most common tests and provide SPSS and SAS programs to perform them. When they are applicable, our code also computes $100 \times (1 - \alpha)\%$ confidence intervals corresponding to the tests. For testing hypotheses about independent regression coefficients, we demonstrate one method that uses summary data and another that uses raw data (i.e., Potthoff analysis). When the raw data are available, the latter method is preferred, because use of summary data entails some loss of precision due to rounding.

Keywords Correlation · Regression · Ordinary least squares · SPSS · SAS

Electronic supplementary material The online version of this article (doi:10.3758/s13428-012-0289-7) contains supplementary material, which is available to authorized users.

B. Weaver
Human Sciences Division, Northern Ontario School of Medicine, Thunder Bay, ON, Canada P7B 5E1

B. Weaver (✉)
Centre for Research on Safe Driving, Lakehead University, Thunder Bay, ON, Canada P7B 5E1
e-mail: bweaver@lakeheadu.ca

K. L. Wuensch
Department of Psychology, East Carolina University, Greenville, NC, USA 27858-4353

Introduction

Several textbooks and articles describe methods for testing hypotheses concerning Pearson correlations and coefficients from ordinary least squares (OLS) regression models (e.g., Howell, 2013; Kenny, 1987; Potthoff, 1966; Raghunathan, Rosenthal, & Rubin, 1996; Steiger, 1980). However, we are not aware of any single resource that describes all of the most common procedures. Furthermore, many of the methods described in those various resources have not yet been implemented in standard statistical software packages such as SPSS and SAS. In some cases, data analysts may find stand-alone programs that perform the desired tests.[1] However, such programs *can* be relatively difficult to use (e.g., if they are old 16-bit DOS programs, they may not run on modern computers), or they may not provide all of the desired output (e.g., one program we found reports a *z*-test result, but not the corresponding *p*-value). It would be much more convenient if one could carry out all of these tests using one's usual statistical software. With that in mind, the twofold purpose of this article is to provide a single resource that briefly reviews the most common methods for testing hypotheses about Pearson correlations and OLS regression coefficients and to provide SPSS and SAS code that performs the calculations. When they are applicable, our code also computes $100 \times (1 - \alpha)\%$ confidence intervals (CIs) corresponding to the statistical tests.

We describe the various methods in this order: methods concerning (1) *single* parameters (e.g., testing the significance of a correlation), (2) two *independent* parameters (e.g., the difference between two independent correlations), (3) k independent parameters, where $k \geq 2$ (e.g., testing the

[1] For example, James Steiger's Multicorr program (http://www.statpower.net/Software.html) can be used to perform "single sample comparisons of correlations"; and Calvin Garbin's FZT program (http://psych.unl.edu/psycrs/statpage/comp.html) can be used to compute "a variety of r and R^2 comparison tests."

equivalence of three correlations), and (4) two *nonindependent* parameters (e.g., the difference between two nonindependent correlations). In all cases, SPSS and SAS programs to carry out the computations are provided as part of the online supplementary material, along with the output they generate. (The data files, code, and output are also available on the authors' Web sites: https://sites.google.com/a/lakeheadu.ca/bweaver/Home/statistics/spss/my-spss-page/weaver_wuensch and http://core.ecu.edu/psyc/wuenschk/W&W/W&W-SAS.htm.) Users can select the desired confidence level for CIs (when they are applicable) by setting the value of a variable called alpha (e.g., set alpha = .05 to obtain a 95 % CI, alpha = .01 to obtain a 99 % CI, etc.).

To illustrate the various methods, we use the lung function data set from Afifi, Clark, and May's (2003) book *Computer-Aided Multivariate Analysis*. We chose this data set for two reasons: (1) It contains variables suitable for demonstrating all of the methods we discuss, and (2) readers can easily download it in several formats (SAS, Stata, SPSS, Statistica, S-Plus, and ASCII) from the UCLA Academic Technology Services Web site (http://www.ats.ucla.edu/stat/spss/examples/cama4/default.htm). The data are from the UCLA study of chronic obstructive pulmonary disease. Afifi and coauthors described this file as "a subset including [nonsmoking] families with both a mother and a father, and one, two, or three children between the ages of 7 and 17 who answered the questionnaire and took the lung function tests at the first time period." The variables we use are area of the state (four levels) plus height (in inches) and weight (in pounds) for both fathers (variable names FHEIGHT and FWEIGHT) and mothers (MHEIGHT and MWEIGHT). Note that the initial F and M for the height and weight variables stand for *father's* and *mother's*, not *female* and *male*.

Input data for most of the code we provide consist of summary statistics that we computed using the lung function data. For example, we computed within each of the four different regions a correlation matrix for father's height, father's weight, mother's height, and mother's weight (variables FHEIGHT, FWEIGHT, MHEIGHT, and MWEIGHT). Table 1 shows those four correlation matrices. We also carried out some regression analyses, the results of which are displayed later in the article.

Methods for single parameters

Testing the null hypothesis that ρ = a specified value

The correlation matrices shown in Table 1 include a *p*-value for each correlation. If those same correlations appeared in a report or article that did not include *p*-values, one could work out the *p*-values by computing a *t*-test on each of the Pearson *r* values, as shown in Eq. 1. Under the null hypothesis that $\rho = 0$, the test statistic *t* is asymptotically distributed as *t* with $df = n - 2$:[2]

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \tag{1}$$

When a null hypothesis that specifies a *nonzero* value for ρ is tested, things are more complicated. As Howell (2013) put it, "When $\rho \neq 0$, the sampling distribution of *r* is not approximately normal (it becomes more and more skewed as $\rho \Rightarrow \pm 1.00$), and its standard error is not easily estimated" (p. 284). Fortunately, there is a straightforward solution to this problem: One can apply Fisher's (1921) *r*-to-*z* transformation to both *r* and ρ. Equation 2 shows the application of Fisher's transformation to *r*, and Eq. 3 shows the inverse transformation from r' to *r*.[3] Fisher showed that the sampling distribution of r' is approximately normal with variance equal to $1/(n-3)$, where *n* is the sample size. Taking the square root of that variance yields the standard error of r' (see Eq. 4):

$$r' = (0.5)\log_e\left|\frac{1+r}{1-r}\right| \tag{2}$$

$$r = \frac{e^{r'} - e^{-r'}}{e^{r'} + e^{-r'}} = \frac{e^{2r'} - 1}{e^{2r'} + 1} \tag{3}$$

$$s_{r'} = \sqrt{\frac{1}{n-3}}. \tag{4}$$

The final step is to compute a *z*-test (see Eq. 5). The *p*-value for this z-test is obtained in the usual fashion (i.e., using the standard normal distribution). This *z*-test *can* be used even when the null hypothesis states that $\rho = 0$ (and our

[2] The symbol ρ is the Greek letter *rho*. It is used to represent the population correlation.

[3] Because the sampling distribution of the transformed value is approximately normal, Fisher (1921) called it *z*. Following Howell (2013) and many other authors, we call it r' instead, in order to avoid confusion with the *z*-test value to be reported shortly. (Some authors use z_r rather than *z*, for the same reason.)

Table 1 Pearson correlations computed using the height and weight variables for fathers and mothers in the lung function data file

Area of state			Height of father in inches	Weight of father in pounds	Height of mother in inches	Weight of mother in pounds
Burbank (n = 24)	Height of father in inches	Pearson	1	.628**	.164	−.189
		Sig. (2-tailed)		.001	.443	.376
	Weight of father in pounds	Pearson	.628**	1	−.145	−.201
		Sig. (2-tailed)	.001		.499	.346
	Height of mother in inches	Pearson	.164	−.145	1	.624**
		Sig. (2-tailed)	.443	.499		.001
	Weight of mother in pounds	Pearson	−.189	−.201	.624**	1
		Sig. (2-tailed)	.376	.346	.001	
Lancaster (n = 49)	Height of father in inches	Pearson	1	.418**	.198	.065
		Sig. (2-tailed)		.003	.172	.660
	Weight of father in pounds	Pearson	.418**	1	−.181	.299*
		Sig. (2-tailed)	.003		.214	.037
	Height of mother in inches	Pearson	.198	−.181	1	.040
		Sig. (2-tailed)	.172	.214		.786
	Weight of mother in pounds	Pearson	.065	.299*	.040	1
		Sig. (2-tailed)	.660	.037	.786	
Long Beach (n = 19)	Height of father in inches	Pearson	1	.438	.412	.114
		Sig. (2-tailed)		.061	.079	.641
	Weight of father in pounds	Pearson	.438	1	−.032	.230
		Sig. (2-tailed)	.061		.898	.343
	Height of mother in inches	Pearson	.412	−.032	1	.487*
		Sig. (2-tailed)	.079	.898		.035
	Weight of mother in pounds	Pearson	.114	.230	.487*	1
		Sig. (2-tailed)	.641	.343	.035	
Glendora (n = 58)	Height of father in inches	Pearson	1	.589**	.366**	.071
		Sig. (2-tailed)		.000	.005	.596
	Weight of father in pounds	Pearson	.589**	1	.330*	.209
		Sig. (2-tailed)	.000		.011	.115
	Height of mother in inches	Pearson	.366**	.330*	1	.364**
		Sig. (2-tailed)	.005	.011		.005
	Weight of mother in pounds	Pearson	.071	.209	.364**	1
		Sig. (2-tailed)	.596	.115	.005	

code computes it), but in that case, the *t*-test shown in Eq. 1 is preferred.

Equation 6 shows how the standard error of r' (Eq. 4) can be used to compute a CI for ρ'. The $z_{\alpha/2}$ in Eq. 6 represents the critical value of z for a two-tailed test with α set to the desired level. For a 95 % CI, for example, $\alpha = .05$, and $z_{\alpha/2} = 1.96$. The *inverse* of the *r*-to-*z* transformation (Eq. 3) is used to convert the lower and upper confidence limits for ρ' into confidence limits for ρ:

$$z = \frac{r' - \rho'}{s_{r'}} = \frac{r' - \rho'}{\sqrt{\frac{1}{n-3}}} \tag{5}$$

$$100(1-\alpha)\%\ \text{CI for}\ \rho' = \mathrm{r}' \pm \mathrm{z}_{\alpha/2}\mathrm{s}_{\mathrm{r}'}. \tag{6}$$

Whereas the choice of test statistic (*t* vs. *z*) depends on whether the null hypothesis specifies that $\rho = 0$ versus some nonzero value, the computation of confidence limits for ρ does not. The method shown in Eq. 6 is used to compute a CI regardless of the value of ρ under the null hypothesis.

Our code for illustrating these methods requires the following input variables: r (the observed Pearson *r*), rho (the population correlation according to H_0), n (the sample size), alpha (the value used to determine the confidence level for the CI on rho), and Note, a text field in which a brief

description of the data can be entered. The SPSS code for this situation has the following DATA LIST command[4]:

```
DATA LIST LIST / r rho (2f5.3) n (f5.0) alpha (f5.3) Note (a30).
BEGIN DATA
.628 .000 24 .05 "rho=0|95% CI|Bur"
.418 .000 49 .05 "rho=0|95% CI|Lan"
.438 .000 19 .05 "rho=0|95% CI|L Beach"
.589 .000 58 .05 "rho=0|95% CI| Glen"
.628 .650 24 .05 "rho=.65|95% CI|Bur"
.628 .650 24 .01 "rho=.65|99% CI|Bur"
.418 .650 49 .05 "rho=.65|95% CI|Lan"
.418 .650 49 .01 "rho=.65|99% CI|Lan"
.438 .650 19 .05 "rho=.65|95% CI|L Beach"
.438 .650 19 .01 "rho=.65|99% CI|L Beach"
.589 .650 58 .05 "rho=.65|95% CI|Glen"
.589 .650 58 .01 "rho=.65|99% CI|Glen"
END DATA.
```

The correlations entered in variable r are the correlations between father's height and father's weight for the four areas of the state (see Table 1). The first four rows of input set rho = 0, whereas the last eight rows set rho = .650.[5] Therefore, our code uses the *t*-test shown in Eq. 1 for only the first four rows, whereas the *z*-test in Eq. 5 is computed for every row. Note too that the value of alpha is .05 in some rows of input data and .01 in others. Our code computes 95 % CIs where alpha = .05 and 99 % CIs where alpha = .01.[6] All CIs are computed via Eq. 6. The output from our SPSS code is listed below.

```
     r    rho     n     t    df   p_t      z   p_z  alpha Lower Upper Note

  .628   .000    24 3.785    22  .001  3.382  .001   .050  .301  .823 rho=0|95% CI|Bur
  .418   .000    49 3.154    47  .003  3.020  .003   .050  .155  .626 rho=0|95% CI|Lan
  .438   .000    19 2.009    17  .061  1.879  .060   .050 -.020  .744 rho=0|95% CI|L Beach
  .589   .000    58 5.454    56  .000  5.014  .000   .050  .390  .735 rho=0|95% CI| Glen

  .628   .650    24  .        .  .     -.170  .865   .050  .301  .823 rho=.65|95% CI|Bur
  .628   .650    24  .        .  .     -.170  .865   .010  .174  .862 rho=.65|99% CI|Bur

  .418   .650    49  .        .  .    -2.238  .025   .050  .155  .626 rho=.65|95% CI|Lan
  .418   .650    49  .        .  .    -2.238  .025   .010  .065  .678 rho=.65|99% CI|Lan

  .438   .650    19  .        .  .    -1.222  .222   .050 -.020  .744 rho=.65|95% CI|L Beach
  .438   .650    19  .        .  .    -1.222  .222   .010 -.172  .805 rho=.65|99% CI|L Beach

  .589   .650    58  .        .  .     -.735  .462   .050  .390  .735 rho=.65|95% CI|Glen
  .589   .650    58  .        .  .     -.735  .462   .010  .317  .771 rho=.65|99% CI|Glen

* When rho = 0, the t-test is preferred to the z-test.
* The confidence level for CI = (1-alpha)*100.
```

The *t*-test results in the first four rows of output indicate that the correlation between height and weight (for fathers) is statistically significant in all four areas except area 3, Long Beach. Note too that the *p*-values for those correlations (.001, .003, .060, and .000) agree almost perfectly with the *p*-values reported in Table 1. The only differences (e.g., .060 vs. .061 for Long Beach) are due to loss of precision resulting from our use of summary data.

[4] Users who wish to analyze their own data can do so by replacing the data lines between BEGIN DATA and END DATA and then running the syntax.

[5] We don't know the value of the actual population correlation between height and weight of the fathers. We chose .650 because it was convenient for producing a mix of significant and nonsignificant *z*-tests.

[6] In general, our code computes CIs with confidence level = $100(1-\alpha)\%$.

Table 2 Parameter estimates for four simple linear regression models with father's height regressed on father's weight; father's height was centered on 60 in. (5 ft)

Coefficients[a]

Area		Unstandardized coefficients		Standardized coefficients	t	Sig.	95.0 % confidence interval for B	
		B	Std. error	Beta			Lower bound	Upper bound
Burbank (n = 24)	(Constant)	142.011	10.664		13.317	.000	119.896	164.127
	Height of father (centered on 60 in)	4.179	1.105	.628	3.781	.001	1.887	6.472
Lancaster (n = 49)	(Constant)	148.053	11.142		13.288	.000	125.638	170.468
	Height of father (centered on 60 in)	3.709	1.177	.418	3.151	.003	1.341	6.078
Long Beach (n = 19)	(Constant)	144.038	18.250		7.893	.000	105.535	182.541
	Height of father (centered on 60 in)	3.749	1.866	.438	2.009	.061	−.187	7.685
Glendora (n = 58)	(Constant)	130.445	10.228		12.753	.000	109.955	150.935
	Height of father (centered on 60 in)	5.689	1.044	.589	5.451	.000	3.598	7.780

[a] Dependent Variable: weight of father in pounds

In the final eight rows of output, where $\rho = .650$ under the null hypothesis, only the *z*-test is computed. Note that we included the input data for each area twice, first with alpha = .05 and again with alpha = .01. Thus, the first line of output for each area displays a 95 % CI, and the second a 99 % CI. The *z*-test result is unaffected by the value of alpha, which is why the same test result appears twice for each area. Only in area 1 (Lancaster) does the observed correlation differ significantly from .650, $z = -2.238$, $p = .025$.

Testing the hypothesis that *b* = a specified value

The data we use as an illustration in this section come from four simple linear regression models (one for each area) with father's weight regressed on father's height. In order to make the intercepts more meaningful, we first centered height on 60 in. (5 ft).[7] Parameter estimates for the four models are shown in Table 2.

In his discussion of this topic, Howell (2013) began by showing that the standard error of *b* (s_b) can be computed from the standard error of *Y* given *X* ($s_{Y|X}$), the standard deviation of the *X* scores (s_X), and the sample size (*n*). Given that $s_{Y|X} = \sqrt{MS_{error}}$, or the *root mean square error* (RMSE), s_b can be computed as shown in Eq. 7:

$$s_b = \frac{RMSE}{s_X\sqrt{n-1}} = \frac{RMSE}{\sqrt{SS_X}}. \tag{7}$$

However, it is extremely difficult to imagine circumstances under which one would have the RMSE from the regression model (plus the sample size and the standard deviation of *X*), but *not* the standard error of *b*. Therefore, we do not provide code to compute the standard error of *b* as shown in Eq. 7. Instead, we simply take the standard error of *b* from the regression output and plug it into Eq. 8, which shows a *t*-test for the null hypothesis that *b**, the population parameter corresponding to *b*, is equal to a specified value.[8] The *m* in the subscript is the number of predictor variables, not including the constant, and $n - m - 1$ equals the degrees of freedom for the *t*-test.[9] The standard error of *b* is also used to compute a $100(1-\alpha)$% CI for *b** (Eq. 9):

$$t_{n-m-1} = \frac{b - b^*}{s_b} \tag{8}$$

$$100(1-\alpha)\%\ \text{CI for b}^* = \text{b} \pm \text{t}_{\alpha/2}\text{s}_\text{b}. \tag{9}$$

[7] In other words, we used a transformed height variable equal to height minus 60 in. If we had used the original height variable, the constant from our model would have given the fitted value of weight when height = 0, which would be nonsensical. With height centered on 60 in., the constant gives the fitted value of weight when height = 60 in.

[8] We follow Howell (2013) in using *b** rather than β to represent the parameter corresponding to *b*. We do this to avoid "confusion with the standardized regression coefficient," which is typically represented by β.

[9] Although some authors use *p* to represent the number of predictors in a regression model, we use *m* in this context in order to avoid confusion with the *p*-value.

As we saw earlier, when testing hypotheses about ρ, we can use the *t*-test shown in Eq. 1 when the null hypothesis states that $\rho = 0$; but when the null hypothesis states that $\rho =$ some *nonzero* value, we must apply Fisher's *r*-to-*z* transformation to both *r* and ρ and then use the *z*-test shown in Eq. 5. For regression coefficients, on the other hand, the *t*-test shown in Eq. 8 can be used regardless of the value of b^*. In other words, when $b^* = 0$, we will get the usual *t*-test shown in the table of regression coefficients. To confirm this, we plugged the displayed values of the intercept and slope into our implementation of Eq. 8 and set $b^* = 0$. Doing so produced the following output:

b	bstar	se	t	df	p	alpha	CI_Lower	CI_Upper	Note
142.011	.000	10.664	13.317	22	.000	.050	119.895	164.127	Int, Bur
148.053	.000	11.142	13.288	47	.000	.050	125.638	170.468	Int, Lan
144.038	.000	18.250	7.892	17	.000	.050	105.534	182.542	Int, L Beach
130.445	.000	10.228	12.754	56	.000	.050	109.956	150.934	Int, Glen
4.179	.000	1.105	3.782	22	.001	.050	1.887	6.471	Slope, Bur
3.709	.000	1.177	3.151	47	.003	.050	1.341	6.077	Slope, Lan
3.749	.000	1.866	2.009	17	.061	.050	-.188	7.686	Slope, L Beach
5.689	.000	1.044	5.449	56	.000	.050	3.598	7.780	Slope, Glen

Apart from some rounding error, the results of these *t*-tests match those shown in Table 2. Note that alpha = .05 on every line, so all CIs are 95 % CIs.

Now suppose that we have reason to believe that the *true* population values for the intercept and slope are 145 and 3.5, respectively, and we wish to compare our sample values with those parameters. Plugging the observed intercepts and slopes into our SPSS implementation of Eq. 8 with $b^* = 145$ for intercepts and $b^* = 3.5$ for slopes, we get the output listed below:

b	bstar	se	t	df	p	alpha	CI_Lower	CI_Upper	Note
142.011	145.000	10.664	-.280	22	.782	.050	119.895	164.127	Int, Bur
148.053	145.000	11.142	.274	47	.785	.050	125.638	170.468	Int, Lan
144.038	145.000	18.250	-.053	17	.959	.050	105.534	182.542	Int, L Beach
130.445	145.000	10.228	-1.423	56	.160	.050	109.956	150.934	Int, Glen
4.179	3.500	1.105	.614	22	.545	.050	1.887	6.471	Slope, Bur
3.709	3.500	1.177	.178	47	.860	.050	1.341	6.077	Slope, Lan
3.749	3.500	1.866	.133	17	.895	.050	-.188	7.686	Slope, L Beach
5.689	3.500	1.044	2.097	56	.041	.050	3.598	7.780	Slope, Glen

Looking first at the results for the intercepts, we would fail to reject the null hypothesis (that $b^* = 145$) in all four cases, because all *p*-values are greater than .05. For the slopes, on the other hand, we would reject the null hypothesis (that $b^* = 3.5$) for Glendora, $t(56) = 2.097$, $p = .041$, but not for any of the other three areas (where all *t*-ratios are <1 and all *p*-values are ≥ .545).

Methods for two independent parameters

We now shift our focus to tests and CIs for the difference between two independent parameters.

Testing the difference between two independent correlations

When the correlation between two variables is computed in two independent samples, one may wish to test the null hypothesis that the two population correlations are the same (H_0: $\rho_1 = \rho_2$). To test this null hypothesis, we use a simple extension of the method for testing the null that ρ = a specified value. As in that case, we must apply Fisher's *r*-to-*z* transformation to convert the two sample correlations into r' values. As is shown in Eq. 4, the standard error of an r' value is $\sqrt{1/(n-3)}$. Squaring that expression (i.e., removing the square root sign) gives the variance of the

sampling distribution of r'. The variance of the difference between two independent r' values is the sum of their variances.[10] Taking the square root of that sum of variances yields the standard error of the difference between two independent r' values (see Eq. 10). That standard error is used as the denominator in a z-test (see Eq. 11):

$$s_{r'_1-r'_2} = \sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}} \tag{10}$$

$$z = \frac{r'_1 - r'_2}{s_{r'_1-r'_2}} = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}. \tag{11}$$

We illustrate these computations using several independent pairs of correlations from Table 1.[11] In each case, we compare the values for Lancaster and Glendora, the two areas with the largest sample sizes. Plugging the needed values into our implementation of Eq. 11 gave us the output shown below:

```
     r1      r2      rp1     rp2   rpdiff  sediff      z      p Note

   .418    .589    .445    .676    -.231    .200 -1.155   .248 r(FHT,FWT), Lan v Glen
   .040    .364    .040    .381    -.341    .200 -1.709   .087 r(MHT,MWT), Lan v Glen
   .198    .366    .201    .384    -.183    .200  -.917   .359 r(FHT,MHT), Lan v Glen
   .299    .209    .308    .212     .096    .200   .482   .630 r(FWT,MWT), Lan v Glen
  -.181    .330   -.183    .343    -.526    .200 -2.632   .008 r(FWT,MHT), Lan v Glen
   .065    .071    .065    .071    -.006    .200  -.030   .976 r(FHT,MWT), Lan v Glen
   .490    .360    .536    .377     .159    .138  1.156   .248 Zou (2007) Example 1

* rp1 = r-prime for r1; rp2 = r-prime for r2.
* rpdiff = rp1-rp2; sediff = SE(rp1-rp2).
* FHT = Father's height; MHT = Mother's height
```

In the *Note* column, the initial F and H stand for *father's* and *mother's* respectively, and HT and WT stand for *height* and *weight*. Thus, the *r(FHT,FWT)* on the first line indicates that the correlation between father's height and father's weight has been computed for both Lancaster and Glendora and the two correlations have been compared. The *rp1* and *rp2* columns give the r' values corresponding to r_1 and r_2. (Standard errors for *rp1* and *rp2* are also computed but are not listed here, in order to keep the output listing to a manageable width.) The *rpdiff* and *sediff* columns show the numerator and denominator of Eq. 11. The null hypothesis (that $\rho_1 - \rho_2 = 0$) can be rejected only for the test comparing the correlations between father's weight and mother's height, $z = -2.632$, $p = .008$. For all other comparisons, the p-values are greater than .05.

Our code also computes $100 \times (1 - \alpha)$% CIs for ρ_1, ρ_2, and $\rho_1 - \rho_2$. CIs for ρ_1 and ρ_2 are obtained by computing CIs for ρ'_1 and ρ'_2 (see Eq. 6) and then back-transforming them (Eq. 3). The CI for $\rho_1 - \rho_2$ is computed using Zou's (2007) modified asymptotic (MA) method.[12] The first listing below shows CIs for ρ_1 and ρ_2, and the second listing shows the CI for $\rho_1 - \rho_2$. (We included Zou's example in order to verify that our code for his method was correct.) Alpha = .05 in all cases, so they are all 95 % CIs.

[10] More generally, the variance of the difference is the sum of the variances minus two times the covariance. But when the samples are independent, the covariance is equal to zero.

[11] Readers may wonder why we do not compare the correlation between height and weight for fathers with the same correlation for mothers. Given that there are matched pairs of fathers and mothers, those correlations are not independent. Therefore, it would be inappropriate to use this method for comparing them. However, we do compare those two correlations later, using the *ZPF* statistic, which takes into account the dependency.

[12] We use Zou's MA method because his simulations demonstrate that for the situations listed below, it provides coverage much closer (on average) to the nominal confidence level than do the more conventional methods: (1) comparing two independent correlations, (2) comparing two correlated correlations with one variable in common, and (3) comparing two correlated correlations with no variables in common.

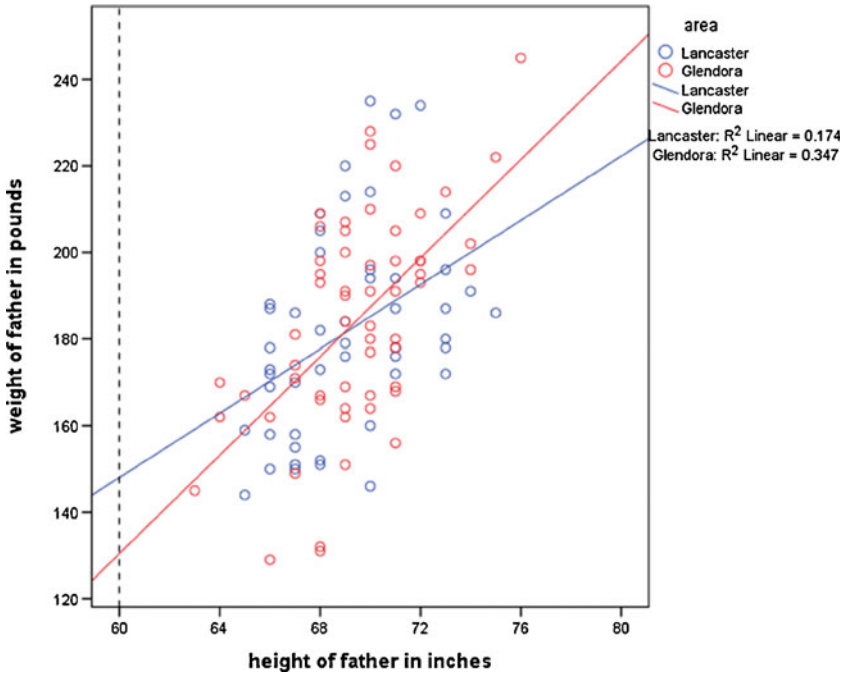

Fig. 1 The relationship between fathers' heights and weights in the Lancaster and Glendora samples (blue and red symbols respectively). Height was centered on 60 in.; therefore, the intercepts for the two models (148.053 and 130.45) occur at the intersections of the two regression lines with the dashed line at height = 60

```
     r1   Lower1   Upper1       r2   Lower2   Upper2   alpha Note

   .418     .155     .626     .589     .390     .735    .050 r(FHT,FWT), Lan v Glen
   .040    -.244     .318     .364     .117     .569    .050 r(MHT,MWT), Lan v Glen
   .198    -.088     .454     .366     .119     .570    .050 r(FHT,MHT), Lan v Glen
   .299     .019     .535     .209    -.052     .443    .050 r(FWT,MWT), Lan v Glen
  -.181    -.440     .106     .330     .078     .542    .050 r(FWT,MHT), Lan v Glen
   .065    -.220     .340     .071    -.191     .323    .050 r(FHT,MWT), Lan v Glen
   .490     .355     .605     .360     .162     .530    .050 Zou (2007) Example 1

* CIs for rho1 and rho2.
* FHT = Father's height; MHT = Mother's height.
```

```
     r1       r2 Lower_diff Upper_diff   alpha Note

   .418     .589     -.472       .117    .050 r(FHT,FWT), Lan v Glen
   .040     .364     -.674       .048    .050 r(MHT,MWT), Lan v Glen
   .198     .366     -.520       .188    .050 r(FHT,MHT), Lan v Glen
   .299     .209     -.275       .442    .050 r(FWT,MWT), Lan v Glen
  -.181     .330     -.846      -.130    .050 r(FWT,MHT), Lan v Glen
   .065     .071     -.387       .374    .050 r(FHT,MWT), Lan v Glen
   .490     .360     -.087       .359    .050 Zou (2007) Example 1

* CI for (rho1 - rho2) computed using Zou's (2007) method.
* FHT = Father's height; MHT = Mother's height.
```

Testing the difference between two independent regression coefficients

If one has the results for OLS linear regression models from two independent samples, with the same criterion and explanatory variables used in both models, there may be some interest in testing the differences between corresponding coefficients in the two models.[13] The required test is a simple extension of the *t*-test described earlier for testing the null hypothesis that b^* = a specified value (see Eq. 8).

[13] If one has the raw data for both samples, the same comparisons can be achieved more directly by running a single model that uses all of the data and includes appropriate interaction terms. We will demonstrate that approach shortly.

As was noted earlier, when one is dealing with two independent samples, the variance of a difference is the sum of the variances, and the standard error of the difference is the square root of that sum of variances. Therefore, the standard error of the difference between b_1 and b_2, two independent regression coefficients, is computed as shown in Eq. 12, where the two terms under the square root sign are the squares of the standard errors for b_1 and b_2. This standard error is used to compute the *t*-test shown in Eq. 13 and to compute the 100(1 − α)% CI (Eq. 14). The *t*-test has $df=n_1+n_2-2m-2$ (where m = the common number of predictor variables in the two regression models, not including the constant).[14] Some books (e.g., Howell, 2013) give the degrees of freedom for this *t*-test as n_1+n_2-4. That is because they are describing the special case where $m = 1$ (i.e., the two regression models have only one predictor variable). And of course, $n_1+n_2-2(1)-2=n_1+n_2-4$.

$$s_{b_1-b_2} = \sqrt{s_{b_1}^2 + s_{b_2}^2} \tag{12}$$

$$t_{(n_1+n_2-2m-2)} = \frac{b_1 - b_2}{s_{b_1-b_2}} \tag{13}$$

$$100(1-\alpha)\%\ \text{CI for}\left(b_1^* - b_2^*\right) = (b_1 - b_2) \pm t_{\alpha/2} s_{b_1-b_2}. \tag{14}$$

To illustrate, we use the results for Lancaster and Glendora shown in Table 2 and also depicted graphically in Fig. 1. Specifically, we compare the regression coefficients (both intercept and slope) for Lancaster and Glendora. Plugging the coefficients and their standard errors (and sample sizes) into our code for Eq. 13, we get the output listed below:

b1	b2	bdiff	sediff	t	df	p	Note
148.053	130.445	17.608	15.125	1.164	103	.247	Int, Lan v Glen
3.709	5.689	-1.980	1.573	-1.259	103	.211	Slope, Lan v Glen

The *bdiff* and *sediff* columns show the difference between the coefficients and the standard error of that difference—that is, the numerator and denominator of Eq. 13. Since both *p*-values are greater than .05, the null hypothesis cannot be rejected in either case. The next listing shows the CIs for *bdiff*. Because alpha = .05 on both lines of output, these are 95 % CIs:

b1	b2	bdiff	sediff	alpha	CI_Lower	CI_Upper	Note
148.053	130.445	17.608	15.125	.050	-12.388	47.604	Int, Lan v Glen
3.709	5.689	-1.980	1.573	.050	-5.100	1.140	Slope, Lan v Glen

The method we have just shown is fine in cases where one does not have access to the raw data but does have access to the required summary data. However, when the raw data are available, one can use another approach that provides more accurate results (because it eliminates rounding error). The approach we are referring to is sometimes called *Potthoff analysis* (see Potthoff, 1966).[15] It entails running a hierarchical regression model. The first step includes only the predictor variable of primary interest (height in this case). On the second step, $k - 1$ indicator variables are added to differentiate between the k independent groups. The products of those indicators with the main

[14] In Eqs. 12–14, the subscripts on b_1 and b_2 refer to which *model* the coefficients come from, not which explanatory variable they are associated with, as is typically done for models with two or more explanatory variables.

[15] Also see these unpublished documents on the second author's Web site: http://core.ecu.edu/psyc/wuenschk/docs30/CompareCorrCoeff.pdf, http://core.ecu.edu/psyc/wuenschk/MV/multReg/Potthoff.pdf).

predictor variable are also added on step 2. In this case, we have $k = 2$ groups (Lancaster and Glendora), so we add only one indicator variable and one product term on step 2. (We chose to use an indicator for area 2, Lancaster, thus making Glendora the reference category.) The SPSS commands to run this model were as follows, with fweight = father's weight, fht60 = father's height centered on 60 in., A2 = an indicator for area 2 (Lancaster), and FHTxA2 = the product of fht60 and A2:

```
* Compute indicators for Area and Area x Height products.
DO REPEAT a = A1 to A4 / Int = FHTxA1 to FHTxA4 / # = 1 to 4 .
+ COMPUTE a = (area EQ #). /* This computes A1 to A4 indicators.
+ COMPUTE int = a * fht60. /* This computes the 4 AxHt products.
END REPEAT.
FORMATS A1 to A4 (f1) / FHTxA1 to FHTxA4 (f5.0).
COMPUTE A24 = any(area,2,4). /* Lancaster & Glendora.
filter by A24. /* Use only records from Lancaster & Glendora.
REGRESSION
  /STATISTICS COEFF OUTS CI(95) R ANOVA CHANGE
  /DEPENDENT fweight
  /METHOD=ENTER fht60 /ENTER A2 FHTxA2.
```

The F-test on the change in R^2 (from step 1 to step 2) tests the null hypothesis of *coincidence*, which states that the two population regression lines are identical (i.e., they have the same intercept *and* the same slope). In the table of coefficients for the full model (step 2), the t-test for the area 2 indicator variable tests the null hypothesis that the population *intercepts* are the same, and the t-test for the height × A2 product term tests the null hypothesis that the two population *slopes* are equal. (The t-test for height in the full model tests the null hypothesis that the population slope = 0 for the reference group—that is, the group for which the area 2 indicator variable = 0.)

We ran that hierarchical regression analysis for the Lancaster and Glendora data and found that the change in R^2 from step 1 to step 2 = .011, $F(2, 103) = 0.816$, $MS_{\text{residual}} = 44{,}362.179$, $p = .445$. Therefore, the null hypothesis of coincidence of the regression lines cannot be rejected. Normally, we would probably stop at this point, because there is no great need to compare the slopes and intercepts separately if we have already failed to reject the null hypothesis of coincident regression lines. However, in order to compare the results from this Potthoff analysis with results obtained earlier via Eq. 13, we shall proceed.

The regression coefficients for both steps of our hierarchical model are shown in Table 3. Looking at the step 2, the coefficient for the area 2 indicator is equal to the difference between the intercepts for Burbank and Glendora (see Table 2). The t-test for the area 2 indicator is not statistically significant, $t(103) = 1.168$, $p = .245$. Therefore, the null hypothesis that the two population intercepts are equal cannot be rejected. The coefficient for the height × A2 product term gives the difference between the slopes for Burbank and Glendora. The t-test for this product term is not statistically significant, $t(103) = -1.264$, $p = .209$. Therefore, the null hypothesis that the population slopes are the same cannot be rejected either. Finally, note that apart from rounding error, the results of these two tests match the results we got earlier by plugging summary data into Eq. 13: $t(103) = 1.164$, $p = .247$, for the intercepts; and $t(103) = -1.259$, $p = .211$, for the slopes. (As has been noted, methods that use the raw data are generally preferred over methods that use summary data, because the former eliminate rounding error.)

Methods for *k* independent parameters

On occasion, one may wish to test a null hypothesis that says that three or more independent parameters are all equivalent. This can be done using the test of heterogeneity that is familiar to meta-analysts (see Fleiss, 1993, for more details). The test statistic is often called Q[16] and is computed as follows:

$$Q = \sum_{i=1}^{k} W_i \left(Y_i - \overline{Y}\right)^2, \tag{15}$$

where k = the number of independent parameters, Y_i = the estimate for the ith parameter, W_i = the reciprocal of its variance, and $\overline{Y}$ = a weighted average of the k parameter estimates, which is computed as shown in Eq. 16. When the

[16] Meta-analysts often describe this statistic as *Cochran's Q* and cite Cochran (1954). This may cause some confusion, however, because *Cochran's Q* often refers to a *different* statistic used to compare k related dichotomous variables, where $k \geq 3$. That test is described in Cochran (1950).

Table 3 Parameter estimates for a hierarchical regression model with height entered on step 1 and an area 2 (Lancaster) indicator and its product with height both entered on step 2

Coefficient[a]

Step		Unstandardized coefficients		Standardized coefficients	t	Sig.	95.0 % confidence interval for B	
		B	Std. error	Beta			Lower bound	Upper bound
1	(Constant)	138.793	7.510		18.481	.000	123.902	153.684
	Height of father (centered on 60 in)	4.771	.778	.513	6.130	.000	3.228	6.314
2	(Constant)	130.445	10.511		12.410	.000	109.598	151.292
	Height of father (centered on 60 in)	5.689	1.073	.612	5.304	.000	3.562	7.816
	Area 2 indicator	17.608	15.075	.367	1.168	.245	−12.289	47.505
	Height × A2	−1.979	1.566	−.403	−1.264	.209	−5.086	1.127

a. Dependent Variable: weight of father in pounds

null hypothesis is true (i.e., when all population parameters are equivalent), Q is distributed (approximately) as chi-square with $df = k - 1$.

$$\overline{Y} = \frac{\sum W_i Y_i}{\sum W_i}. \tag{16}$$

An example using regression coefficients

We illustrate this procedure using output from the four simple linear regression models summarized in Table 2. Using the method described above to test the null hypothesis that the four population *intercepts* are all the same, we get $Q = 1.479$, $df = 3$, $p = .687$. And testing the null hypothesis that the *slopes* are all the same, we get $Q = 1.994$, $df = 3$, $p = .574$. Therefore, we cannot reject the null hypothesis in either case.

Because the raw data are available in this case, we can also test the null hypothesis that all slopes are the same by performing another Potthoff analysis, like the one described earlier. With $k = 4$ groups (or areas), we will need three (i.e., $k - 1$) indicator variables for area and three product terms. The test of coincidence will contrast the full model with a model containing only the continuous predictor variable (height). The test of intercepts will contrast the full model with a model from which the $k - 1$ indicator variables have been removed. The test of slopes will contrast the full model with a model from which the k-1 interaction terms have been dropped.

Using SPSS, we ran a hierarchical regression model with height entered on step 1. On step 2, we added three indicators for area plus the products of those three indicators with height. The SPSS REGRESSION command for this analysis was as follows:

```
USE ALL.
FILTER OFF. /* use all 4 areas again.

REGRESSION
  /STATISTICS COEFF OUTS CI(95) R ANOVA CHANGE
  /DEPENDENT fweight
  /METHOD=ENTER fht60
  /TEST (fht60) (A1 A2 A3) (FHTxA1 FHTxA2 FHTxA3).
```

Table 4 shows the ANOVA summary table for this model, and Table 5 shows the parameter estimates. Because we used the TEST method (rather than the default ENTER method) for step 2 of the REGRESSION command, the ANOVA summary table includes the multiple degree of freedom tests we need to test the null hypotheses that all intercepts and all slopes are the same (see the "Subset Tests" section in Table 4). For SAS code that produces the same results, see the online supplementary material or the second author's Web site (http://core.ecu.edu/psyc/wuenschk/W&W/W&W-SAS.htm).

The R^2 values for steps 1 and 2 of our hierarchical regression model were .272 and .286, respectively, and the

Table 4 ANOVA summary table for the hierarchical regression model with height entered on step 1 and three area indicators and their products with height entered on step 2

ANOVA[d]

Step			Sum of squares	df	Mean square	F	Sig.	R square change
1	Regression		23221.739	1	23221.739	55.189	.000[a]	
	Residual		62274.134	148	420.771			
	Total		85495.873	149				
2	Subset Tests	Height of father (centered on 60 in)	12117.765	1	12117.765	28.182	.000[b]	.142
		Area 1 indicator, Area 2 indicator, Area 3 indicator	631.267	3	210.422	.489	.690[b]	.007
		Height × A1, Height × A2, Height × A3	850.075	3	283.358	.659	.579[b]	.010
	Regression		24438.781	7	3491.254	8.120	.000[c]	
	Residual		61057.092	142	429.980			
	Total		85495.873	149				

a. Predictors: (Constant), Height of father (centered on 60 in)

b. Tested against the full model

c. Predictors in the Full Model: (Constant), Height of father (centered on 60 in), Area 3 indicator, Area 1 indicator, Height × A2, Height × A1, Height × A3, Area 2 indicator

d. Dependent Variable: weight of father in pounds

change in R^2 from step 1 to 2 was equal to .014, $F(6, 142) = 0.472$, $p = .828$.[17] Therefore, the null hypothesis of coincident regression lines cannot be rejected. Nevertheless, we shall report separate tests for the intercepts and slopes in order to compare the results from this analysis with those we obtained earlier via Eq. 15. The test for homogeneity of the intercepts is the second Subset Test in Table 4—that is, the combined test for the area 1, area 2, and area 3 indicators. It shows that the null hypothesis of homogeneous intercepts cannot be rejected, $F(3, 142) = 0.489$, $p = .690$. When testing this same hypothesis via Eq. 15, we got $Q = 1.479$, $df = 3$, $p = .687$. The test of homogeneity of the slopes in the Potthoff analysis is the third Subset Test in Table 4—that is, the combined test for the three product terms. It shows that the null hypothesis of homogeneous slopes cannot be rejected, $F(3, 142) = 0.659$, $p = .579$. Earlier, using Eq. 15, we got $Q = 1.994$, $df = 3$, $p = .574$, when testing for homogeneity of the slopes.

Note that for both of these tests, the p-values for the Q and the F-tests are very similar. The differences are partly due to rounding error in the computation of Q (where we rounded the coefficients and their standard errors to three decimals) and partly due to the fact that the denominator degrees of freedom for the F-tests are less than infinite. For a good discussion of the relationship between F and χ^2 tests (bearing in mind that Q is approximately distributed as χ^2 when the null hypothesis is true), see Gould's (2009) post on the Stata FAQ Web site (http://www.stata.com/support/faqs/stat/wald.html).

Finally, we should clarify how the coefficients and t-tests for the full model (Table 5, step 2) are interpreted. The intercept for the full model is equal to the intercept for area 4 (Glendora), the omitted reference group (see Table 2 for confirmation). The coefficients for the three area indicators give the differences in intercepts between each of the other three areas and area 4 (with the area 4 intercept subtracted from the other intercept in each case). None of those pairwise comparisons are statistically significant (all p-values $\geq .244$). The coefficient for height gives the *slope* for area 4, and the coefficients for the three product terms give differences in slope between each of the other areas and area 4 (with the area 4 slope subtracted from the other slope). None of the pairwise comparisons for slope are statistically significant either (all p-values $\geq .208$).

An example using correlation coefficients

When using the test of heterogeneity with correlations, it is advisable to first apply Fisher's r-to-z transformation. To illustrate, we use the correlation between father's height and father's weight in Table 1. The values of that correlation in the four areas were .628, .418, .438, and .589 (with sample sizes of 24, 49, 19, and 58, respectively). The r' values for these

[17] The three "R Square Change" values in Table 4 give the change in R^2 for removal of each of the three subsets of predictors from the final (full) model. They do not give the change in R^2 from step 1 to step 2 of the hierarchical model.

Table 5 Parameter estimates for a hierarchical regression model with height entered on step 1 and three area indicators and their products with height entered on step 2

Coefficients[a]

Step	Unstandardized coefficients		Standardized coefficients	t	Sig.	95.0 % confidence interval for B	
	B	Std. error	Beta			Lower bound	Upper bound
1 (Constant)	140.491	5.844		24.039	.000	128.942	152.040
Height of father (centered on 60 in)	4.492	.605	.521	7.429	.000	3.297	5.687
2 (Constant)	130.445	10.502		12.420	.000	109.684	151.206
Height of father (centered on 60 in)	5.689	1.072	.660	5.309	.000	3.570	7.807
Area 1 indicator	11.566	15.758	.178	.734	.464	−19.584	42.717
Area 2 indicator	17.608	15.062	.346	1.169	.244	−12.167	47.383
Area 3 indicator	13.593	19.591	.189	.694	.489	−25.135	52.321
Height × A1	−1.510	1.622	−.226	−.931	.354	−4.716	1.697
Height × A2	−1.979	1.565	−.375	−1.265	.208	−5.073	1.114
Height × A3	−1.940	2.002	−.266	−.969	.334	−5.897	2.017

a. Dependent Variable: weight of father in pounds

correlations are .7381, .4453, .4698, and .6761. These are the Y_i values we will use in Eqs. 15 and 16. The variance of the sampling distribution of r' is equal to $1/(n-3)$, so the W_i values needed in Eqs. 15 and 16 are simply n_i-3 (i.e., 21, 46, 16, and 55). Plugging these W_i and Y_i values into Eq. 16 yields $\overline{Y}$ equal to .5847. Solving Eq. 15 for these data results in Q = 2.060, df = 3, p = .560. Therefore, the null hypothesis that the four population correlations are equal cannot be rejected.

Finally, we should point out that when the procedure described here is used to test the equivalence of *two* correlations, the result is identical to that obtained via the z-test for comparing two independent correlations ($z^2 = Q$). For example, when we used this procedure to compare the correlation between father's weight and mother's height for Lancaster, $r = -.181$, n = 49, p = .214, with the same correlation for Glendora, r = .330, n = 58, p = .011, we got Q = 6.927, df = 1, p = .008. Comparing these same two correlations earlier using Eq. 11, we got $z = -2.632$, p = .008.

Methods for two nonindependent parameters

In this section, we describe two standard methods for comparing two nonindependent correlations. These methods are applicable when both of the correlations to be compared have been computed using the same sample. One method is for the situation where the two correlations have a variable in common (e.g., r_{12} vs. r_{13}), and the other for the situation where there are no variables in common (e.g., r_{12} vs. r_{34}). (The first situation is sometimes described as *overlapping*, and the second as *nonoverlapping*.)

Two nonindependent correlations with a variable in common

Hotelling (1931) devised a test for comparing two nonindependent correlations that have a variable in common, but Williams (1959) came up with a better test, which is still in use today. Although Williams actually described it as an F-test, it is more commonly presented as a t-test nowadays.[18] Equation 17 shows the formula for Williams's t-test:

$$t_{n-3} = (r_{12} - r_{13})\sqrt{\frac{(n-1)(1+r_{23})}{2\left(\frac{n-1}{n-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}} \quad \text{where} \quad |R| = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} \tag{17}$$

To illustrate Williams's (1959) test, we use correlations reported in Table 1. Within each of the four

[18] Because Williams's (1959) test statistic was distributed (approximately) as F, with df = 1 and $n-3$, its square root is distributed (approximately) as t with $df = n-3$.

areas, we wish to compare r_{12} and r_{13}, with X_1 = father's height, X_2 = mother's height, and X_3 = mother's weight. Thus, the comparisons we wish to make are as follows: .164 vs. −.189 (Burbank), .198 vs. .065 (Lancaster), .412 vs. .114 (Long Beach), and .366 vs. .071 (Glendora). The r_{23} values for the four areas (i.e., the correlations between mother's height and mother's weight) are .624, .040, .487, and .364, respectively. Plugging the appropriate values into Eq. 17 yields the results listed below. The CI included in the results is a CI on $\rho_{12} - \rho_{13}$ computed using Zou's (2007) modified asymptotic method. (Zou's example 2 is included in order to confirm that our code produces his result.)

```
     r12      r13      r23        t    df        p    Lower    Upper    alpha Note

    .164    -.189     .624    2.043    21     .054    -.008     .666     .050 Burbank
    .198     .065     .040     .663    46     .511    -.257     .510     .050 Lancaster
    .412     .114     .487    1.295    16     .214    -.162     .726     .050 Long Beach
    .366     .071     .364    2.082    55     .042     .011     .564     .050 Glendora
    .396     .179     .088    1.381    63     .172    -.093     .517     .050 Zou (2007) Example 2

* The CI reported in variables Lower and Upper is the 100*(1-alpha)% CI.
* It was computed using the method described by Zou (2007) in his article
* in Psychological Methods.
```

These results indicate that the difference between the two correlated correlations is statistically significant only in area 4, Glendora, $t_{55} = 2.082$, $p = .042$. As was expected, that is also the only case in which the 95 % CI for $\rho_{12} - \rho_{13}$ does not include 0.

Two nonindependent correlations with no variables in common

Pearson and Filon (1898) devised a method for comparing two nonindependent correlations with no variables in common, but a revised version of it by Steiger (1980) yields a "theoretically better test statistic" (Raghunathan et al., 1996, p. 179). Pearson and Filon's original statistic is often called PF and is calculated as shown in Eq. 18.

$$PF = \frac{r_{12} - r_{34}}{s_{r_{12} - r_{34}}} = \frac{r_{12} - r_{34}}{\sqrt{\frac{\left(1 - r_{12}^2\right)^2 + \left(1 - r_{34}^2\right)^2 - k}{n}}}$$

where

$$k = (r_{13} - r_{23}r_{12})(r_{24} - r_{23}r_{34}) + (r_{14} - r_{13}r_{34})(r_{23} - r_{13}r_{12}) + (r_{13} - r_{14}r_{34})(r_{24} - r_{14}r_{12}) + (r_{14} - r_{12}r_{24})(r_{23} - r_{24}r_{34}) \quad (18)$$

The modified version of the Pearson–Filon statistic, which is usually called *ZPF*, can be calculated using Eq. 19. The *Z* in *ZPF* is there because this statistic is calculated using r' values (obtained via Fisher's *r*-to-*z* transformation) in the numerator[19]:

$$ZPF = \frac{r'_{12} - r'_{34}}{s_{r'_{12} - r'_{34}}} = \frac{r'_{12} - r'_{34}}{\sqrt{\left(1 - \frac{k}{2\left(1 - r_{12}^2\right)\left(1 - r_{34}^2\right)}\right)\left(\frac{2}{n-3}\right)}}. \quad (19)$$

To illustrate this method, let r_{12} = the correlation between father's height and weight and r_{34} the correlation between *mother's* height and weight and compare r_{12} and r_{34} in each of the four areas separately, but also for all of the data, collapsing across area.[20] The correlations within each area are shown in Table 1. Collapsing across area, the correlation between height and weight is .521 ($p < .001$) for fathers and .318 ($p < .001$) for mothers, with $n = 150$ for both. Plugging those values into Eq. 18 yields the results shown

[19] The reason this statistic is called *ZPF* is that Fisher used z to symbolize correlations that had been transformed using his *r*-to-*z* transformation. As was noted earlier, many current authors use r' rather than *z*, to avoid confusion with *z*-scores or *z*-test values.

[20] As was noted earlier, the lung function data file has matched pairs of fathers and mothers, which is why the correlation between height and weight for fathers is not independent of the same correlation for mothers.

below. The 100 × (1 − α)% CI shown in these results was computed using Zou's (2007) method. (Zou's third example was included to ensure that his method has been implemented correctly in our code.)

r12	r34	PF	ZPF	p_PF	p_ZPF	Lower	Upper	alpha	Note
.628	.624	.023	.022	.982	.983	-.373	.382	.050	F v M, Bur
.418	.040	2.129	2.027	.033	.043	.011	.716	.050	F v M, Lan
.438	.487	-.208	-.191	.835	.848	-.550	.452	.050	F v M, L Beach
.589	.364	1.614	1.582	.107	.114	-.054	.507	.050	F v M, Glen
.521	.316	2.255	2.236	.024	.025	.025	.384	.050	F v M, All
.396	.189	1.375	1.338	.169	.181	-.096	.501	.050	Zou Example 3

* F v M = Fathers versus Mothers.
* CI for (rho1 - rho2) computed using Zou's (2007) method.

The *PF* and *ZPF* columns show the Pearson–Filon and modified Pearson–Filon statistics, respectively, and the *p_PF* and *p_ZPF* columns show the corresponding *p*-values. Thus, the difference between the two correlated correlations is statistically significant only for the sample from Lancaster (*p* for *ZPF* = .043) and for the analysis that uses data from all four areas (*p* for *ZPF* = .025). Because alpha = .05 on all rows, all CIs are 95 % CIs.

Summary

Our goal in writing this article was twofold. First, we wished to provide, in a *single* resource, descriptions and examples of the most common procedures for statistically comparing Pearson correlations and regression coefficients from OLS models. All of these methods have been described elsewhere in the literature, but we are not aware of any single book or article that discusses all of them. In the past, therefore, researchers or students who have used these tests may have needed to track down several resources to find all of the required information. In the future, by way of contrast, they will be able to find all of the required information in this one article.

Our second goal was to provide actual *code* for carrying out the tests and computing the corresponding 100 × (1 − α) CIs, where applicable.[21] Most if not all of the books and articles that describe these tests (including our own article) present *formulae*. But more often than not, it is left to readers to translate those formulae into code. For people who are well-versed in programming, that may not present much of a challenge. However, many students and researchers are *not* well-versed in programming. Therefore, their attempts to translate formulae into code are liable to be very time consuming and error prone, particularly when they are translating some of the more complicated formulae (e.g., Eq. 17 in the present article).

Finally, we must acknowledge that resampling methods provide another means of comparing correlations and regression coefficients. For example, Beasley et al. (2007) described two bootstrap methods for testing a null hypothesis that specifies a nonzero population correlation. Such methods are particularly attractive when distribution assumptions for asymptotic methods are too severely violated or when sample sizes are small. However, such methods cannot be used if one has only summary data; they require the raw data. Fortunately, in many cases, the standard methods we present here do work quite well, particularly when the samples are not too small.

In closing, we hope that this article and the code that accompanies it will prove to be useful resources for students and researchers wishing to test hypotheses about Pearson correlations or regression coefficients from OLS models or to compute the corresponding CIs.

Acknowledgments We thank Dr. John Jamieson for suggesting that an article of this nature would be useful to researchers and students. We thank Drs. Abdelmonem A. Afifi, Virginia A. Clark, and Susanne May for allowing us to include their lung function data set with this article. And finally, we thank three anonymous reviewers for their helpful comments on an earlier draft of the manuscript.

Competing interests None of the authors have any competing interests.

[21] Although we provide code for SPSS and SAS only, users of other statistics packages may also find it useful, since there are many commonalities across packages. For example, the first author was able to translate SAS code for certain tests into SPSS syntax without difficulty, and the second author was able to translate in the opposite direction without difficulty.

References

Afifi, A. A., Clark, V., & May, S. (2003). *Computer-aided multivariate analysis* (4th Ed.). London, UK: Chapman & Hall/CRC. (ISBN-10: 1584883081; ISBN-13: 978-1584883081).

Beasley, W. H., DeShea, L., Toothaker, L. E., Mendoza, J. L., Bard, D. E., & Rodgers, J. L. (2007). Bootstrapping to test for nonzero population correlation coefficients using univariate sampling. *Psychological Methods, 12,* 414–433.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika, 37,* 256–266.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3–32.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research, 2,* 121–145.

Gould, W. (2009, July). *Why does test sometimes produce chi-squared and other times F statistics? how are the chi-squared and F distributions related? [Support FAQ]* Retrieved from http://www.stata.com/support/faqs/statistics/chi-squared-and-f-distributions/

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics, 2,* 360–378.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Wadsworth.

Kenny, D. A. (1987). *Statistics for the social and behavioral sciences*. Boston, MA: Little, Brown and Company.

Pearson, K., & Filon, L. G. N. (1898). Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Transactions of the Royal Society London (Series A), 191,* 229–311.

Potthoff, R. F. (1966). *Statistical aspects of the problem of biases in psychological tests. (Institute of Statistics Mimeo Series No. 479)*. Chapel Hill: University of North Carolina, Department of Statistics. URL: http://www.stat.ncsu.edu/information/library/mimeo.archive/ISMS_1966_479.pdf

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1,* 178–183.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society (Series B), 21,* 396–399.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12,* 399–413.