Comparing Correlations and Regressions
Notes by Alan Pickering (29.2.2004)

Introduction

Some of this material is covered in Howell, Chapter 9 (in the 5th edition, pp. 276-281). There are different situations in which one might want to compare correlations/regressions:

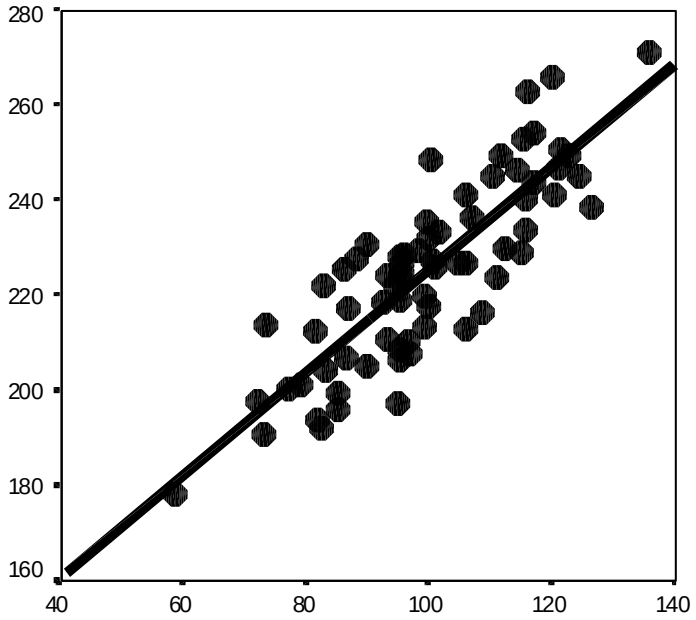
1. Comparing the correlations/regressions between variables *x* and *y* in **different groups** of subjects
2. Comparing correlations/regressions **within a single group** of subjects
   (a) Correlation/regressions between variable *j* and *k* vs. correlation between variables *j* and *h*.
   (b) Correlation/regression between variable *j* and *k* vs. correlation between variables *h* and *m*.

Note that when we have two variables (*x* and *y*) the significance test for the correlation between them gives identical results to the significance test on the regression coefficient. However, comparing two correlations is not the same as comparing two regressions. The regression coefficient is the slope of the best-fitting line relating the dependent variable to the predictor. The correlation indexes the degree of spread around that line. So it is entirely possible for two regression lines to have identical slopes (same regression) but the data to be tightly clustered around one regression line (high correlation) and significantly less tightly clustered around the second line (i.e., different correlations). Conversely, the two regression lines may have very different slopes but the size of the correlation (degree of closeness to the line) may be very similar. This is illustrated in the SPSS outputs on the following pages. These data are available in a dataset (on web and J drive) called **regs_and_corrs.sav**. There are two groups of 70 subjects (group 1 and 2), and 3 variables (*indepvar*, *depvar1*, & *depvar2*). You might try some of the analyses, explained in the notes below, using these data. In particular:

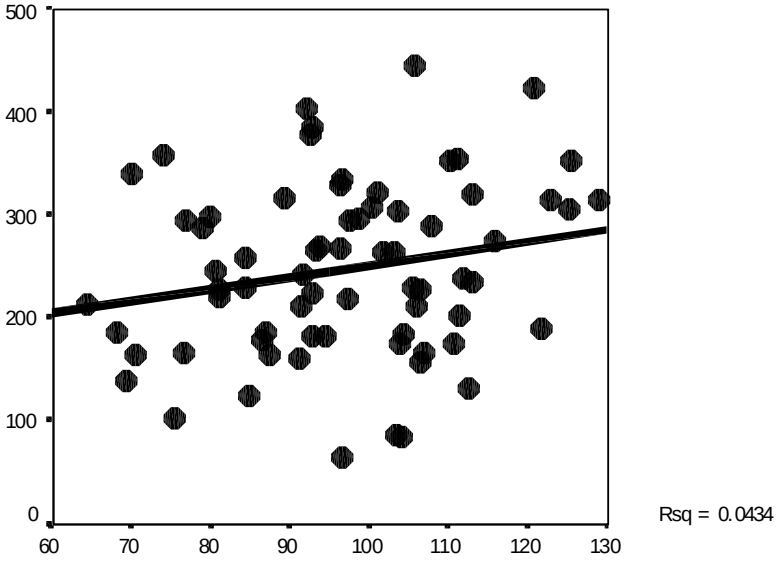
- compare the correlation between *depvar1* and *indepvar* for group 1 with that for group 2;
- compare the regression B coefficient for *indepvar* in predicting *depvar1* in group 1 with that for group 2;
- repeat the above two analyses for the relationship between *depvar2* and *indepvar*
- in group 1 only compare the correlation between *depvar1* and *indepvar* with that between *depvar2* and *indepvar*;
- repeat the above for group 2 only.

In doing the above analyses you will find it useful to employ the syntax commands including in the file **compcorr_syntax.sps** (available on the J drive and via the web). You will probably find it useful to create your own syntax for computing the Fisher Z statistic.

GROUP: 1.00



GROUP: 2.00



Note the above two relationships have very different correlations. What about the regression coefficients? The output below shows that the regression B values are very similar.

GROUP = 1.00

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	116.698	8.467		13.783	.000
	INDEPVAR	1.082	.084	.842	12.878	.000

a. Dependent Variable: DEPVAR1

b. GROUP = 1.00

GROUP = 2.00

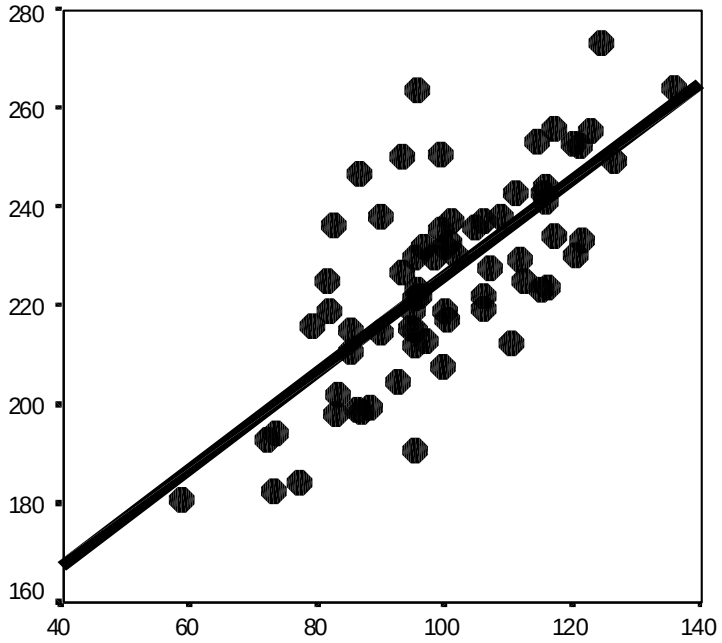
Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134.406	64.354		2.089	.040
	INDEPVAR	1.154	.657	.208	1.757	.083

a. Dependent Variable: DEPVAR1

b. GROUP = 2.00

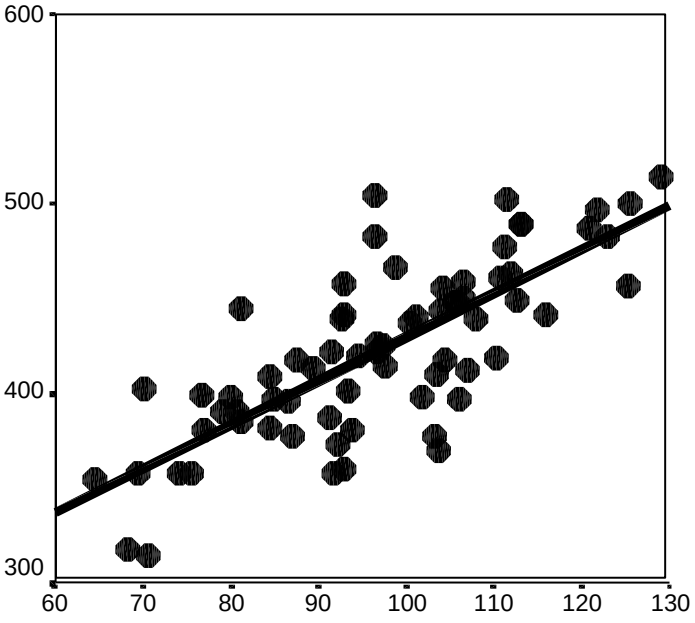
GROUP: 1.00



Rsq = 0.5245

INDEPVAR

GROUP: 2.00



DEPVAR2

Rsq = 0.5868

INDEPVAR

Note the correlations are very similar, but the output below shows that the regressions are very different (i.e., have very different B values)

GROUP = 1.00

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	128.070	11.375		11.259	.000
	INDEPVAR	.977	.113	.724	8.661	.000

a. Dependent Variable: DEPVAR2

b. GROUP = 1.00

GROUP = 2.00

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	198.191	23.056		8.596	.000
	INDEPVAR	2.312	.235	.766	9.828	.000

a. Dependent Variable: DEPVAR2

b. GROUP = 2.00

Procedures

(i) Comparing Correlations Across Independent Groups Of Subjects

2 correlation coefficients: Fisher's *Z* transformation of the correlation coefficient *r* is used. This is not the *Z* transformation we use when standardising. It is often denoted by *r'*

$$r' = 0.5*\log_e\{\text{Abs}([1 + r]/[1 - r])\}$$

where Abs(*x*) is the absolute value of *x* (i.e., ignore sign). The standard error of *r'* is denoted $s_{r'}$ and has a value of 1/sqrt(*N* - 3), where sqrt(*x*) means the square root of *x*, and *N* is the number of subjects involved in the correlation. We can calculate a z-statistic for the difference between two *r'* values:

$$Z = (r_1' - r_2') / \text{sqrt}([1/(N_1 - 3)] + 1/[N_2 - 3])$$

Howell gives a worked example. We also use Fisher's *Z* transformation to test whether a sample correlation is equal to any specific value, i.e. could the sample value we have obtained have come from a population with a correlation of a particular value. Fisher's *Z* can also be employed to put confidence limits around a sample estimate of a correlation.

2 or more correlation coefficients: There is also a simple formula for this given by Kullback (1958, pp. 321). The information statistic in favour of the hypothesis (H_1) that the correlations (between variables 1 and 2) differ across *m* samples of subjects, relative to the hypothesis (H_2) that the correlations are equal in all the samples is given by the formula:

$$2I(H_1{:}H_2) = \Sigma_{i=1 \text{ to } m} N_i \log_e [(1 - \{r_{12}\}^2)/(1 - \{r_{i12}\}^2)]$$

where r_{i12} = the correlation between variables 1 and 2 in sample *i*
N_i = the number of subjects in sample *i*
r_{12} = $\{\Sigma_{i=1 \text{ to } m} (N_i * r_{i12})\}/N$
N = $\Sigma_{i=1 \text{ to } m} N_i$

$2I(H_1{:}H_2)$ is distributed as χ^2 with $(m - 1)$ degrees of freedom. You can look up a *p*-value for χ^2 using SPSS. We will carry out an example of this in the computer class.

(ii) Comparing Regressions Across Independent Groups Of Subjects

2 regression coefficients: This is covered in Howell. Let b_1 and b_2 be the regression coefficients for the separate regressions predicting *y* using *x* in groups 1 and 2 respectively. Subscripts 1 and 2 will be used to denote values calculated in each group separately. Then one can test the difference between the regressions using a *t*-test:

$$t = (b_1 - b_2)/se(b_1 - b_2)$$

where *se*(*x*) is the standard error of *x*. The degrees of freedom for this *t*-test are:

$df = df_1 + df_2 = (N_1 - 2) + (N_2 - 2)$, for groups of size N_1 and N_2 respectively. We can look up a *p*-value for this *t*-statistic using SPSS. We will carry out an example of this in the computer class. To estimate $se(b_1 - b_2)$ we use the formula (analogous to a *t*-test formula for comparing independent group means) that:

$$se(b_1 - b_2) = s_{resid} * \text{sqrt}(1/[s_{x1}^2*(N_1 -1)] +1/[s_{x2}^2*(N_2 -1)])$$

where the variances of the predictor variable, in each group, are given by s_{x1}^2 and s_{x2}^2. The pooled residual (or error) variance, s_{resid}^2, is obtained from the residual sums of squares (SS_{resid}) for the regression for each group separately (this can be read from the SPSS regression output).

$$s_{resid}^2 = (SS_{resid1} + SS_{resid2})/(df_1 + df_2)$$

As with the familiar *t-test* for means, this formula uses a pooled error estimate, and so assumes homogeneity of error variances across the two groups.

2 or more regression coefficients: This can be done in SPSS directly using the GLM programs, using a so-called "homogeneity of regression" model. We met these models before as one needs to use them to check the "homogeneity of regression" assumption of (M)ANCOVA. Within calculation accuracy, this should give the same result (only as an *F*-ratio) as the method given above, when there are two groups.

(iii) Comparing Correlations Within A Single Sample

(a) *Compare the correlation between variables j and h (r_{jh}) vs. correlation between variables j and k (r_{jk}).* The formulae are given and explained in the paper by Steiger (1980), and given out in the lecture on a handwritten sheet. You must calculate the relevant correlations in your sample of *N* subjects: r_{jh}, r_{jk} and r_{kh}.

If you have a large sample, then you can calculate a statistic Z_1 (which can be tested against a standard normal-- Z – distribution).

Hotelling proposed a T_1 statistic which is simpler to calculate but this should NOT be used. As Steiger says: "it is virtually useless". Strangely, this statistic is recommended in several well-known textbooks on correlations.

Williams' T_2 statistic deals with the problems of Hotelling's T_1 statistic. The formula is given in Howell. T_2 has a *t*-distribution with df = N-3.

If the sample is not large, or the sample correlations are near to 0 or 1, then Z_1 does not work well. *Dunn* and *Clark* proposed a revised statistic (Z_1*) to improve matters. *Steiger* himself proposed a modification to the Dunn and Clark formula to give a new statistic (Z_1bar*). Steiger argues that T_2, Z_1* and Z_1bar* all work equivalently well with small samples.

Fortunately, these statistics are available in a special computer programme I have written: "**compcor1.exe**". All you need is the values of r_{jh}, r_{jk} and r_{hk} (from an SPSS printout). This programme is available on drive J or on the web page and will run on any pc. We will carry out an example in the computer class.

(b) *Compare the correlation between variables j and k (r_{jk}) vs. correlation between variables h and m (r_{hm}).* The formulae are given and explained in the paper by Steiger (1980). One measures the relevant correlations in your sample of N subjects: r_{jh}, r_{jk}, r_{jm}, r_{hk}, r_{hm} and r_{km}. The recommended statistics are analogous to those for reviewed for the correlations with a variable in common [(iii) (a) above]. The statistics are called Z_2 (for large samples) and Z_2* or *Z2bar** (for small samples, or samples with correlations near 0 or 1). For the exact formulae see Steiger. These statistics are not available in **compcor1.exe**. It is possible to use these statistics with multiple correlation results as well, using the extension of correlation comparisons to regressions noted in (iv) below.

(iv) Comparing Multiple Correlations Within A Single Sample

As pointed out by Tabachnick and Fidell (in their Multiple Regression chapter; Section 5.6.2.5; pp. 145-147, in the 2001 4th edition) you can extend this approach to multiple correlations. That is, one can ask whether variable *j* has a stronger regression based on one set of predictors (set *k*) compared with the regression based on another set of predictors (set *h*; no variables are common to both sets *k* and *h*). We can get the multiple correlations R_{jh} and R_{jk} from the regression printout. However, for the formulae above, we need to know R_{hk}. There is a simple trick to obtain this. The multiple correlation in a regression, *R*, is actually the same as the correlation between the dependent variable and the predicted value of the dependent value (based on the independent variables). The dependent variable is *y* and the predicted value of *y* based on the *h* set of predictors is y'_h (and is y'_k for the *k* set of predictors). Then R_{jh} is the correlation between *y* and y'_h and R_{jk} is the correlation between *y* and y'_k. The remaining correlation that we need, R_{hk}, is given by the correlation between y'_h and y'_k. So, we can use the SPSS regression options to save the predicted values from the 2 separate regressions (one based on the set *h* of predictors, the other based on the set *k*). Then we can get the necessary correlations from the regression printout plus a correlation between the saved predicted values. Finally, we can plug these values into **compcor1**, exactly as we did with simple correlations.

References

Kullback, S. (1959). *Information theory and statistics*. New York: John Wiley & Sons. (Republished in 1997 by Dover Publications, Inc).

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