Semipartial (Part) and Partial Correlation

This discussion borrows heavily from Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, by Jacob and Patricia Cohen (1975 edition; there is also an updated 2003 edition now).

Overview. Partial and semipartial correlations provide another means of assessing the relative "importance" of independent variables in determining Y. Basically, they show how much each variable uniquely contributes to R^2 over and above that which can be accounted for by the other IVs. We will use two approaches for explaining partial and semipartial correlations. The first relies primarily on formulas, while the second uses diagrams and graphics. To save paper shuffling, we will repeat the SPSS printout for our income example:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
INCOME	24.4150	9.78835	20
EDUC	12.0500	4.47772	20
JOBEXP	12.6500	5.46062	20

Correlations

		INCOME	EDUC	JOBEXP
Pearson Correlation	INCOME	1.000	.846	.268
	EDUC	.846	1.000	-.107
	JOBEXP	.268	-.107	1.000

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.919[a]	.845	.827	4.07431

a. Predictors: (Constant), JOBEXP, EDUC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1538.225	2	769.113	46.332	.000[a]
	Residual	282.200	17	16.600		
	Total	1820.425	19			

a. Predictors: (Constant), JOBEXP, EDUC

b. Dependent Variable: INCOME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-7.097	3.626		-1.957	.067	-14.748	.554					
	EDUC	1.933	.210	.884	9.209	.000	1.490	2.376	.846	.913	.879	.989	1.012
	JOBEXP	.649	.172	.362	3.772	.002	.286	1.013	.268	.675	.360	.989	1.012

a. Dependent Variable: INCOME

Approach 1: Formulas. One of the problems that arises in multiple regression is that of defining the contribution of each IV to the multiple correlation. One answer is provided by the semipartial correlation sr and its square, sr^2. (NOTE: Hayes and SPSS refer to this as the part correlation.) Partial correlations and the partial correlation squared (pr and pr^2) are also sometimes used.

Semipartial correlations. Semipartial correlations (also called part correlations) indicate the "unique" contribution of an independent variable. Specifically, the squared semipartial correlation for a variable tells us how much R^2 will decrease if that variable is removed from the regression equation. Let

H = the set of all the X (independent) variables,
G_k = the set of all the X variables *except* X_k

Some relevant formulas for the semipartial and squared semipartial correlations are then

$$sr_k == b'_k * \sqrt{1 - R^2_{X_k G_k}} = b'_k * \sqrt{Tol_k}$$

$$sr_k^2 = R^2_{YH} - R^2_{YG_k} = b'^2_k * (1 - R^2_{X_k G_k}) = b'^2_k * Tol_k$$

That is, to get X_k's unique contribution to R^2, first regress Y on all the X's. Then regress Y on all the X's except X_k. The difference between the R^2 values is the squared semipartial correlation. Or alternatively, the standardized coefficients and the Tolerances can be used to compute the semipartials and squared semipartials. Note that

- The more "tolerant" a variable is (i.e. the less highly correlated it is with the other IVs), the greater its unique contribution to R^2 will be.
- Once one variable is added or removed from an equation, all the other semipartial correlations can change. The semipartial correlations only tell you about changes to R^2 for one variable at a time.
- Semipartial correlations are used in Stepwise Regression Procedures, where the computer (rather than the analyst) decides which variables should go into the final equation. We will discuss Stepwise regression in more detail shortly. For now, we will note that, in a forward stepwise regression, the variable which would add the largest increment to R^2 (i.e. the variable which would have the largest semipartial correlation) is added next (provided it is statistically significant). In a backwards stepwise regression, the variable which would produce the smallest decrease in R^2 (i.e. the variable with the smallest semipartial correlation) is dropped next (provided it is not statistically significant.)

For computational purposes, here are some other formulas for the <u>two IV case only</u>:

$$sr_1 = \frac{r_{Y1} - r_{Y2}\, r_{12}}{\sqrt{1 - r_{12}^2}} = \frac{r_{Y1} - r_{Y2}\, r_{12}}{\sqrt{Tol_1}} = b'_1 \sqrt{1 - r_{12}^2} = b'_1 \sqrt{Tol_1}$$

$$sr_2 = \frac{r_{Y2} - r_{Y1}\, r_{12}}{\sqrt{1 - r_{12}^2}} = \frac{r_{Y2} - r_{Y1}\, r_{12}}{\sqrt{Tol_2}} = b'_2 \sqrt{1 - r_{12}^2} = b'_2 \sqrt{Tol_2}$$

For our income example,

$$sr_1 = \frac{r_{Y1} - r_{Y2}\, r_{12}}{\sqrt{1 - r_{12}^2}} = \frac{.846 - .268 * -.107}{\sqrt{1 - (-.107)^2}} = .8797 = b_{k'} \sqrt{Tol_k} = .884438 * \sqrt{.988578} = .879373,$$

$$sr_1^2 = .879373^2 = .7733 = R_{Y12}^2 - r_{Y2}^2 = .845 - .268^2 = .7732,$$

$$sr_2 = \frac{r_{Y2} - r_{Y1}\, r_{12}}{\sqrt{1 - r_{12}^2}} = \frac{.268 - .846 * -.107}{\sqrt{1 - (-.107)^2}} = .3606 = b_{2'} \sqrt{Tol_2} = .362261 * \sqrt{.988578} = .360186$$

$$sr_2^2 = .360186^2 = .1297 = R_{Y12}^2 - r_{Y1}^2 = .845 - .846^2 = .1293$$

Compare these results with the column SPSS labels "part corr." Another notational form of sr_1 used is $r_{y(1 \cdot 2)}$.

Also, referring back to our general formula, it may be useful to note that

$$R_{YH}^2 = R_{YG_k}^2 + sr_k^2,$$

$$R_{YG_k}^2 = R_{YH}^2 - sr_k^2$$

That is, when Y is regressed on all the Xs, R^2 is equal to the squared correlation of Y regressed on all the Xs except X_k plus the squared semipartial correlation for X_k; and, if we would like to know what r^2 would be if a particular variable were excluded from the equation, just subtract sr_k^2 from R_{YH}^2. For example, if we want to know what R^2 would be if X_1 were eliminated from the equation, just compute $R_{YH}^2 - sr_1^2 = .845 - .772 = .072 = R_{Y2}^2$; and, if we want to know what R^2 would be if X_2 were eliminated from the equation, compute $R_{YH}^2 - sr_2^2 = .845 - .130 = .715 = R_{Y1}^2$.

Partial Correlation Coefficients. Another kind of solution to the problem of describing each IV's participation in determining r is given by the partial correlation coefficient pr, and its square, pr^2. The squared partial r answers the question "How much of the Y variance which is not estimated by the other IVs in the equation is estimated by this variable?" The formulas are

$$pr_k = \frac{sr_k}{\sqrt{1 - R^2_{YG_k}}} = \frac{sr_k}{\sqrt{1 - R^2_{YH} + sr^2_k}}, \quad pr^2_k = \frac{sr^2_k}{1 - R^2_{YG_k}} = \frac{sr^2_k}{1 - R^2_{YH} + sr^2_k}$$

Note that, since the denominator cannot be greater than 1, partial correlations will be larger than semipartial correlations, except in the limiting case when other IVs are correlated 0 with Y in which case sr = pr.

In the two IV case, pr may be found via

$$pr_1 = \frac{sr_1}{\sqrt{1 - r^2_{Y2}}} = \frac{sr_1}{\sqrt{1 - R^2_{Y12} + sr^2_1}}, \quad pr_2 = \frac{sr_2}{\sqrt{1 - r^2_{Y1}}} = \frac{sr_2}{\sqrt{1 - R^2_{Y12} + sr^2_2}}$$

In the case of our income example,

$$pr_1 = \frac{sr_1}{\sqrt{1 - r^2_{Y2}}} = \frac{.879373}{\sqrt{1 - .268^2}} = .91276, \; pr^2_1 = .91276^2 = .83314,$$

$$pr_2 = \frac{sr_2}{\sqrt{1 - r^2_{Y1}}} = \frac{.360186}{\sqrt{1 - .846^2}} = .67554, \; pr^2_2 = .67554^2 = .45635$$

(To confirm these results, look at the column SPSS labels "partial".) These results imply that 46% of the variation in Y (income) that was left unexplained by the simple regression of Y on X1 (education) has been explained by the addition here of X2 (job experience) as an explanatory variable. Similarly, 83% of the variation in income that is left unexplained by the simple regression of Y on X2 is explained by the addition of X1 as an explanatory variable.

A frequently employed form of notation to express the partial r is $r_{Y1 \cdot 2}$ pr_k^2 is also sometimes called the partial coefficient of determination for X_k.

WARNING. In a multiple regression, the metric coefficients are sometimes referred to as the *partial regression coefficients.* These should not be confused with the *partial correlation coefficients* we are discussing here.

Alternative formulas for semipartial and partial correlations:

$$sr_k = \frac{T_k * \sqrt{1 - R^2_{YH}}}{\sqrt{N - K - 1}}$$

$$pr_k = \frac{T_k}{\sqrt{T_k^2 + (N - K - 1)}}$$

Note that the only part of the calculations that will change across X variables is the T value; therefore the X variable with the largest partial and semipartial correlations will also have the largest T value (in magnitude).

Examples:

$$sr_1 = \frac{T_1 * \sqrt{1 - R^2_{YH}}}{\sqrt{N - K - 1}} = \frac{9.209 * \sqrt{1 - .845}}{\sqrt{17}} = \frac{3.6256}{4.1231} = .879$$

$$sr_2 = \frac{T_2 * \sqrt{1 - R^2_{YH}}}{\sqrt{N - K - 1}} = \frac{3.772 * \sqrt{1 - .845}}{\sqrt{17}} = \frac{1.4850}{4.1231} = .360$$

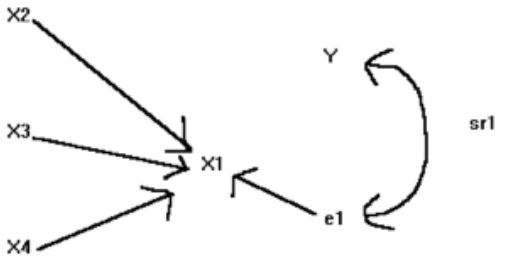
$$pr_1 = \frac{T_2}{\sqrt{T_1^2 + (N - K - 1)}} = \frac{9.209}{\sqrt{9.209^2 + 17}} = \frac{9.209}{10.0899} = .913$$

$$pr_2 = \frac{T_2}{\sqrt{T_2^2 + (N - K - 1)}} = \frac{3.772}{\sqrt{3.772^2 + 17}} = \frac{3.772}{5.5882} = .675$$

Besides making obvious how the partials and semipartials are related to T, these formulas may be useful if you want the partials and semipartials and they have not been reported, but the other information required by the formulas has been. Once I figured it out (which wasn't easy!) I used the formula for the semipartial in the `pcorr2` routine I wrote for Stata.

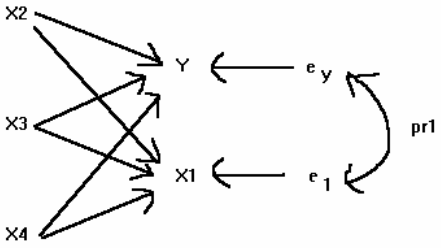
Approach 2: Diagrams and Graphics. Here is an alternative, more visually oriented discussion of what semipartial and partial correlations are and what they mean.

Following are graphic representations of semipartial and partial correlations. Assume we have independent variables X_1, X_2, X_3, and X_4, and dependent variable Y. (Assume that all variables are in standardized form, i.e. have mean 0 and variance 1.) To get the semipartial correlation of X1 with Y, regress X_1 on X_2, X_3, and X_4. The residual from this regression (i.e. the difference between the predicted value of X1 and the actual value) is e_1. The semipartial correlation, then, is the correlation between e_1 and Y. It is called a semipartial correlation because the effects of X_2, X_3, and X_4 have been removed (i.e. "partialled out") from X_1 but not from Y.



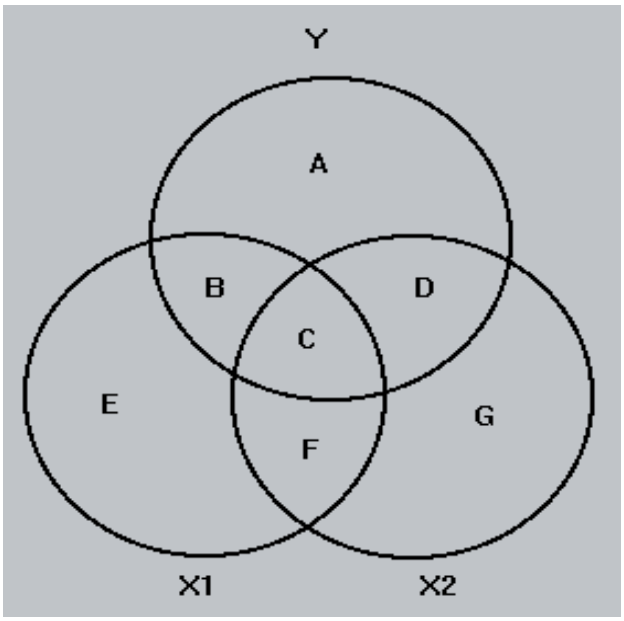
Semipartial (Part) Correlation

To get the partial correlation of X_1 with Y, regress X_1 on X_2, X_3, and X_4. The residual from this regression is again e_1. Then, regress Y on X_2, X_3, and X_4 (but NOT X_1). The residual from this regression is e_y. The partial correlation is the correlation between e_1 and e_y. It is called a partial correlation because the effects of X_2, X_3, and X_4 have been "partialled out" from both X_1 and Y.



Partial Correlation

Semipartial (Part) Correlations. To better understand the meaning of semipartial and squared semipartial correlations, it will be helpful to consider the following diagram (called a "ballantine"). [NOTE: This ballantine describes our current problem pretty well. Section 3.4 of the 1975 edition of Cohen and Cohen gives several other examples of how the Xs and Y can be interrelated, e.g. X1 and X2 might be uncorrelated with each other, or they might be negatively correlated with each other but positively correlated with Y.]



In this diagram, the variance of each variable is represented by a circle of unit area (i.e. each variable is standardized to have a variance of 1). Hence,

$A + B + C + D = s_y^2 = r_{yy} = 1$,
$(B + C)/ (A + B + C + D) = B + C = r_{Y1}^2$,
$(C + D)/ (A + B + C + D) = C + D = r_{Y2}^2$,
$(C + F)/ (B + C + E + F) = (C + F)/ (C + D + F + G) = C + F = r_{12}^2$,
$(B + C + D) / (A + B + C + D) = B + C + D = r_{Y12}^2$

That is, the overlapping of 2 circles represents their squared correlation, e.g. r_{12}^2. The total area of Y covered by the X_1 and X_2 areas represents the proportion of Y's variance accounted for by the two IVs, r_{Y12}^2. The figure shows that this area is equal to the sum of the areas designated B, C, and D. (NOTE: Don't confuse the A and B used in the diagram with the a and b we use for regression coefficients!) The areas B and D represent those portions of Y overlapped uniquely by X_1 and X_2, respectively, whereas area C represents their simultaneous overlap with Y. The "unique" areas, expressed as proportions of Y variance, are squared semipartial correlation coefficients, and each equals the increase in the squared multiple correlation which occurs when the variable is added to the other IV. Thus,

$$sr_1^2 = B = (B + C + D) - (C + D) = R_{Y12}^2 - R_{Y2}^2 ,$$

$$sr_2^2 = D = (B + C + D) - (B + C) = R_{Y12}^2 - R_{Y1}^2$$

The semipartial correlation sr_1 is the correlation between all of Y and X_1 from which X_2 has been partialled. It is a *semi*partial correlation since the effects of X_2 have been removed from X_1 but not from Y. "Removing the effect" is equivalent to subtracting from X_1 the X_1 values estimated from X_2, that is, to working with $x_1 - \hat{x}_1$ (where $\hat{x}_1$ is estimated by regressing X_1 on X_2). That is, $x_1 - \hat{x}_1$ is the *residual* obtained by regressing X_1 on X_2. We will denote this as e_1. Hence, $sr_1 = r_{ye1}$. sr_k^2 is the amount that r^2 is increased by including X_k in the multiple regression equation (or alternatively, it is the amount that r^2 would go down if X_k were eliminated from the equation.)

In terms of our diagram,

$s_y^2 = A + B + C + D = 1$, (because Y is standardized)
$r_{y1}^2 = (B + C)/ (A + B + C + D) = B + C$,
$sr_1^2 = B / (A + B + C + D) = B$.

Thus, we remove the area C from X_1 but not from Y.

Another notational form of sr_1 used is $r_{y(1\bullet2)}$, the 1•2 being a shorthand way of expressing X_1 from which X_2 has been partialled.

Partial Correlation Coefficients. Another kind of solution to the problem of describing each IV's participation in determining r is given by the partial correlation coefficient pr, and its square, pr^2. The squared partial correlation pr_1^2 may be understood best as the proportion of the variance of Y not associated with X_2 which is associated with X_1. That is,

$$pr_1^2 = \frac{B}{A + B} = \frac{(B + C + D) - (C + D)}{(A + B + C + D) - (C + D)} = \frac{R_{Y12}^2 - r_{Y2}^2}{1 - r_{Y2}^2} = \frac{sr_1^2}{1 - r_{Y2}^2}$$

$$pr_2^2 = \frac{D}{A + D} = \frac{(B + C + D)\ (B + C)}{(A + B + C + D)\ (B + C)} = \frac{R_{Y12}^2 - r_{Y1}^2}{1 - r_{Y1}^2} = \frac{sr_2^2}{1 - r_{Y1}^2}$$

More generally, we can say that

$$pr_k = \frac{sr_k}{\sqrt{1 - R_{YG_k}^2}} = \frac{sr_k}{\sqrt{1 - R_{YH}^2 + sr_k^2}}, \quad pr_k^2 = \frac{sr_k^2}{1 - R_{YG_k}^2} = \frac{sr_k^2}{1 - R_{YH}^2 + sr_k^2}$$

The numerator for pr_1^2 is the squared semipartial correlation coefficient; however, the base includes not all of the variance as in sr_1^2, but only that portion of Y variance which is not associated with X_2, that is, $1 - r_{Y2}^2$. Thus, the squared partial r answers the question "How much of the Y variance which is not estimated by the other IVs in the equation is estimated by this variable?" Note that, since the denominator cannot be greater than 1, partial correlations will be larger than semipartial correlations, except in the limiting case when other IVs are correlated 0 with Y in which case sr = pr.

Another way of viewing the partial correlation is that pr_1 is the correlation between X_1 from which X_2 has been partialled and Y from which X_2 has also been partialled (i.e., the correlation between $\hat{x}_{1\bullet 2}$ and $\hat{y}_{\bullet 2}$). A frequently employed form of notation to express the partial r is $r_{Y1\bullet 2}$, which conveys that X_2 is being partialled from both Y and X_1, in contrast to the semipartial r, which is represented as $r_{Y(1\bullet 2)}$. pr_k^2 is also sometimes called the <u>partial coefficient of determination for X_k</u>.

In terms of our diagram,

$s_y^2 = A + B + C + D = 1$, (because Y is standardized)
$r_{y1}^2 = (B + C)/ (A + B + C + D) = B + C$,
$sr_1^2 = B / (A + B + C + D) = B$
$pr_1^2 = B / (A + B)$

Thus, in the squared semipartial correlation, areas which belong to X_2 and which overlap either X_1 or Y (C and D) are removed from X_1 but not Y. In the squared partial correlation, areas which belong to X_2 are removed from both X_1 and Y.