

NOTICE. Unless otherwise indicated, all materials on this page and linked pages at the blue.temple.edu, astro.temple.edu, and rbtaylor.net addresses are the sole property of Ralph B. Taylor and © 1999-2001 by Ralph B. Taylor. All users have the right to freely access and copy these pages provided that they: acknowledge the source, do not make changes on any pages, and do not charge more than copying costs for distribution.

FI=NOT0109\_CC

NOTES ON CORRELATION AND REGRESSION:  
INTRODUCTION TO PARTIAL CORRELATION AND MULTIPLE REGRESSION

INTRODUCTION TO PARTIAL CORRELATION

Hamilton covers this material in Chapter 3.

You've already been there

You have already been introduced to the topic of partial correlation, and partial correlation.

A couple of weeks ago in lab you were looking at a plot of the residuals of VIOLRA90 after you have predicted VIOLRA90 with VIOLRA85.

These residuals represented the portion of the 1990 violent crime rate in each state that could not be predicted from 1985 violent crime rates.

As some of you remarked, looking at the flat regression line that resulted when you tried to predict the VIOLRA90 residuals with VIOLRA85 the correlation was zero. There was no relationship between VIOLRA90-RESID and VIOLRA85.

In other words, any relationship between VIOLRA85 and VIOLRA90-RESID had been removed by the regression operation. The scatter you were observing around the flat regression line is that **left over** portion completely unrelated to the predictor.

So we can say things like:

- We are looking at the portion of VIOLRA90 that **has been partialled** with respect to VIOLRA85
- The variation we see on the Y axis (VIOLRA90-RESID) is the variation observed **after we have controlled** for the impact of VIOLRA85 on VIOLRA90

Now suppose you had this question. You wondered what the impact would be of a low high school graduation rate, circa 1984, on subsequent changes in the violent crime rate. You hypothesized that if a state has a low HS graduation rate a few years later the violent crime rate will go up.

In other words, what you are asking is:

What are the impacts of HSRATE82 on VIOLRA90 after controlling for impacts of VIOLRA85 on VIOLRA90

If you do a simple regression with VIOLRA90-RESID as your dependent variable you can answer this question. But there is another way you can do this as well. *You can enter both predictors at the same time.* The next example gives you a hypothetical scenario where we could do this.

### Scenario

Imagine the following situation. (*This is a hypothetical scenario, and the data are fictitious*). You are involved in a courtwatchers program, trying to get more people put away for a longer time, and generally seeing to it that victims rights and concerns are addressed adequately in sentencing. Since you are working in a cooperative relationship with the DA's office, you have also had access to PSI reports.

Over the last week when you sat in court you saw 50 convicted felons -- all of whom were convicted of armed robbery -- receive sentences of incarceration. The sentences ranged from 3 months to 60 months. (Notice the instant offense is the same for all these offenders.)

From the PSI you know two things about each of these offenders. First, you know, for each, how many prior convictions they had, as an adult.

In addition, you know the degree to which the victim was harmed in the crime.

- 1 = victim not harmed at all
- 2 = slight injury, no hospitalization required
- 3 = hospitalization required, less than 2 days
- 4 = hospitalization required, more than 2 days

Given this information about the 50 cases, you go home and compute the correlation between each pair of variables. Your correlations are:

- X1 = number of prior convictions
- X2 = degree of harm
- Y1 = length of incarceration in months

	X1	X2	Y1
X1	--		
X2	.50	--	
Y1	.30	.25	--

OF COURSE you have looked carefully at your three bivariate scattergrams and convinced yourself that there are no problems with any of the key assumptions of regression; everything is ok in terms of homoscedasticity, average residual = 0 **at different ranges of X**, residuals not being correlated with X values, and so on.

BRAINBUSTER QUESTION: how about the assumption of random sampling?

As a courtwatcher, you suspect that the court is not taking into account victim harm, to the extent it should. You realize that X1 and X2 are correlated. Felons who are career criminals do more victim bashing.

Since X1 and X2 are correlated the correlation between X2 and Y does not reflect the unique influence of X2 on Y.

And, what you would really like to know is: what is the unique contribution of victim harm to length of incarceration.

This is a case for: PARTIAL CORRELATION

(Hamilton (p. 69, 70) calls these partial effects.)

You might also want to know: putting aside offender history, every unit increase in victim harm results in how many more months added on to the felon's sentence? In other words, is the judge taking enough account of victim harm; is she weighing the harm appropriately in her sentencing?

This is a case for PARTIAL REGRESSION SLOPE

You might also want to know, pondering over your data late at night with your Bud Lite: does the judge's sentencing make sense? Is she being arbitrary? To what extent do her sentence lengths assigned reflect these very important aspects of the offender and the crime itself? Or, is the judge being capricious, and if so, what is the extent of her capriciousness?

This is a case for the MULTIPLE CORRELATION

\*\*\*

#### GENERAL OVERVIEW ON LOGIC OF PARTIALS AND MULTIPLE R

There are three general purposes to getting into bizarre realms that take us beyond the bivariate (two variable) case.

1. CONTROL FOR EXTRANEOUS VARIABLE. You might want to get rid of the influence of another variable because you know it is influencing the variable of interest.

You might have many reasons for wanting to do this. You might say, for example, does the relationship hold across all levels of the control variable.

We are literally going to subtract out the influence of the control variable, we're going to wash it away.

In the language of the above scenario: what is the relationship between victim harm and sentence length when I control for the extent to which the offender is a career criminal?

NOTE: you can have, if you wish, as many control variables as you wish. In the scenario we are working with here, you can add: race, history of drug problem, history of juvenile offending, and so on. The number is limited only by your theoretical imagination, your data collecting abilities, and the sturdiness of your computing. BUT: it is important, before you go adding control variables willy-nilly, to think through THEORETICALLY why or how they might influence things.

ALSO: as we will see, introducing additional predictors can sometimes create problems. And, it gets us away from simple, parsimonious models (Hamilton, p. 72).

Good control variables are:

- correlated with the outcome
- related to the other predictors
- theoretically relevant

2. ISOLATE EXACT ROLE OF KEY PREDICTOR. You are fixated on one predictor, let's say victim harm, and you really want to know the specific influence of that variable on the outcome. That is, if I 'throw away' the portion of victim harm contaminated by or related to offender history, I have some measure of how much victim harm, by itself, contributed to the outcome.

The simple correlation above -- or zero order correlation -- (.25) doesn't tell me. Why? Because harm in the instant offense is related to offender history -- number of previous convictions ( $r = .50$ ).

I somehow have to look at the portion of harm's influence on the outcome -- sentence length -- that is independent of the portion of harm linked to conviction history.

I want to know its partial slope on the outcome, and its partial correlation with the outcome. (Darlington also talks about semipartial correlation.)

3. USE MORE PREDICTORS TO EXPLAIN MORE OF THE OUTCOME. Sometimes more is more. You might say to yourself: if I have two variables, and they are both correlated with an outcome, altogether, or taken in toto, how much of the outcome do they account for or 'explain'?

It makes sense to expect that taken together two variables will explain more of the outcome than either variable alone.

In a practical sense, with more variables, your prediction or explanation will improve; the amount of residual variance will decrease. So: this is another case for going into multiple correlation. To see if you can explain more of your Y variable than with just one predictor.

NOTE: you cannot simply add up your zero order r's to get the multiple r. This is because X1 and X2 are related to one another. In other word, to some extent, X1 and X2 are trying to predict the same portions of Y variance. You need to take this into account.

NOTE: If X1 and X2 are PERFECTLY UNCORRELATED you could add up the simple correlations to get the multiple correlation.

#### SIMPLE REGRESSION IS JUST A SPECIAL CASE OF MULTIPLE REGRESSION

Hark back to those days of yesteryear when we were discussing simple regression. No doubt you recall that familiar equation:

(Eq. 1)  $y = a + b_1X_1 + e$

Now just imagine for a moment that there were numerous other predictors we could have used in this regression. We might have even had them in our data set, but we chose not to use them. In other words, we decide beforehand that other variables we may or may not have had at hand had no influence on the outcome. We eliminated them from consideration.

In other words, Eq. 1 is just a simple and very special case of the more general equation:

(Eq. 2)  $y = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n + e$

where you have simplified by setting

$$b_2 = 0$$

$$b_3 = 0$$

.

.

.

$$b_n = 0$$

In other words, there are all these other potential predictors -- ranging up to a total of n predictors if you like -- whose potential influence on Y you have set to 0 by forcing their slopes to be zero.

*In the cases of X2 to Xn you have forced their slopes to be 0 simply by not including them in the regression in the first place!*

Imagine the above courtwatcher case where you only looked at victim harm as a predictor, and didn't consider any other factors in the case.

Please note thus that your slope b1 has a meaning much more specific than we noted in Eq. 1.

What b1 really equals is:

*the effect of X1 on Y when other predictors, X2 through Xn are forced to have no influence on Y, that is, when the slopes of X2 through Xn are forced to 0.*

Likewise for a; it is interpreted in the same restricted manner.

Now comes the fun. Suppose you do away with the restrictions. You decide you are no longer going to force the constants b<sub>2</sub> through b<sub>n</sub> to be set to 0. YOU WILL LET THE DATA DECIDE.

Note that this changes the interpretation of the slope, b.

b<sub>n</sub> is now a partial slope, based on the partial correlation between X<sub>n</sub> and Y.

#### REGRESSION LINE BECOMES A REGRESSION PLANE

Consider the equation:

(Eq. 3)  $Y = a + b_1X_1 + b_2X_2 + e$

and the regression line defined by:

(Eq. 4)  $Y_{\text{predicted}} = a + b_1X_1 + b_2X_2$

We can still use Eq. 4 to compute points on a regression line, but that line has become a plane. (With more than 2 predictors it becomes a hyperplane. See Blalock pp. 452 - 453 for more details)

In other words, for every possible combination of X<sub>1</sub> and X<sub>2</sub> values we can compute a Y-predicted value

You can also compute variances of the y predicted scores, the y residual scores, and so on. (The residuals capture distance in two dimensions between the regression plane and the data point.)

You can also plot 2 regression lines; the line showing the slope of X<sub>1</sub> on Y, and the line showing the slope of X<sub>2</sub> on Y.

But, each slope is now a partial slope if X<sub>1</sub> and X<sub>2</sub> are correlated with each other.

### WHAT IS A PARTIAL SLOPE?

"A partial slope can be interpreted as the hypothetical change that would occur in the dependent variable if one of the independent variables were to change by one unit and if the other independent variables were to remain constant" (Blalock, p. 479)

In the scenario we started the handout with: imagine there are several offenders with previous convictions = 2.

You could do an analysis just on this sample and find out what the effects of victim harm on sentence length are, for offenders where previous convictions = 2; then the slope of sentence length on harm would be a partial slope, applicable for offenders with 2 previous convictions.

This is one way of doing it.

But note: it would apply only to cases where previous convictions = 2. You would have to act similarly, for each group of offenders who had a certain number of previous convictions. This would soon get tedious. Also, you might run into the problem where you only have one offender at one level of previous convictions.

There's also a second problem with this approach. You end up with a bunch of partial slopes rather than just one. You want just ONE partial slope for harm on sentence length.

### HOW TO GET IT

Once you know the slope of every variable on every other one you can get your partial slope, and your new a constant. Darlington talks about this in chapter 2. See also Blalock's Eq. 19.9.

$$(Eq. 5) \quad b_{ij.k} = \frac{b_{ij} - ((b_{ik}) * (b_{kj}))}{1 - b_{jk} * b_{kj}}$$

where:

I = dependent variable

j = predictor of interest

k = predictor 'controlled' (shown by .k)

Let's try and unpack equation 5 conceptually.

You start in the numerator with the 'raw' slope of the predictor variable on the outcome (b<sub>ij</sub>).

Now note what we take away from that. We first consider how much the control variable influences the outcome (bik). Then we multiply this term by the slope of the predictor on the control variable. So what you're taking away is ((how much the control correlates with the outcome) \* (how much the predictor determines the control variable)).

Now on to the denominator. What you are doing is controlling for how much j and k are intercorrelated.

(Eq. 6) 
$$r_{jk}^2 = b_{jk} * b_{kj}$$

So the denominator reduces to, in the special two predictor case:

$$(1 - r_{jk}^2)$$

BRAINBUSTER QUIZ: if j and k are completely uncorrelated, what does Eq. 6 reduce to?

BRAINBUSTER QUIZ: If j and k correlate perfectly with each other, what does Eq. 6 reduce to?

So, once you have your raw or zero-order correlations, and your raw b weights, or regression coefficients, you can calculate partial slopes immediately.

Note that these are unstandardized, so your interpretation is in terms of raw units of the I variable and the j variable.

BRAINBUSTER QUIZ: what would the formula be if k was the predictor of interest, and j was the control variable?

#### GETTING PARTIAL SLOPES OR STANDARDIZED REGRESSION COEFFICIENTS

We can also derive standardized partial slopes or standardized partial regression coefficients, or beta weights.

You can get these from your b weights, by 'controlling' for the scale of the variables, as in the equation (Blalock 19.13):

(Eq. 7) 
$$\text{beta}(ij.k) = b(ij.k) * \frac{\text{sd}(j)}{\text{sd}(i)}$$

Or, you can get your betas by working directly with the correlations (Blalock 19.15):

(Eq. 8) 
$$\text{betaij.k} = \frac{r_{ij} - (r_{ik} * r_{jk})}{1 - r_{jk}^2}$$

NOTE: you can use Eq. 7 and 8 to cross check yourselves, as I'm sure you will want to do.

REMEMBER how the interpretation of the standardized slopes is different from the interpretation of the

unstandardized slopes.

BRAINBUSTER QU: Use Eq. 8 to answer the burning courtwatcher's question about the effects of victim harm on sentence length, controlling for felon conviction history.

### THINKING ABOUT PARTIAL CORRELATION

Let's talk a bit more about partial correlations. WE CAN DEFINE THEM LOTS OF DIFFERENT WAYS.

1. They represent "a single measure summarizing the degree of relationship between two variables controlling for a third." (Blalock, 457-458). (Or a third and a fourth. Or a third and a fourth and a fifth, and so on.)

2. "The partial correlation between Y and X<sub>1</sub> controlling for X<sub>2</sub>, can be defined as the correlation between the residuals of the regressions of Y on X<sub>2</sub>, and X<sub>1</sub> on X<sub>2</sub>." (Blalock, p. 457)

To state 2. another way, regress the control variable on the outcome, and then figure out the portion of the predictor that is uncorrelated with the outcome, and then determine the portion of that which will be correlated with the residuals left after using the control to predict the outcome.

The Cohen & Cohen ballantine approach gives you another way of thinking about partials and semipartial. THEIR PERSPECTIVE IS ALSO VERY HELPFUL.

If you have 1 control variable (k) then the partial correlation is referred to as a first order partial correlation.

If you have two control variables, then the partial is a second order partial.

And so on.

### GETTING PARTIAL CORRELATIONS

If:

- i = dependent variable
- j = predictor of interest
- k = predictor 'controlled'

then the formula for getting the first order partial can be expressed as (Blalock 19.3):

(Eq. 9)

$$r_{ij.k} = \frac{r_{ij} - (r_{ik} * r_{jk})}{((Sq. Rt. (1 - r_{ik}^2)) * ((Sq. Rt. (1 - r_{jk}^2)))}$$

Note the similarities between Eq. 9 and Eq. 5.

Compare and contrast Eq. 9 with Eq. 8.

Conceptually, do you see what the pieces are in Eq. 9?

---

BRAINBUSTER QUIZ: Looking at Eq. 9,

\* what happens as the correlation between the control variable and the outcome variable approaches 0 ? What parts of equation are being influenced?

\* what happens as the correlation between the predictor and the control variable approaches zero? What parts of equation are being influenced?

\* as the correlation between j and k approaches 1?

---

### THE MULTIPLE R

The multiple correlation coefficient - R - is a measure of the goodness of fit of the least squares plane or hyperplane, to the data at hand. It expresses the correlation between Y and  $Y_{pred}$ .

With one predictor, we tried to get a line to fit the data points.

With two predictors, we try to get a plane - a board - to fit the data points which are plotted in three dimensions

With three predictors, we try to get a hyperplane to fit the data, which now exists in more than three dimensions.

On the plane or hyperplane are predicted scores. We can calculate the variance of these predicted scores.

Despite Twilight Zone overtones, in each case we can figure out what the distance is between the actual data points, and the regression plane or hyperplane. We calculate the discrepancy between each data point and each regression line (work in one dimension at a time), add up, and square the total distances. Adding up across data points and dividing by N we can get the variance and SD of these residuals.

The square of the multiple correlation ( $R^2$ ) will be equal to the percentage of the variation in the Y variable explained by all the independent variables.

In other words, it still holds that:

$$(Eq. 10) \quad R^2 = \frac{\text{variance explained}}{\text{total variance of Y}}$$

And:

$$(Eq. 11) \quad (1 - R^2) = \frac{\text{variance NOT explained}}{\text{total variance}}$$

And, of course:

variance predicted scores = variance explained

variance residuals = variance NOT explained

So: one way to get your R<sup>2</sup> is to compute your predicted scores, using b1, b, and a, and then compute the variance of those scores, the variance of the original Y variable, and plug into Eq. 10. **Alternatively, SPSS calculates these for us directly, and we can save them to the file.**

Or: once you have predicted scores, compute residuals (y - ypredicted), then determine their variance, and use Eq. 11. **Same point as above; we can save to file with SPSS** and then get their statistics.

A second way is to use the correlations. With two predictors (Blalock, 19.16 and 19.17):

$$(Eq. 12) \quad R^2_{y.12} = \begin{array}{l} \text{proportion} \\ \text{explained} \\ \text{by} \\ \text{X1} \end{array} + \begin{array}{l} \text{additional} \\ \text{portion} \\ \text{explained by} \\ \text{X2} \end{array} * \begin{array}{l} \text{proportion} \\ \text{unexplained by} \\ \text{X1} \end{array}$$

NOTE: in above equation do multiplication first.

In other words:

$$(Eq. 13) \quad R^2_{y.12} = r^2_{y1} + (r^2_{y2.1} * (1 - r^2_{y1}))$$

NOTE: if you want you can re-arrange the above equation and solve for r<sub>2y2.1</sub>. This gives you a second way to check your partial correlation.

Note, that multiple R is strictly defined as: Sq. Root R<sup>2</sup>. Therefore, since R<sup>2</sup> is working off of squared correlations, R can never ever be negative. Sometimes, however, adjusted R<sup>2</sup> can be negative though...

A third way to get the multiple R is to work only in terms of the raw correlations (Blalock, Eq. 19.20).

(Eq. 14)

$$R^2_{i.jk} = \frac{r^2_{ij} + r^2_{ik} - (2 * r_{ij} * r_{ik} * r_{jk})}{1 - r^2_{jk}}$$