Research Design - -Topic 10

Multiple Regression and Multiple Correlation



1. Overview, General Rationale, and Applications

2. The case of two predictors

   Basic Equations
   Partial and Semi-partial (Part) correlation
   Multiple correlation and correlations between predictors

3. The case of many predictors

   Tests of Significance
   Relation of $R^2$ to semi-partial correlations
   Multiple correlation and $R^2$ (proportion of variance)
   Expected value of $R^2$ and Shrinkage
   Venn Diagrams and Dimensionality
   Example
   Running SPSS Regression
   Interpretation

1

## Applications and Implications

It is often said that multiple correlation can be used to identify good predictors. This is not the case. Multiple correlation does not identify predictors of a criterion. It identifies variables that add to prediction. There is a difference. Note that:

***The Pearson product moment correlation between a variable and the criterion can be considered a measure of prediction. The correlation coefficient is the regression coefficient in standard score form.***

***The regression coefficient in multiple regression is a measure of the extent to which a variable adds to the prediction of a criterion, given the other variables in the equation. It is not a correlation coefficient.***

2

Multiple Correlation was introduced by Yule (1897) as an extension of bivariate regression to assess linear relations involving a number of independent variables. The intent was to improve prediction over the bivariate case.

Since then, there have been many applications, including:

1. Establishment of a prediction equation
2. Selection of a subset of "predictors"
3. Analysis of variance
4. Curve fitting
5. Assessing mediation
6. Assessing moderation
7. Path analysis

3

Multiple regression is an equation linking a criterion variable (X) to a set of other variables. For example, one might wish to predict grades in a subject (the criterion) with a number of other variables such as GRE-Verbal, GRE-Quantitative, and Height.

The general form of the regression equation in raw score form is:

$$X' = b_0 + b_1V_1 + b_2V_2 + \ldots + b_kV_k$$

In standard score form, the equation is:

$$Z'_X = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_k Z_k$$

4

Multiple Correlation is the Pearson product moment correlation of the obtained and predicted values of X.

$$R = \frac{\sum (X - \overline{X})(X^{'} - \overline{X}^{'})}{nS_X S_{X'}}$$

And with a bit of algebra: $R = \sqrt{\beta_1 r_{1X} + \beta_2 r_{2X} + ... + \beta_k r_{kX}}$

(i.e., the multiple correlation is equal to the square root of the sum of the product of the standardized regression coefficient for each predictor times its correlation with the criterion.)

And that: $\beta_k = \frac{b_k S_k}{S_X}$

(i.e., the standardized regression coefficient is equal to the unstandardized regression coefficient times the standard deviation of the predictor divided by the standard deviation of the criterion.)

5

**Basic Equations**:

Raw score form $X_1^{'} = b_0 + b_2 X_2 + b_3 X_3 + ... + b_k X_k$

Standard score form $Z_1^{'} = \beta_2 Z_2 + \beta_3 Z_3 + ... + \beta_k Z_k$

The square of the multiple correlation is equal to the variance of the predicted Z scores such that:

$$R^2 = \frac{\sum Z_1^{'2}}{N} = \frac{\sum (\beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4)^2}{N}$$

$$= \beta_2^2 + \beta_3^2 + \beta_4^2 + 2\beta_2\beta_3 r_{23} + 2\beta_2\beta_4 r_{24} + 2\beta_3\beta_4 r_{34}$$

i.e., the square of the multiple correlation is equal to the sum of the squared standardized regression coefficients plus two times the product of the correlation between each pair of predictors times their regression coefficients.

6

**Matrix equations**. Matrix notation is often used with multiple regression and correlation. The following examples consider the use of 3 predictors.

The squared multiple correlation is written as:

$$R^2_{1.234} = \beta_2 r_{12} + \beta_3 r_{13} + \beta_4 r_{14}$$

which can be expressed as the product of two vectors as:

$$= \begin{bmatrix} r_{12} & r_{13} & r_{14} \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = R_{1j}\beta_j$$

where $\beta_j$ is defined as: $\beta_j = R_{jj}^{-1} R_{j1}$

i.e., the product of the inverse of the matrix of correlations of the predictors with the vector of correlations of the criterion with the predictors. That is:

$$R_{jj} R_{jj}^{-1} = I$$

Thus, in matrix terms: $R^2_{1.234} = R_{1j} R_{jj}^{-1} R_{j1}$

7

The Case of Two Predictors

The regression equation describes a plot in three dimensional space as indicated on the next slide. The plot shows the model in raw score form based on the regression equation as follows:

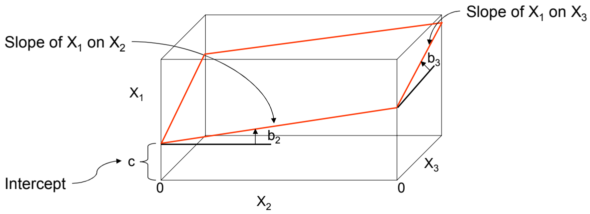
$$X_1^{'} = C + b_2 X_2 + b_3 X_3$$

where $b_2 = \frac{S_{X_1}\beta_2}{S_{X_2}}$ and $b_3 = \frac{S_{X_1}\beta_3}{S_{X_3}}$

and $C = \overline{X}_1 - b_2 S_{X_2} - b_3 S_{X_3}$

8

## A two predictor illustration of the model

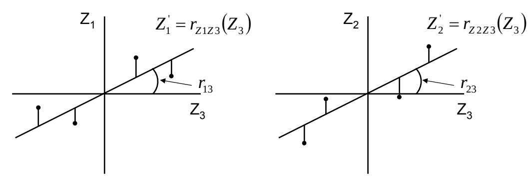



In the diagram, X2 and X3 are shown to be orthogonal (i.e., independent of each other), but generally the predictors are correlated. Thus, to ensure independence, we calculate the regression coefficients on residualized variables. This involves the constructs of partial and semipartial correlation.

9

**1. <u>Partial Correlation</u> – Plots in Standard Score Form**

$$r_{12.3} = \frac{\sum(Z_1 - Z_1^{'})(Z_2 - Z_2^{'})}{nS_{Z_1 - Z_1^{'}}S_{Z_2 - Z_2^{'}}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$




**2. <u>Semipartial (part) Correlation</u>**

$$r_{1(2.3)} = \frac{\sum Z_1(Z_2 - Z_2^{'})}{nS_{Z_2 - Z_2^{'}}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{23}^2}}$$

10

## The regression equations in standard score form

$$Z_{X_1}^{'} = \beta_2 Z_{X_2} + \beta_3 Z_{X_3}$$

Where: $$\beta_2 = \frac{r_{12} - r_{13}r_{23}}{1-r_{23}^2} = \frac{r_{1(2.3)}}{\sqrt{1-r_{23}^2}}$$

$$\beta_3 = \frac{r_{13} - r_{12}r_{23}}{1-r_{23}^2} = \frac{r_{1(3.2)}}{\sqrt{1-r_{23}^2}}$$

Thus: Beta coefficients can be shown to equal the semipartial correlation of the criterion with a predictor divided by the standard error of estimate of that predictor in standard score form as predicted by the other predictor.

Note that: $R_{1.23} = \sqrt{\beta_2 r_{12} + \beta_3 r_{13}}$

11

## Relation of Multiple Correlation to Relations Among Predictors

Other things being equal, it can be shown that the the multiple correlation decreases as the correlation between predictors increases. Consider the case where

$r_{12} = .6 \qquad r_{13} = .5$

It can be shown that:

$$r_{23} = r_{12}r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2}$$

Thus: $r_{23} = .30 \pm \sqrt{.48} \qquad \therefore -.393 \le r_{23} \le .993$

Thus, we can consider the values of $\beta_2$, $\beta_3$, and $R_{1.23}$ when $r_{23}$ varies from -.30 to .90. Applying the formulae would produce the following answers

| $r_{23}$ | $\beta_2$ | $\beta_3$ | $R_{1.23}$ | $r_{Z1(Z_2+Z_3)}$ |
|---|---|---|---|---|
| -.30 | .824 | .747 | .932 | .930 |
| -.20 | .729 | .646 | .872 | .870 |
| .00 | .600 | .500 | .781 | .781 |
| .20 | .521 | .396 | .715 | .710 |
| .40 | .476 | .310 | .664 | .657 |
| .60 | .469 | .219 | .625 | .615 |
| .80 | .556 | .056 | .601 | .580 |
| .90 | .789 | -.211 | .607 | .564 |

Correlation of Z1 with Z2 + Z3 (see Topic 15)

12

## The Case of Many Predictors

Multiple correlation and multiple regression can be conducted with any number of predictors though it is wise to keep the number manageable. It is generally recognized that the dependent variable should be continuous (and normal) but predictors can be both continuous and categorical (while distributional characteristics will influence the results). The following discussion will focus primarily on three predictors though the generalizations apply to any number of predictors.

Multiple regression can be performed by entering all of the predictors of interest in one step or by using a hierarchical method in which the researcher enters the predictors in some predetermined manner either one at a time or in groups. There are also a number of indirect methods where the computer enters the predictors in a manner determined by the data. Some of these are discussed below.

13

## Relation of R² to semipartial correlations

$$R^2_{1.234} = \beta_2 r_{12} + \beta_3 r_{13} + \beta_4 r_{14}$$
$$= r^2_{12} + r^2_{1(3.2)} + r^2_{1(4.32)}$$

It can be shown that:

$$Semipartial\ correlation^2 = r^2_{1(4.23)} = \frac{t^2_{b_4}(1 - R^2_{1.234})}{N - 3 - 1}$$

and

$$Semipartial\ correlation^2 = r^2_{1(3.2)} = \frac{t^2_{b_3}(1 - R^2_{1.23})}{N - 2 - 1}$$

or

$$Partial\ correlation^2 = r^2_{14.23} = \frac{r^2_{1(4.23)}}{1 - R^2_{4.23}}$$

Etc…

14

The test of significance of the multiple R is:

$$F = \frac{R^2 / p}{(1 - R^2)/(N - p - 1)}$$ With df: v1 = p, v2 = N-p-1

The test of significance of the increase in the Multiple R when adding variables to an existing regression equation is:

$$F = \frac{\left(R^2_2 - R^2_1\right)/(p_2 - p_1)}{(1 - R^2_2)/(N - p_2 - 1)}$$ With df: $v1 = p_2 - p_1$, $v2 = N - p_2 - 1$.

Where:

$R^2_2$ = R² with $p_2$ predictors    $R^2_1$ = R² with $p_1$ predictors

$$p_2 > p_1$$
$$N = total\ number\ of\ subjects$$

15

## Tests of Significance of the Regression Coefficients

Test of significance of b:

$$t_b = \frac{b}{SE_b} \quad @\ df = N - p - 1$$

It will be recalled that this is the square root of the F for R² change when the predictor is added to the equation.

Test of significance of β

$$t = \frac{\beta_j}{SE_{\beta j}} = \frac{\beta_j}{\sqrt{\frac{1 - R^2_{Y.1,2,..k}}{N - k - 1}}\sqrt{\frac{1}{1 - R^2_{j.1,2,..k}(omitting\ j)}}}$$

Where: $df = N - p - 1$

Both tests will yield the same value of t.

16

The Wherry (1931) adjustment is based on defining the mean square for the residual as the sum of squares divided by the degrees of freedom rather than N-1. This adjustment is used in SPSS and is considered an unbiased estimate of the population value.

$$R^2_{adjusted} = 1-(1-R^2)\frac{N-1}{N-p-1}$$

Given $R^2 = .50$, N=50, p = 10 **$R^2_{adjusted} = .37$**

The Stein (1960) (correlation model) adjustment is an estimate of the expected value of a series of cross validation samples where the predictors are considered random (i.e., the X's can take any value).

$$R^2_{adjusted} = 1-\left[\frac{N-1}{N-p-1}\right]\left[\frac{N-2}{N-p-2}\right]\left[\frac{N+1}{N}\right]\left[1-R^2\right]$$

Given $R^2=.50$, N=50, p=10 **$R^2_{adjusted} = .19$**

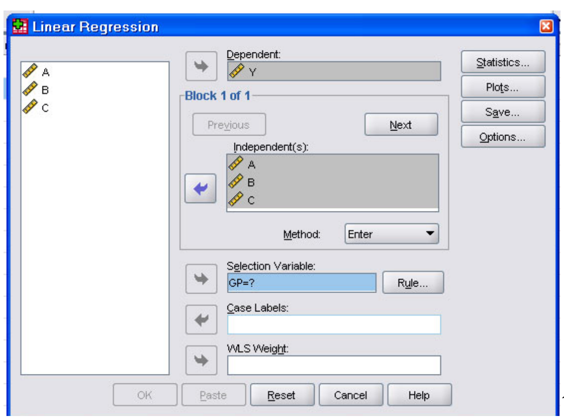
17

## Data Used in the Next Example

| A | B | C | Y |
|---|---|---|---|
| 24 | 29 | 21 | 49 |
| 24 | 33 | 23 | 54 |
| 27 | 30 | 23 | 56 |
| 26 | 26 | 19 | 45 |
| 23 | 32 | 20 | 48 |
| 25 | 27 | 19 | 39 |
| 26 | 30 | 20 | 50 |
| 30 | 33 | 24 | 55 |
| 29 | 32 | 20 | 55 |
| 25 | 30 | 20 | 50 |
| 21 | 27 | 18 | 39 |
| 24 | 29 | 21 | 52 |
| 28 | 30 | 21 | 56 |
| 23 | 30 | 22 | 52 |
| 23 | 30 | 20 | 50 |
| 26 | 30 | 20 | 51 |
| 23 | 29 | 19 | 43 |
| 24 | 31 | 18 | 48 |
| 23 | 31 | 21 | 51 |
| 27 | 31 | 21 | 57 |

18

## Windows Setup




19

**Correlations**

| | | y | a | b | c |
|---|---|---|---|---|---|
| Pearson Correlation | y | 1.000 | .601 | .706 | .738 |
| | a | .601 | 1.000 | .331 | .434 |
| | b | .706 | .331 | 1.000 | .580 |
| | c | .738 | .434 | .580 | 1.000 |
| Sig. (1-tailed) | y | . | .003 | .000 | .000 |
| | a | .003 | . | .077 | .028 |
| | b | .000 | .077 | . | .004 |
| | c | .000 | .028 | .004 | . |
| N | y | 20 | 20 | 20 | 20 |
| | a | 20 | 20 | 20 | 20 |
| | b | 20 | 20 | 20 | 20 |
| | c | 20 | 20 | 20 | 20 |

The multiple correlation obtained with all three predictors is:

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | 0.85864[a] | 0.73727 | 0.68801 | 2.92773 |

a. Predictors: (Constant), c, a, b

20

Often, when conducting these analyses, researchers will report the multiple correlation, and the regression coefficients.

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | -25.390 | 11.647 | | -2.180 | 0.045 |
| | a | .703 | 0.325 | 0.309 | 2.161 | 0.046 |
| | b | 1.073 | 0.445 | 0.382 | 2.412 | 0.028 |
| | c | 1.248 | 0.541 | 0.382 | 2.306 | 0.035 |

a. Dependent Variable: y

In the present example, all three regression coefficients are significant. This does not mean they are significant predictors (this information is contained in the correlation matrix). It means only that given the criterion and these three variables each one adds significantly to prediction. Add or subtract one or more predictors or change the dependent variable and the results can change drastically. This is because the multiple correlation is made up of more than the information given in the regression coefficients.

21

## On the interpretation of regression coefficients

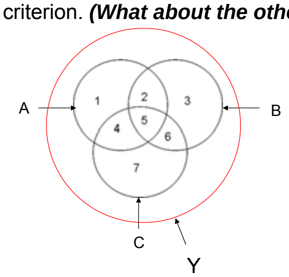
Unstandardized regression coefficients are the weights applied to the original measures. As such they are expressed in the unit of measurement of the variable, and thus are not directly comparable. They indicate the amount of change in the dependent variable for a unit change in the predictor. Tests of significance of regression coefficients are performed on these coefficients.

Standardized regression coefficients are the weights applied to the standardized measures and define the amount of change in the standardized dependent variable for a unit change in the standardized predictor. They are unitless, and are weights of variables that have a mean of 0, and a standard deviation of 1. They are not, however, directly comparable. Given the relationships among the variables, it is possible for one Beta to be larger than another, though the second may be significant and the first not.

22

## Venn Diagrams and the Dimensionality of Multiple Correlation

Consider a three predictor problem, A, B, and C. The following Venn diagram shows the sources of variance overlapping the criterion. ***(What about the other stuff?)***




Contributions to the variance of the criterion of each of the following components

1. Uniquely to A
3. Uniquely to B
7. Uniquely to C
2. Uniquely to that common to A and B
4. Uniquely to that common to A and C
6. Uniquely to that common to B and C
5. Uniquely to that common to A,B, & C.

23

This can be demonstrated by computing the multiple correlations with each pair of predictors and calculating the component accounted for by each segment in Slide 23.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | 0.81276[a] | 0.66058 | 0.62065 | 3.22834 |

a. Predictors: (Constant), c, b

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | 0.80110[a] | 0.64176 | 0.59962 | 3.31663 |

a. Predictors: (Constant), a, c

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | 0.80618[a] | 0.64993 | 0.60875 | 3.27859 |

a. Predictors: (Constant), b, a

24

Segment 1. $$\hat{R}^2_A = R^2_{A,B,C} - R^2_{B,C} = .73727 - .66058 = .07669$$

Segment 3. $$\hat{R}^2_B = R^2_{A,B,C} - R^2_{A,C} = .73727 - .64176 = .09551$$

Segment 7. $$\hat{R}^2_C = R^2_{A,B,C} - R^2_{A,B} = .73727 - .64993 = .08734$$

Segment 4.
$$\hat{R}^2_{AC} = R^2_{A,B,C} - R^2_B - \hat{R}^2_A - \hat{R}^2_C = .73727 - .70577^2 - .07669 - .08734 = .07513$$

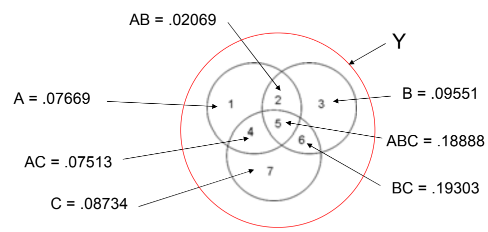
Segment 6.
$$\hat{R}^2_{BC} = R^2_{A,B,C} - R^2_A - \hat{R}^2_B - \hat{R}^2_C - = .73727 - .60116^2 - .09551 - .08734 = .19303$$

Segment 2.
$$\hat{R}^2_{AB} = R^2_{A,B,C} - R^2_C - \hat{R}^2_B - \hat{R}^2_A = .73727 - .73782^2 - .09551 - .07669 = .02069$$

Segment 5
$$\hat{R}^2_{ABC} = R^2_{A,B,C} - \hat{R}^2_A - \hat{R}^2_B - \hat{R}^2_C - \hat{R}^2_{AB} - \hat{R}^2_{AC} - \hat{R}^2_{BC}$$
$$= .73727 - .07669 - .09551 - .08734 - .07513 - .19303 - .02069 = .18888$$



## Contributions to the Squared Multiple Correlation




Note:

$r^2_{AY}$ = .07669 + .07513 + .18888 + .02069 = .36139 = .60116² = $r^2_{AY}$

$r^2_{BY}$ = .09551 + .02069 + .18888 + .19303 = .49811 = .70577² = $r^2_{BY}$

$r^2_{CY}$ = .08734 + .07513 + .18888 + .19303 = .54438 = 73782² = $r^2_{CY}$



## Comments on these values

Segments 1,3, and 7 are squared semipartial multiple correlations, thus they are always positive. Moreover, their tests of significance are identical to tests of significance of the corresponding regression coefficients.

Segments 2,4,5, and 6 are residuals and thus can be positive or negative. There are no obvious tests of significance, but the segments sum to $R^2$, thus it is possible to estimate the proportion that each contributes to $R^2$.

| | | | |
|---|---|---|---|
| Segment 1 | .07669/.73727, …10.40% | Unique A | 35.2% |
| Segment 3 | .09551/.73727 … 12.95% | Unique B | |
| Segment 7 | .08734/.73727, …11.85% | Unique C | |
| Segment 4 | .02069/.73727 … 2.81% | Unique AB | |
| Segment 6 | .07513/.73727, …10.19% | Unique AC | |
| Segment 2 | .19303/.73727, …26.18% | Unique BC | |
| Segment 5 | .18888/.73727, …25.62% | Unique ABC | |
| Total | 100.00% | | |



## Methods of Indirect Entry

Methods of indirect entry use a series of steps followed by the computer to determine the least number of predictors that give almost as good a level of prediction as the full number of predictors. These methods are good for the purpose intended, but are not of any value for interpretation.

**Stepwise Inclusion**

Step 1. The predictor that correlates highest with the criterion is the first predictor.

Step 2. Partial correlations, removing the effects of the first predictor, are contrasted, and the predictor with the highest partial correlation is the next predictor.

Step 3. The regression equation is determined, and each regression coefficient is tested for significance. If any are not significant, they are removed.

Step 4. Calculate the partial correlations removing effects of all predictors in the regression equation after Step 3. If none are significant, stop; otherwise go to step 3.

**Forward Inclusion**

The computer follows the previous steps except that the regression coefficients are not tested for significance after each equation is determined. As a result, some coefficients may not be significant in the final equation.

**Backward Elimination**

Step 1. All variables are entered into the equation, and R is calculated.

Step 2. Test regression coefficients for significance. If any are not significant remove the predictor that has the highest alpha.

Step 3. Calculate new R and test regression coefficients for significance. If any are not significant, remove the one with the highest alpha. Repeat until all regression coefficients are significant.