Introduction to Multiple Regression

Dale E. Berger

Claremont Graduate University

January 2014

Overview

Multiple regression is a flexible method of data analysis that may be appropriate whenever a quantitative variable (the Y, dependent, or criterion variable) is to be examined in relationship to any other factors (expressed as X, or independent, or predictor variables). Relationships may be nonlinear, independent variables may be quantitative or qualitative, and one can examine the effects of a single X variable or multiple X variables with or without the effects of other X variables taken into account (Berger, 2004; Cohen, Cohen, West, & Aiken, 2003).

Multiple Regression Models and Significance Tests

Many practical questions involve the relationship between a dependent or criterion variable of interest (call it Y) and a set of k independent variables or predictor variables (call them X_1, X_2, X_3,..., X_k), where the scores on all variables are measured for all N cases. For example, you might be interested in predicting job performance (Y) using information on years of experience (X_1), performance in a training program (X_2), and performance on an aptitude test (X_3). The score for an individual on a variable X_j is indicated as X_{ij}.. A multiple regression equation for predicting Y for an individual i can be expressed as follows:

$$(1) \qquad \hat{Y}_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + B_3 X_{i3}$$

To apply the equation, each X_j score for an individual case is multiplied by the corresponding B_i value, the products are added together, and the constant B_0 is added to the sum. The result is $\hat{Y}_i$, the predicted Y value for the case.

For a given set of data, the values for A and the B_js are determined mathematically to minimize the sum of squared deviations between predicted `$\hat{Y}_i$ and the actual Y scores. Calculations are quite complex, and best performed with the help of a computer, although simple cases with only one or two predictors can be solved by hand with special formulas.

The correlation between $\hat{Y}_i'$ and the actual Y value is also called the multiple correlation coefficient, $R_{y.12...k}$, or simply R. Thus, R provides a measure of how well Y can be predicted from the set of X scores. The following formula can be used to test the null hypothesis that in the population there is no linear relationship between Y and prediction based on the set of k X variables from N cases:

$$(2) \quad F = \frac{R^2_{y.12...k} / k}{(1 - R^2_{y.12...k}) / (N - k - 1)}, \quad df = k,\ N - k - 1 .$$

For the statistical test to be accurate, a set of assumptions must be satisfied. The key assumptions are that cases are sampled randomly and independently from the population, and that the deviations of Y values from the predicted Y values are normally distributed with equal variance for all predicted values of Y.

Alternatively, the independent variables can be expressed in terms of standardized scores where Z_1 is the z score of variable X_1, etc. The regression equation then simplifies to:

(3) $Z_{Y'} = ß_1Z_1 + ß_2Z_2 + ß_3Z_3$.

The value of the multiple correlation R and the test for statistical significance of R are the same for standardized and raw score formulations.

Test of R Squared Added

An especially useful application of multiple regression analysis is to determine whether a set of variables (Set B) contributes to the prediction of Y beyond the contribution of a prior set (Set A). The statistic of interest here, R squared added, is the difference between the R squared for both sets of variables ($R^2_{Y.AB}$) and the R squared for only the first set ($R^2_{Y.A}$). If we let k_A be the number of variables in the first set and k_B be the number in the second set, a formula to test the statistical significance of R squared added by Set B is:

$$\textbf{(4)} \quad F = \frac{(R^2_{Y.AB} - R^2_{Y.A}) / k_B}{(1 - R^2_{Y.AB}) / (N - k_A - k_B - 1)}, \quad df = k_B, N - k_A - k_B - 1.$$

Each set may have any number of variables. Notice that Formula (2) is a special case of Formula (4) where k_A=0. If k_A=0 and k_B=1, we have a test for a single predictor variable, and Formula (4) becomes equivalent to the square of the t test formula for testing a simple correlation.

Example: Prediction of Scores on a Final Examination

An instructor taught the same course several times and used the same examinations each time. The composition of the classes and performance on the examinations was very stable from term to term. Scores are available on a final examination (Y) and two midterm examinations (X_1 and X_2) from an earlier class of 28 students. The correlation between the final and the first midterm, r_{Y1}, is .60. Similarly, r_{Y2}=.50 and r_{12}=.30. In the current class, scores are available from the two midterm examinations, but not from the final. The instructor poses several questions, which we will address after we develop the necessary tools:

a) What is the best formula for predicting performance on the final examination from performance on the two midterm examinations?

b) How well can performance on the final be predicted from performance on the two midterm examinations?

c) Does this prediction model perform significantly better than chance?

d) Does the second midterm add significantly to prediction of the final, beyond the prediction based on the first midterm alone?

Regression Coefficients: Standardized and Unstandardized

Standard statistical package programs such as SPSS REGRESSION can be used to calculate statistics to answer each of the questions in the example, and many other questions as well. Because there are only two predictors, special formulas can be used to conduct an analysis without the help of a computer.

With standardized scores, the regression coefficients are:

$$(5) \quad \beta_1 = \frac{r_{Y1} - (r_{Y2})(r_{12})}{1 - (r_{12})^2}, \ and \quad \beta_2 = \frac{r_{Y2} - (r_{Y1})(r_{12})}{1 - (r_{12})^2}.$$

Using the data from the example, we find:

$$\beta_1 = \frac{.6 - (.5)(.3)}{1 - (.3)(.3)} = .49, \ and \ \beta_2 = \frac{.5 - (.6)(.3)}{1 - (.3)(.3)} = .35.$$

We can put these estimates of the beta weights into Formula (3) to produce a prediction equation for the standardized scores on the final examination. For a person whose standardized scores on the midterms are $Z_1 = .80$ and $Z_2 = .60$, our prediction of the standardized score on the final examination is:

$$Z_{Y'} = (\text{ß}_1)(Z_1) + (\text{ß}_2)(Z_2) = (.49)(.80) + (.35)(.60) = .602.$$

Once we have the beta coefficients for standardized scores, it is easy to generate the B_j regression coefficients shown in Formula (1) for prediction using unstandardized or raw scores, because

$$(6) \qquad B_1 = \beta_1 \frac{SD_Y}{SD_{X_1}}, \ B_2 = \beta_2 \frac{SD_Y}{SD_{X_2}}, \ and \ A = \bar{Y} - (B_1)(\bar{X}_1) - (B_2)(\bar{X}_2).$$

It is important that B_j weights not be compared without proper consideration of the standard deviations of the corresponding X_j variables. If two variables, X_1 and X_2, are equally predictive of the criterion, but the SD for the first variable is 100 times larger than the SD for the second variable, B_1 will be 100 times smaller than B_2! However, the beta weights for the two variables would be equal.

To apply these formulas, we need to know the SD and mean for each test. Suppose the mean is 70 for the final, and 60 and 50 for the first and second midterms, respectively, and SD is 20 for the final, 15 for the first midterm, and 10 for the second midterm. We can calculate $B_1 = (.49)(20/15) = .653$ and $B_2 = (.35)(20/10) = .700$, and $A = 70 – (.653)(60) – (.700)(50) = - 4.18$.

Thus, the best formula for predicting the score on the final in our example is

$$\hat{Y}_i = - 4.18 + .653 X_1 + .700 X_2$$

For someone with a score of 75 on the first midterm and 60 on the second midterm, the predicted score on the final is - 4.18 + (.653)(75) + (.700)(60) = 86.795, or about 87.

Multiple Correlation with Two Predictors

The strength of prediction from a multiple regression equation is nicely measured by the square of the multiple correlation coefficient, R^2 . In the case of only two predictors, R^2 can be found by using the formula

(7) $$R^2_{Y.12} = \frac{r_{Y1}{}^2 + r_{Y2}{}^2 - 2(r_{Y1})(r_{Y2})(r_{12})}{1 - r^2_{12}}.$$

In our example, we find

$$R^2_{Y.12} = \frac{(.6)^2 + (.5)^2 - 2(.6)(.5)(.3)}{1 - (.3)^2} = \frac{.43}{.91} = .473.$$

One interpretation of $R^2_{Y.12}$ is that it is the proportion of Y variance that can be explained by the two predictors. Here the two midterms can explain (predict) 47.3% of the variance in the final test scores.

Tests of Significance for R

It can be important to determine whether a multiple regression coefficient is statistically significant, because multiple correlations calculated from observed data will always be positive. When many predictors are used with a small sample, an observed multiple correlation can be quite large, even when all correlations in the population are actually zero. With a small sample, observed correlations can vary widely from their population values. The multiple regression procedure capitalizes on chance by assigning greatest weight to those variables which happen to have the strongest relationships with the criterion variables in the sample data. If there are many variables from which to choose, the inflation can be substantial. Lack of statistical significance indicates that an observed sample multiple correlation could well be due to chance.

In our example we observed R^2=.473. We can apply Formula (2) to test for statistical significance:

$$F = \frac{.473 / 2}{(1-.473) / (28-2-1)} = 11.2, \quad df = 2,25.$$

The tabled $F_{(2,\,25,\,.01)}$ = 5.57, so our findings are highly significant ($p < .01$). In fact, $p < .001$ because tabled $F_{(2,\,25,\,.001)}$ = 9.22.

Tests of Significance for R Squared Added

The ability of any single variable to predict the criterion is measured by the simple correlation, and the statistical significance of the correlation can be tested with the t-test, or with an F-test using Formula (2) with k=1. Often it is important to determine if a second variable contributes reliably to prediction of the criterion after any redundancy with the first variable has been removed.

In our example, we might ask whether the second midterm examination improves our ability to predict the score on the final examination beyond our prediction based on the first midterm alone. Our ability to predict the criterion with the first midterm (X_1) alone is measured by $(r_{Y1})^2 = (.6)^2 = .360$, and with both X_1 and X_2 our ability to predict the criterion is measured by $R^2 = .473$. The increase is our ability to predict the criterion is measured by the increase in R squared, which is also called “R squared added.” In our example R squared added = (.473 - .360) = .113. We can test R squared added for statistical significance with Formula (4), where Set A consists of the first midterm exam (X_1), and Set B consists of the second midterm exam (X_2). For our example we find

$$F = \frac{(.473 - .360)/1}{(1 - .473)/(28 - 1 - 1 - 1)} = 5.36, \quad df = 1, 25.$$

The *p*-value is .029, so our finding is statistically significant with p < .05, but not p < .01. We can conclude that the second midterm does improve our ability to predict the score on the final examination beyond our predictive ability using only the first midterm score.

Measures of Partial Correlation

The increase of R^2 when a single variable (B) is added to an earlier set of predictors (A) is identical to the square of the semipartial correlation of Y and B with the effects of set A removed from B. Semipartial correlation is an index of the unique contribution of a variable above and beyond the influence of some other variable or set of variables. It is the correlation between the criterion variable (Y) and that part of a predictor variable (B) which is independent of the first set of predictors (A). In comparison, partial correlation between Y and B is calculated by statistically removing the effects of set A from both Y and B. Partial and semipartial correlation have similar interpretations, and identical tests of statistical significance. If one is significant, so is the other.

The tests of statistical significance for both standardized and unstandardized regression coefficients for a variable Xj are also identical to the tests of significance for partial and semipartial correlations between Y and Xj if the same variables are used. This is because the null hypotheses for testing the statistical significance of each of these four statistics (B, beta, partial correlation, and semipartial correlation) have the same implication: The variable of interest does not make a unique contribution to the prediction of Y beyond the contribution of the other predictors in the model.

When two predictor variables are highly correlated, neither variable may add much unique predictive power beyond the other. The partial and semipartial correlations will be small in this case. The beta and B weights will not necessarily be small, but our estimates of these weights will be unstable. That is, the weight that each variable is given in the regression equation is somewhat arbitrary if the variables are virtually interchangeable. This instability of the estimates of beta and B is reflected in the tests of statistical significance, and the F tests will be identical to the F tests of the partial and semipartial correlations.

In the special case of two predictors, the standard error for beta (which is the same for both betas when there are only two predictors) can be calculated with the following formula and applied to our example:

$$(8) \qquad SE \ \ for\ beta\,(two\ \ predictors) = \sqrt{\frac{1 - R^2}{(N - k - 1)(1 - r_{12}^2)}} = \sqrt{\frac{1 - .473}{(25)(1 - .3^2)}} = .152.$$

Each beta can be tested for statistical significance using a t-test with df = N – k - 1, where t = beta / (SE for beta) = ß / $SE_{ß}$. For our second variable, this leads to t(25 df) = .352 /.152 = 2.316. If we wished to conduct an F test, we could square the t value and use $F_{(I,\ N\text{-}k\text{-}1)}$. For our data, this produces $F_{(1,25)}$ = 5.36 which is the same value that we obtained when we tested the R square added by X_2.

These tests give the same level of statistical significance because they each test whether X_2 contributes beyond X_1 in predicting Y. Thus, it is critically important to include the context when interpreting a beta weight. It is misleading to interpret a beta value without consideration of what else is in the model.

Confidence Intervals

Although p-values are used conventionally, they are not very informative because they are quite unstable and they give little information about the effect size. Often, the researcher is most interested in estimating the effect size and the margin of error in that estimate. That is exactly the function of a confidence interval.

Often the statistic of interest can be assumed to have a reasonably normal sampling distribution. In that case, all we need is a standard error for the statistic and we can create a confidence interval.

For the first midterm exam in the current example, $\beta_1 = .490$ and SE $_{\beta 1} = .152$ with df = 25. For a 95% confidence interval, we use $t_{.05,\,25} = 2.060$. The confidence interval for the population beta is

Probability [β_1 - (SE $_{\beta 1}$)($t_{.05,\,25}$) < population β_1 < β_1 + (SE $_{\beta 1}$)($t_{.05,\,25}$)] = 95%

Probability [.490 – (.152)(2.060) < population β_1 < .490 – (.152)(2.060)] = 95%

Probability [.177 < population β_1 < .803] = 95%

This confidence interval allows us to exclude zero as a plausible value for beta, but it also indicates that we have not pinned down the population value of β_1 with much precision. Of course, a larger sample size would produce a narrower confidence interval, if all else remains the same.

The relationship between standardized and unstandardized regression weights is

$$B_1 = \beta_1 \frac{SD_Y}{SD_{X_1}} \; and \; SE_{B1} = SE_{\beta 1} \frac{SD_Y}{SD_{X_1}}.$$

Thus, B_1 = (.490)(20/15) = .653 and SE $_{B1}$ = (.152)(20/15) = .203.

A 95% confidence interval for the population value of B_1 is

Probability [.653 – (.203)(2.060) < population B_1 < .653 – (.203)(2.060)] = 95%

Probability [.235 < population B_1 < 1.071] = 95%

Tolerance and Multicollinearity

Notice the effect of a large r_{12} on the SE for beta in Formula (8). As r_{12} approaches 1.0, the SE for beta grows very rapidly. If you try to enter two predictor variables that are perfectly correlated (r_{12}=1.0), the regression program may abort because calculation of the SE involves division by zero. When any one predictor variable can be predicted to a very high degree from the other predictor variables, we say there is a problem of multicollinearity, indicating a situation where estimates of regression coefficients are very unstable.

The SE for an unstandardized regression coefficient, B_j, can be obtained by multiplying the SE for the beta by the ratio of the SD for Y divided by the SD for the X_j variable:

$$SE_{B_j} = \frac{SD_Y}{SD_{X_j}} (SE_{\beta_j}). \qquad (9)$$

The t-test of statistical significance of B_j is t = (observed B_j)/(SE for B_j) with df=N-k-1, which is N-3 when there are two predictors. With more than two predictor variables (k > 2), the standard error for beta coefficients can be found with the formula:

$$SE_{B_j} = \sqrt{\frac{1 - R_Y^2}{(N - k - 1)(1 - R_{(j)}^2)}}. \qquad (10)$$

where R^2_Y indicates the multiple correlation using all k predictor variables, and $R^2_{(j)}$ indicates the multiple correlation predicting variable X_j using all of the remaining (k-1) predictor variables. The term $R^2_{(j)}$ is an index of the redundancy of variable X_j with the other predictors, and is a measure of multicollinearity. Tolerance, as calculated by SPSS and other programs, is equal to (1 - $R^2_{(j)}$). Tolerance close to 1.0 indicates the predictor in question is not redundant with other predictors already in the regression equation, while a tolerance close to zero indicates a high degree of redundancy.

Shrunken R Squared (or Adjusted R Squared)

Multiple R squared is the proportion of Y variance that can be explained by the linear model using X variables in the sample data, but it overestimates that proportion in the population. This is because the regression equation is calculated to produce the maximum possible R for the observed data. Any variable that happens to be correlated with Y in the sample data will be given optimal weight in the sample regression equation. This capitalization on chance is especially serious when many predictor variables are used with a relatively small sample. Consider, for example, sample R^2 = .60 based on k=7 predictor variables in a sample of N=15 cases. An estimate of the proportion of Y variance that can be accounted for by the X variables in the population is called "shrunken R squared" or "adjusted R squared." It can be calculated with the following formula:

$$(11) \qquad Shrunken\, R^2 = \tilde{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1} = 1 - (1 - .6)\frac{14}{7} = .20.$$

Thus, we conclude that the rather impressive R^2 = .60 that was found in the sample was greatly inflated by capitalization on chance, because the best estimate of the relationship between Y and the X variables in the population is shrunken R^2 = .20.

If all *k* predictors are uncorrelated with Y and with each other in the population, the expected value of R^2 is equal to *k* divided by (*N-1*). With *N* =15 and *k* = 7, the expected value of R^2 = .467.

$$(12) \qquad E(R^2) = k\,/\,(N-1)$$

A shrunken R squared equal to zero corresponds exactly to F = 1.0 in the test for statistical significance. If the formula for shrunken R squared produces a negative value, this indicates that the observed R^2 is smaller than expected if R^2 = 0 in the population, and the best estimate of the population value of R is zero. It is important to have a large number of cases (N) relative to the number of predictor variables (k). One rule of thumb is $N > 50 + 8k$ when testing R^2 and $N > 104 + k$ when testing individual B_j values (Green, 1991). In exploratory research the N:k ratio may be lower, but as the ratio drops it becomes increasingly risky to generalize regression results beyond the sample. If predictors are correlated with each other, even larger samples are desirable (Maxwell, 2000).

Stepwise Vs. Hierarchical Selection of Variables

Another pitfall, which can be even more serious, is inflation of the sample R^2 due to selection of the best predictors from a larger set of potential predictors. The culprit here is the "stepwise" regression option that is included in many statistical programs. For example, in SPSS REGRESSION it is very easy for the novice to use stepwise procedures whereby the computer program is allowed to choose a small set of the best predictors from the set of all potential predictors. The problem is that the significance levels reported by the computer program do not take this into account!!! As an extreme example, suppose you have 100 variables that are complete nonsense (e.g., random numbers), and you use them in a stepwise regression to predict some criterion Y. By chance alone about half of the sample correlations will be at least slightly positive and half at least slightly negative. Again, by chance one would expect that about 5 would be "statistically significant" with p<.05. The stepwise regression program will find all of the variables that happen to contribute “significantly” to the prediction of Y, and the program will enter them into the regression equation with optimal weights. The test of significance reported by the program will probably show that the R^2 is highly significant when, in fact, all correlations in the population are zero.

Of course, in practice, one does not plan to use nonsense variables, and the correlations in the population are not all zero. Nevertheless, stepwise regression procedures can produce greatly inflated tests of significance if you do not take into account the total number of variables that were considered for inclusion. Until 1979 there was no simple way to deal with this problem. A procedure that was sometimes recommended for tests of statistical significance was to set k equal to the total number of variables considered for inclusion, rather than set k equal to the number of predictors actually used. This is a very conservative procedure because it assumes that the observed R would not have grown larger if all of the variables had been used instead of a subset.

A more accurate test of significance can be obtained by using special tables provided by Wilkinson (1979). These tables provide values of R squared that are statistically significant at the .05 and .01 levels, taking into account sample size (N), number of predictors in the equation (k), and total number of predictors considered by the stepwise program (m). SPSS and other programs will not compute the correct test for you.

Another problem with stepwise regression is that the program may enter the variables in an order that makes it difficult to interpret R squared added at each step. For example, it may make sense to examine the effects of a training program after the effects of previous ability have already been considered, but the reverse order is less interpretable.

In practice, it is almost always preferable for the researcher to control the order of entry of the predictor variables. This procedure is called “hierarchical analysis,” and it requires the researcher to plan the analysis with care, prior to looking at the data. The double advantage of hierarchical methods over stepwise methods is that there is less capitalization on chance, and the careful researcher will be assured that results such as R squared added are interpretable. Stepwise methods should be reserved for exploration of data and hypothesis generation, but results should be interpreted with proper caution.

For any particular set of variables, multiple R and the final regression equation do not depend on the order of entry. Thus, the regression weights in the final equation will be identical for hierarchical and stepwise analyses after all of the variables are entered. At intermediate steps, the B and beta values as well as the R squared added, partial and semipartial correlations can be greatly affected by variables that have already entered the analysis.

Categorical Variables

Categorical variables, such as religion or ethnicity, can be coded numerically where each number represents a specific category (e.g., 1=Protestant, 2=Catholic, 3=Jewish, etc.). It would be meaningless to use a variable in this form as a regression predictor because the size of the numbers does not represent the amount of some characteristic. However, it is possible to capture all of the predictive information in the original variable with **c** categories by using (c-1) new variables, each of which will pick up part of the information.

For example, suppose a researcher is interested in the relationship between ethnicity (X1) and income (Y). If ethnicity is coded in four categories (e.g., 1=Euro-Americans; 2=African-Americans; 3=Latino-Americans; and 4=Other), the researcher could create three new variables that each pick up one aspect of the ethnicity variable. Perhaps the easiest way to do this is to use “dummy” variables, where each dummy variable (D_j) takes on only values of 1 or 0 as shown in Table 1.

Dummy variables are easy to create in SPSS using RECODE. For example:

```
RECODE  X1 (1=1)(ELSE = 0) INTO D1.

RECODE X1 (2=1)(ELSE = 0) INTO D2.

RECODE X1 (3=1)(ELSE = 0) INTO D3.
```

Table 1: Dummy Coding of Ethnicity

	Criterion	Ethnicity	Dummy variables		
Case	(Y)	(X1)	D1	D2	D3
----	---------	---------	--	--	--
1	25	1	1	0	0
2	18	2	0	1	0
3	21	3	0	0	1
4	29	4	0	0	0
5	23	2	0	1	0
6	13	4	0	0	0
:	:	:	:	:	:

In this example, D1=1 for Euro-Americans and D1=0 for everyone else; D2=1 for African-Americans and D2=0 for everyone else; D3=1 for Latino-Americans and D3=0 for everyone else. A person who is not a member of one of these three groups will be given the code of 0 on all three dummy variables. One can examine the effects of ethnicity by entering all three dummy variables into the analysis simultaneously as a set of predictors. The R squared added for these three variables as a set can be measured, and tested for significance using Formula (4). The F test for significance of the R squared added by the three ethnicity variables is identical to the F test one would find with a one-way analysis of variance on ethnicity. In both analyses the null hypothesis is that the ethnic groups do not differ in income, or that there is no relationship between income and ethnicity.

If there are four groups, any three can be selected to define the dummy codes. Tests of significance for R squared added by the entire set of (c-1) dummy variables will be identical in each case. Intermediate results and the regression weights will depend on the exact nature of the coding, however. There are other methods of recoding in addition to dummy coding that will produce identical overall tests, but will produce different intermediate results that may be more interpretable in some applications.

A test of the simple correlation of D1 with Y is a test of the difference between Euro-Americans and everyone else on Y. However, when all three dummy variables are in the model, a test of B_1 for Euro-Americans is a test of the difference between Euro-Americans and the reference group ‘’Other,” the group not represented by a dummy variable in the model. It is important to interpret this surprising result correctly. In a multiple regression model, a test of B or beta is a test of the ‘unique’ contribution of that variable, beyond all of the other variables in the model. In our example, D2 accounts for differences between African-Americans and other groups and D3 accounts for differences between Latino-Americans and other groups. Neither of these two variables can separate Euro-Americans from the ‘Other’ reference group. Thus, the unique contribution of variable D1 is to distinguish Euro-Americans from the ‘Other’ group.

To reiterate this important point, suppose we use D1, D2, and D3 to predict family income. Recall that D1=1 for Euro-Americans, D2=1 for African-Americans, D3 = 1 for Latino-Americans, and all of these variables are set to zero in all other cases. Further suppose that SPSS gave us the following regression equation: Predicted family income = $50,000 + $1000xD1 - $8000xD2 - $7000xD3. Thus, the B weight (unstandardized regression weight) for D1 is +$1000. Suppose this coefficient is statistically significant. How do you interpret this significance test and this value?

We can begin by asking, “What is the predicted income for the fourth group, ‘Other?’“ These folks have a score of 0 on D1, D2, and D3, so the predicted family income of the reference group is $50,000. Important note: the constant is always the predicted value for a case that has values of zero on all predictors. Next, we see that the predicted family income for Euro-Americans (for whom D1=1, D2=0, and D3=0) is $50,000 + $1,000 = $51,000. Thus, the B weight of +$1000 indicates how the predicted family income for Euro-Americans differs from the predicted family income for the reference group 'Other.' We can conclude that on average in our sample, the family income of Euro-Americans was $1000 greater than the average family income of minorities who are neither African-American nor Latino-American, and this difference is statistically significant. Thus, although the D1 variable makes the distinction between Euro-Americans and all others, the contribution of D1 when D2 and D3 are also in the model is the distinction between Euro-Americans and the reference group 'Others' which does not include African-Americans nor Latino-Americans. Tricky, and critically important!

Interactions

The interaction of any two predictor variables can be coded for each case as the product of the values for the two variables. The contribution of the interaction can be assessed as R squared added by the interaction term after the two predictor variables have been entered into the analysis. Interpretability of main effects in the presence of an interaction is improved if continuous variables are first 'centered' by subtracting the mean from each observed score. Tests of significance for main effects in a model that includes their interaction term should not be interpreted if the main effects were not centered in the computation of the interaction term.

It is also possible to assess the effects of an interaction of a categorical variable (X_1) with a quantitative variable (X_2). In this case, the categorical variable with **c** categories is recoded into a set of (c-1) dummy variables, and the interaction is represented as a set of (c-1) new variables defined by the product of each dummy variable with X_2. An F test for the contribution of the interaction can be calculated for R squared added by the set of interaction variables beyond the set of (c-1) dummy variables and X_2. The main effects must be in the model when contribution of the interaction is tested.

Multiple Regression and Analysis of Variance

The interaction between two categorical variables can be tested with regression analysis. Suppose the two variables have **c** categories and **d** categories, respectively, and they are recoded into sets of (c-1) and (d-1) dummy variables, respectively. The interaction can be represented by a set of (c-1) (d-1) terms consisting of all possible pairwise products constructed by multiplying one variable in the first set by one variable in the second set. Formula (4) can be used to conduct tests of significance for each set of dummy variables, and for the R squared added by the set of (c-1)(d-1) interaction variables after the two sets of dummy variables for the main effects. The denominator term that is used in Formula (4) for testing the contribution of the interaction set beyond the main effects (the two sets of dummy variables) is exactly equal to the Mean Squares Within Cells in ANOVA. The F tests of statistical significance for the sets of variables and the set of (c-1)(d-1) interaction variables are identical to the corresponding F tests in analysis of variance if the denominator of Formula (4) is replaced by the Mean Squares Within Cells obtained on the last step of the regression analysis.

In most applications the R squared added by each set of dummy variables will depend on the order of entry. Generally, the unique contribution of most variables will be less when they are entered after other variables than when they are entered prior to the other variables. This is described as "nonorthogonality" in analysis of variance. If the number of cases is the same in each of the (c)(d) cells defined by the (c) levels of the first variable and the (d) levels of the second variable, then the analysis of variance is orthogonal, and the order of entry of the two sets of dummy variables does not affect their contribution to prediction.

Missing Data

Missing data causes problems because multiple regression procedures require that every case have a score on every variable that is used in the analysis. There are several old and common ways to deal with missing data, but none is entirely satisfactory. Four of these ways will be described here: pairwise deletion, listwise deletion, deletion of variables, and coding of missingness.

If data are missing completely at random, then it may be appropriate to estimate each bivariate correlation on the basis of all cases that have data on the two variables. This is called "pairwise deletion" of missing data. An implicit assumption is that the cases where data are available do not differ systematically from cases where data are not available. In most applied situations this assumption clearly is not valid, and generalization to the population of interest is risky.

Another serious problem with pairwise deletion is that the correlation matrix that is used for multivariate analysis is not based on any single sample of cases, and thus the correlation matrix may not be internally consistent. Each correlation may be calculated for a different subgroup of cases. Calculations based on such a correlation matrix can produce anomalous results such as $R^2 > 1.0$. For example, if $r_{y1} = .8$, $r_{y2} = .8$, and $r_{12} = 0$, then $R^2_{y.12} = 1.28$! A researcher is lucky to spot such anomalous results, because then the error can be corrected. Errors in the estimate and testing of multivariate statistics caused by inappropriate use of pairwise deletion usually go undetected.

A second procedure is to delete an entire case if information is missing on any one of the variables that is used in the analysis. This is called "listwise deletion," the default option in SPSS and many other programs. The advantage is that the correlation matrix will be internally consistent. A disadvantage is that the number of cases left in the analysis can become very small and not representative of the population of interest. For example, suppose you have data on 9 variables from 100 cases. If a different group of 10 cases is missing data on each of the 9 variables, then only 10 cases are left with complete data. Results from such an analysis will be useless. The N:k ratio is only 10:9 so the sample statistics will be very unstable and the sample R will greatly overestimate the population value of R. Further, those cases that have complete data are unlikely to be representative of the population. Cases that are able (willing?) to provide complete data are unusual in the sample.

A third procedure is simply to delete a variable that has substantial missing data. This is easy to do, but it has the disadvantage of discarding all information that is carried by the variable.

A fourth procedure, popularized by Cohen and Cohen (1983), is to construct a new "missingness" variable (D_j) for every variable (X_j) that has missing data. The Dj variable is a dummy variable where D_j=1 for each case that is missing data on X_j, and D_j=0 for each case that has valid data on X_j. All cases are retained in the analysis; cases that are missing data on X_j are "plugged" with a constant value such as 999 or the mean. In the regression analysis, the missingness variable D_j is entered immediately prior to the X_j variable. The R squared added for the set of two variables indicates the amount of information that is carried by the original variable as it is coded in the sample. The R squared added by D_j can be interpreted as the proportion of variance in Y that can be accounted for by knowledge of whether or not information is available on X_j. The R squared added by X_j indicates predictive information that is carried by cases that have valid data on X_j.

A somewhat surprising fact is that the R squared added by X_j after D_j has been entered does not depend on the value of the constant that was used to indicate missing data on X_j. An advantage of "plugging" missing data with the mean of valid scores on X_j is that then D_j and X_j are uncorrelated: for both levels of D_j (cases with and without data on X_j), the mean value of X_j is equal to the same value. In this case, the order of entry is D_j and X_j does not affect the value of R squared added for either variable. It is important that only one number is used to plug missing data on any one variable. However, after D_j has entered the analysis, the R squared added by X_j plugged with 999 is identical to the R squared added by X_j plugged with the mean. The R squared added by X_j indicates additional predictive information that is carried by cases that have valid data on X_j.

The correlation of the missingness variable with other variables such as the criterion (Y) can be used to test the hypothesis that data are missing at random.

If data are missing on the dependent variable (Y), there is no alternative but to drop the case from consideration. If the loss is truly random, it might be reasonable to include the case for estimating the correlations among the predictors.

It is also important to consider how much data is missing on a variable. With only a small amount of missing data, it generally doesn't matter which method is used. With a substantial portion of data missing, it is important to determine whether the missingness is random or not. In practice, missingness often goes together on many variables, such as when a respondent quits or leaves a page of a survey blank. In such a case, it may be best to use a single missingness variable for several X_j variables. Otherwise, there may be serious multicollinearity problems among the D_j missingness variables.

There is a growing literature on dealing with missingness. Multiple imputation and maximum likelihood estimation are recommended by many statisticians as the best methods currently available. These methods, which are beyond the scope of this paper, include appropriate estimates of error variances and covariances. A good source for these modern methods is Enders (2011).

What to Report

Reasonable people may present different information, depending on their audience and purpose. It is useful to consider four distinct kinds of information. First, we have the simple correlations (r) which tell us how each individual predictor variable is related to the criterion variable, ignoring all other variables. The correlation of Y with an interaction term is not easily interpreted, because this correlation is greatly influenced by scaling of the main effects; it could be omitted from the table with no loss. (See Table 2)

The second type of information comes from R^2 added at each step. Here the order of entry is critical if the predictors overlap with each other. For example, if Sex had been entered alone on Step 1, R^2 added would have been .004**, statistically significant with p<.01. (R^2 added for the first term is simply its r squared.) Because of partial overlap with education, Sex adds only .001 (not significant) when Education is in the model. However, the interaction term adds significantly beyond the main effects (.002*), indicating that we do have a statistically significant interaction between Sex and Education in predicting Occupational Prestige.

Table 2: Regression of Occupational Prestige on Years of Education and Sex (N=1415)

Step	Variable	r	R^2 added	B	SE_B	Beta
1	Education (years)	.520***	.270***	1.668	.318	.518***
2	Sex (M=1; F=2)	-.063**	.001	-6.083	2.689	-.027
3	Educ X Sex	----	.002*	.412	.201	----
	(Constant)			22.403	4.300	

***p<.05; **p<.01; ***p<.001; Cumulative R squared = .273; (Adjusted R squared = .271).**
Note: B and SE_B are from the final model at Step 3, and Beta is from the model at Step 2 (both main effects, but no interaction term).

The third type of information comes from the B weights in the final model. These weights allow us to construct the raw regression equation, and we can use them to compute the separate regression equations for males and females, if we wish. The B weights and their tests of significance on the main effects are not easily interpreted, because they refer to the unique contribution of each main effect beyond all other terms, including the interaction (which was computed as a product of main effects). The test of B for the final term is meaningful, as it is equivalent to the test of R^2 added for the final term. In this case, both tests tell us that the interaction is statistically significant.

The fourth type of information comes from the beta weights on the model that contains only the main effects. This provides a test of the unique contribution of each main effect beyond the other main effects. If the main effects did not overlap at all, the beta weight would be identical to the r value for each variable. Here we see that Sex does not contribute significantly beyond Education to predicting Occupational Prestige (beta = -.027), although its simple r was -.063, p < .01. Alert: Presenting beta weights from the model with main effects only is unconventional, but it does provide useful information.

It is also good to present the cumulative R squared when all variables of interest have been entered into the analysis. A test of significance should be provided for each statistic that is presented, and the sample size should be indicated in the table. Figures can be helpful, especially to display interactions.

Final Advice

Look at your data! An especially good practice is to examine the plot of residuals as a function of Y. An assumption of regression analysis is that residuals are random, independent, and normally distributed. A residual plot can help you spot extreme outliers or departures from linearity. Bivariate scatter plots can also provide helpful diagnostics, but a plot of residuals is the best way to find multivariate outliers. A transformation of your data (e.g., log or square root) may reduce the effects of extreme scores and make the distributions closer to normal.

It is desirable to use few predictors with many cases. With k independent predictors, Green (1991) recommended $N > 50 + 8k$ when testing R^2 and $N > 104 + k$ when testing individual B_j. Larger samples are needed when predictor variables are correlated. If all population correlations are 'medium' (i.e., all ρ_{xy} and ρ_{xx} =.3), N = 419 is required to attain power = .80 with five predictors, but if all ρ_{xy} = .3 and ρ_{xx} =.5, then required N = 1117 (Maxwell, 2000). Statistical significance may not be very meaningful with extremely large samples, but larger samples are preferred because they provide more precise estimates of parameters and smaller confidence intervals.

If you have data available on many variables and you peek at your data to help you select the variables that are the best predictors of your criterion, be sure that your tests of statistical significance take into account the total number of variables that were considered. The problem is even more serious with stepwise regression where the computer does the peeking, and the significance tests provided by SPSS are wrong because they do not adjust for the number predictors considered.

Watch for multicollinearity where one predictor variable can itself be predicted by another predictor variable or set of variables. For example, with two highly correlated predictors you might find that neither beta is statistically significant but each variable has a significant simple r with the criterion and the multiple R is statistically significant. Further, each variable contributes significantly to the prediction when it is entered first, but not when it is entered second. In this case, it may be best to form a composite of the two variables or to eliminate one of the variables.

It is often useful to reduce the number of predictor variables by forming composites of variables that measure the same concept. A composite can be expected to have higher reliability than any single variable. It is important that the composites are formed on the basis of relationships among the predictors, and not on the basis of their relationship with the criterion. Factor analysis can be used to help formulate composites, and reliability analysis can be used to evaluate the cohesiveness of the composite. These analyses can be completed before regression analysis is applied.

Confidence intervals are likely to be more useful than *p*-values, and certainly more useful than a simple significant vs. not significant statement.

Finally, be thoughtful rather than mechanical with your data analysis. Be sure your summaries adequately reflect your data. Get close to your data. Look at distributions, residuals, etc. Don't trust the computer to do justice to your data. One advantage you have over the computer is that you can ask "Does this make sense?" Don't lose this advantage. Remember, to err is human, but to really screw up it takes a computer!

Recommended Sources

Berger, D. E. (2004). Using regression analysis. In Wholey, J. S., Hatry, H. P., & Newcomer, K. E. *Handbook of practical program evaluation* (2nd ed.). San Francisco: John Wiley & Sons.

Cohen, J. and Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences,* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences,* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Publications. [Good source for modern methods of dealing with missing data – maximum likelihood estimation and multiple imputation.]

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499-510. [Simple rules of thumb based on empirical findings.]

Havlicek, L. & Peterson, N. (1977). Effects of the violation of assumptions upon significance levels of the Pearson r. *Psychological Bulletin, 84*, 373-377. [You can get away with a lot - regression is remarkably robust with respect to violating the assumption of normally distributed residuals. However, extreme outliers can distort your findings substantially.]

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5(4)*, 434-458. [When predictors are correlated with each other, larger samples are needed, especially to test contributions of individual variables.]

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Routledge Academic. [accessible, filled with examples and useful advice, SPSS printouts, MANOVA]

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon. [This is an excellent resource for students and users of a range of multivariate methods, including multiple regression.]

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86*, 168-174. [The serious problem of capitalization on chance in stepwise analyses generally is not understood. Wilkinson provides simple tables to deal with this problem.]