A Practical Guide to Multiple Regression

Kathyrne Mueller, PhD

August 2, 2014

This document is about one specific type of multiple regression – OLS (ordinary least squares) multiple regression. OLS multiple regression is an extension of simple linear regression – where the term "simple" refers to only two variables (a criterion variable and a predictor variable). (The term OLS refers to the calculations made to perform the actual regression.) This document is not about logistic regression or any form of hierarchical regression.

In general terms, multiple regression provides information about prediction. Multiple regression does not compare different groups to each other. Multiple regression is not about correlations between variables (although, correlations are a part of the "underpinnings" of multiple regression). Multiple regression asks: Do the values of the variables A, B, and/or C predict the exact value of the variable X? A, B, and C are called predictor variables (IVs); X is called a criterion variable or a response variable (DV).

Multiple regression "subtracts" out the effects of the individual variables. So, the question asked by multiple regression is sometimes phrased as: Does the value of variable A predict the value of the variable X when controlling for the values of variables B and C?

Multiple regression is based on the same general theory as t-test and ANOVA. That means that the same general assumptions apply. Those assumptions are generally related to normal distributions and homogeneity (relative equality) of variance. Ideally, all variables in multiple regression must be measured at the interval or ratio level. Cheating a little (one categorical IV among many interval/ratio IVs) is ok, but the more we "cheat" the less reliable the results become. The DV (the criterion variable) must be measured at the interval or ratio level.

What Statistical Information is Provided by Multiple Regression?

SPSS will give us an intimidating array of statistical findings. But here's what we really need to know.

1. Is the overall regression model significant? The term "model" is used to refer to all of the IVs (predictor variables) under consideration and their ability to actually predict the value of the DV (the criterion or response variable) in a way that is more than just chance. If the overall model is significant, then we know that at least one of the predictor variables can provide real information about the value of the DV (the criterion or response variable).

For example, suppose we ask whether IQ score and annual income predict attitudes about gay marriage. The "model" refers to the two predictor variables (IQ and annual income) and whether they can predict attitudes. If the model is significant, then either IQ or annual income or both help us to predict attitudes better than just taking a shot in the dark.

Significance of the model is tested by an ANOVA. We will see F, df, and p in the SPSS output. We report those values in the text just as we would for the more familiar type of ANOVA that we learned earlier.

2. What is the magnitude of the effect of the overall model? In other words, what is the effect size? Effect size is measured by adjusted R^2. Some people would say that effect size tells us about practical significance (as opposed to statistical significance).

R^2 ranges in value from 0 to 1.0. Think of R^2 as a proportion. All of the variability in our DV (our criterion variable) is the "whole" that we are trying to explain. If we explain 10% of the whole (an R^2 value of .1) that means that our predictor variables explain 10% of the total variability in the DV. Other unknown variables explain much more of the variability – 90%. So, if our R^2 value is .1, then we will make lots of mistakes if we try to predict the value of the DV from our predictor variables – but our prediction will still be better than chance.

If our R^2 value is .5, then we our predictor variables explain half of the variability in our DV. Other unknown variables explain just as much variability. So, if our R^2 value is .5, the accuracy of our predictions is still far from perfect, but the accuracy is much better than chance and much better than the case when our R^2 value was only .1.

3. Which of the predictor variables is significant? Just like the ANOVA that we learned earlier, the ANOVA that tells us that the overall regression model is significant doesn't provide information about "where" exactly the significance lies. Perhaps only one predictor variable is significant. Perhaps all are significant. How do we know which predictor variables are significant?

We answer this question by looking at the B and beta values and their corresponding p values in the SPSS output. The B values are very much like the slope of the line in simple linear regression. For each unit increase in the IV, how much does the DV change? (in multiple regression, we have "subtracted out" any effect of the other predictor variables). A positive value means both variables increase together; a negative value means one variable increases as the other decreases.

Each B value has a corresponding p value that tells us whether that particular IV (predictor variables) is significant or not. As usual, if p is less than .05, then the predictor variable is significant.

4. Are some IVs better predictors than others? In multiple regression, it would be great if we could compare the relative predictive ability of the different IVs. For example, is income a better predictor of attitudes than annual income? How much better? We have two problems in attempting to answer this question.

Problem A. Annual income and IQ are clearly measured very differently – we can't directly compare IQ scores to annual income – that doesn't make any sense. So, we can't directly compare the B values for IQ or income because they are measured in completely different types of units. To solve that problem, we standardize the scores – sort of like creating a z score to standardize scores.

Beta values are the standardized version of the B values. Because the beta values are standardized, we can directly compare them (as long as we are careful about the second problem discussed below). The higher the beta value (disregarding the positive or negative sign), the greater the "predictive ability." Now we can directly compare our IVs to see which one is the best predictor of the DV.

Problem B. Multiple regression can be calculated with several different modifications – enter method, backwards method, forward method, stepwise method, etc. In the "enter" method, all of the predictor

variables are analyzed simultaneously (all of the predictor variables get dumped into one big "model."). If we use the enter method, we can compare the beta values to each other in a way that makes sense.

But some other methods of calculating multiple regression enter the different IVs according to specific rules. And because of those rules, we can't necessarily directly compare the beta values of the different predictor variables.

What are the Hypotheses for Multiple Regression?

The hypotheses for multiple regression can be expressed in several ways. Here is the one that I like best. We use this approach when we predict the value of one variable from the value of multiple other variables. Do you see why it makes sense?

Ho: $\beta_1=\beta_2=\beta_3=\beta_4$ = etc = 0

Of course, you would replace the subscripts with something more meaningful, such as an abbreviation for each specific predictor variable.

Or the null hypothesis might be: when controlling for A, B, and C, $\beta=0$.

What Potential Problems Need to be Avoided?

B**e careful of nonlinear relationships** – OLS multiple regression was designed for linear relationships. That's one reason why the levels of measurement should be interval or ratio – we can only have a true linear relationship if we use an interval or ratio scale of measurement.

But the bigger problem is collinearity – also called multicollinearity. Collinearity occurs when two or more predictor variables are highly correlated with each other. If two predictor variables are highly correlated, that means that we are essentially measuring the same construct twice – and it wouldn't make sense to include the same predictor variable twice in the same regression. Collinearity plays havoc with variability so that it becomes difficult to truly determine whether our regression is significant.

The problem of collinearity is one reason why we almost always see a correlation matrix in a paper that uses multiple regression. The reader should be able to evaluate the correlations among variables to assess the possible presence of collinearity and to correctly interpret the regression.

If you read articles about multiple regression, you may also see specific statistics that help us to assess collinearity. One of these "**collinearity diagnostics"** is called VIF (variance inflation factor). A VIF of 1.0 indicates no collinearity. A VIF of 5.0 or greater indicates a serious problem with collinearity.

Does This Seem Confusing?

There is a lot to consider with multiple regression. So to begin – focus on the basics – focus on the big picture:

Is the regression significant?
What is the effect size?
Which predictor variables are significant?
Do we have a problem with collinearity?