Addendum: Multiple Regression Analysis

(DRAFT 8/2/07)

When conducting a rapid ethnographic assessment, program staff may:

- Want to assess the relative degree to which a number of possible predictive variables influence an outcome of interest in a population.

  For example, in a study of unprotected sexual behavior, the analyst may want to determine whether a number of possible predictive variables are significant, such as sexual orientation, "race"/ethnicity, age group, educational attainment and gender. Findings from such an analysis may help target prevention programming.

- Want to assess the extent to which observed mean differences between sub-groups on an outcome are explained by a third variable;

  Many health status differences attributed to "race"/ethnicity may largely be due to economic differences between the ethnic groups being compared.

- Decide whether a better outcome for a program over the outcome for a comparison group may be confounded by differences between the groups other than the intervention.

  For example, in our initial evaluation of the Delve curriculum, we examined whether significant differences in post-test knowledge scores between Delve users and a comparison group (that was provided consultant assistance only) were influenced by differences in the educational level of participants and/or by disrupting intervening events. Multiple regression analysis indicated that the positive outcome for Delve was independent of these possible confounding variables.

Multiple regression is the analytic strategy of choice for answering questions such as these. It is a general analytic approach, used extensively in quantitative social science research, particularly by economists and sociologists. Multiple regression is based on the general linear model (this is all the math you will get in this section):

$$y = \alpha + \beta(x) + \varepsilon$$

Where:

y is the dependent (outcome) variable of interest;

α is the intercept of y on the x axis (the point on x where on average y is zero)

β is the slope of y on x; for every unit of x, y on average changes this much; and

ε is the error term or disturbance, the amount that needs to be added or subtracted for the average case to match the actual value of y.

This simple regression model (i.e., a simplified depiction of reality to help us better understand a phenomenon) is expanded on in *multiple* regression. While this sounds very technical, if you think about it carefully and work through the material slowly it is easy to understand conceptually and be able to use the results, which is all we really want you to do. The analyst can add several x (independent or predictor) variables, add terms for interactions between predictor variables, or add multiplier terms to account for curvilinear patterns of data. (Curvilinear patterns occur when relationships between measures differ at different points on a distribution, for example a u-shaped curve where there is a strong relationship at the high and low ends of a predictor but not at the mid-range.)

When continuous measures are added to a regression model, there will be several weights (β's) for each case—one for each independent x variable entered into the model. The estimated value for a case is the sum α (constant for all cases) plus the weight for each x times the value of x for that case, plus error (ε). Statistical packages provide a significance test for each of the predictors to assist in determining whether the x variable significantly predicts y or if the relationship observed could be due to chance.

One form of predictor variable (x) is important to discuss. When you have a dichotomy (e.g., yes or no, male or female, person of color or not) coded as either 0 (no) or 1 (yes), the weight for this variable is added to the intercept (α) and essentially changes the point where the line of y on x crosses the zero point on x. Variables such as this are often called indicator or "dummy" variables. Dummy variables can also be included in interaction with continuous variables, for example to look at the relative effect of age on initiation of sexual intercourse among male and female members of a population.

Multiple regression can also handle dichotomous dependent variables by using a variant called "logistic" regression. Logistic regression estimates the odds of an outcome (y) of zero or one for each independent (x) variable in the model. A common dichotomous outcome in hiv/aids prevention programming would be whether or not a person admits to engaging in risky sexual behavior. Other forms of distribution of y, such as ordered categories or very rarely occurring events, can also be handled but expert statistical advice should be obtained by non-researchers before doing so. (Actually, any application of regression analysis is likely to need assistance from a statistical analyst for most users of Delve!)

Limitations of Multiple Regression

Limitations and problems in applying and interpreting multiple regression must be discussed. While it is often said that regression analysis is "robust" to deviations from its assumptions, there are a number of technical statistical assumptions behind multiple regression that are often violated. This is particularly true in research with small samples or which includes many related predictor (x) variables. While these issues have very technical and statistical explanations, we here provide a basic summary. What is important is that users of regression understand that there are many limitations and nuances to its application and interpretation.

A basic assumption is “no specification error.” This means that all relevant variables are included and irrelevant variables are excluded, and that the relationship is in fact linear. Another primary assumption is that the independent variable is not correlated with its error term—that is that there is not a high degree of error at one end of the distribution. However, this may occur when there is poor predictability of an outcome at high levels of a predictor but not at low levels, or the reverse. For example, alcohol consumption generally increases with level of education in the U.S. population—on average, people with higher education are more likely to drink alcohol. However, in a small sample of individuals skewed to those at low educational level, there may be a very high variation in how much those at the lower educational level drink and less variation at the higher level. In this case, the error term would be correlated negatively with the independent measure.

With small samples, cases that are extremes (“outliers”) also can cause misleading results. The solution is to always carefully examine your raw data and decide whether some cases are so extreme that they may be incorrectly recorded or otherwise in error, in which case they should be fixed or excluded. Outliers may also be indications that your sample is too small and that if a larger sample were drawn more apparently extreme cases would emerge.

The situation called “multi-colinearity” occurs when there are two (or more) highly related x variables in a model. The x variables essentially can cancel one another out and a stable estimate cannot be obtained. A good example of this occurred when modeling the predictors of tobacco consumption in states in the U.S. Both median income and a measure of educational attainment (percent of adults over 25 with college degrees or higher) were used in a model. While both measures had a high (negative) relationship to tobacco consumption when examined alone, the model was un-interpretable when both median income an education were included, since they were both highly related to each other and essentially cancelled each other out. (In this case, the solution was to create a single “latent variable” of socio-economic status—SES-- for each state in the U.S., made up of weighted values of education and income.) This SES variable solved the problem of colinearity between education and income and was, as predicted, negatively related to tobacco consumption.

Recommendations

Beyond very simple models with relatively few variables, the user of this website would do well to consult with a statistician/ analyst before attempting to use multiple regression. However, it is important to understand the approach as a consumer of quantitative studies.

If you are using a standard statistical package such as SPSS, SAS, or STATA, multiple regression (including logistical regression) is quite accessible. Excel spreadsheets also can be analyzed using simple regression analysis, which is available in the spreadsheet calculation software.

Multiple regression is a very useful tool in statistical analysis and, once basic descriptive statistics are mastered, regression is the next step in the learning curve.

Example of Use of Multiple Regression in Outcome Evaluation:

Jill Florence Lackey and Associates Study of Milwaukee Public Museum Science Program, in

Lackey, J.F., Borkan, S.S., Torti, V., Welnetz, T., & Moberg, D.P. (2007). The story behind the findings: Yes, the Science Explorations program worked, but why? Curator 50(3).

Jill Florence Lackey and Associates conducted a study of a program to improve the science achievement and scientific career aspirations of middle school mainly Latino and African American girls. Girls were assigned to treatment or comparison groups, with some erosion of the comparison group into the treatment condition. Girls were surveyed annually about their experience in the program, their motivation to continue in science, their knowledge of science, their attitudes about the importance of science, and their intents to pursue a career in science.

In the initial analysis of one-year follow-up data, there were raw mean differences between groups that showed the intervention girls were significantly higher on four of six outcomes than were the comparison girls. This is indicated in Table A-1 by asterisks in the means column. However, both groups also showed significant before-after change on several of the outcomes, as indicated in the baseline to 1 year p-value column. There was concern that there was differential dropout from the research in the two groups which made them non-equivalent at one year follow-up. There was also some indication that the demographic characteristics of the final groups differed.

Thus a multiple regression analysis was undertaken. For each outcome (y) variable, a multiple regression equation was estimated in which the independent (x) variables were the baseline version of the outcome measure, ethnicity of the students (African American or Latino), the school attending from which assignment to conditions was made, and an indicator variable for condition (0 = comparison, 1= intervention group).

The results of this analysis definitively support the overall benefit of the program on increased science knowledge, confidence in one's own scientific ability, grade point average in science, and career consideration in science. In all of these areas, the one-year score for the intervention students was higher than that for comparison students, controlling for baseline differences, ethnicity and school. Interestingly, this result was positive for the program even on an outcome on which the girls overall declined over time—career consideration in science.

The final column of Table A-1 shows the result of the regression analysis for the intervention indicator variable, in the original metric* of the dependent measure. A variable is a significant predictor at the $p < .05$ level if its coefficient is roughly twice the standard error for the coefficient. Thus in Table A-1 on the *science knowledge* row, we see that the average adjusted difference between the intervention and comparison samples at one year is 1.66 points, on a scale with a mean of about 6 and standard deviation (s.d.)

of about 2.3 at baseline. Similarly, science GPA (theoretical range from 0 to 4.0; mean in this sample at baseline of about 3.0 and s.d. of about .76) showed an average adjusted difference at one year of 0.64 points in favor of the intervention students. While career consideration in science decreased in both groups (last row), the intervention group still had a significantly higher mean score.

*NOTE—“standardized” regression coefficients can also be obtained from most statistical packages, which express the results in standard deviation units which can be compared between variables measured in different increments or between different samples.

Table A-1: Example of Regression Analysis for Assessing Program Outcome:
Evaluation of the *Science Explorations* Program

	Baseline				**1 year follow-up**						**Multi-variate analysis**
Change in	Intervention[a] (n=132)		Control[a] (n=84)		Intervention (n=132)			Control (n=84)			Adjusted[b]B, (s.d.) and p value for program effect at 1 yr
	Mean	SD	Mean	SD	Mean	SD	Baseline to 1 yr *p*-value	Mean	SD	Baseline to 1 yr *p*-value	
Science knowledge	6.22	2.28	6.24	2.36	7.04*	1.92	<.001	5.48	2.55	.029	1.66 (0.36), p < .001
Science importance	2.41	0.39	2.38	0.41	2.43	0.36	.632	2.34	0.40	.364	.052 (.063), p = .41
Outside support for science	2.20	0.62	2.26	0.62	2.28	0.57	.231	2.42	0.49	.024	-.098 (.086), p = .26
Science confidence	2.00	0.42	2.07	0.50	2.13*	0.40	.002	1.84	0.40	.001	.308 (.066), p < .001
Science GPA	2.88*	0.80	3.22	0.71	3.23*	0.18	<.001	2.66	0.86	<.001	.637 (.114), p < .001
Career consideration in science	1.74	0.48	1.80	0.56	1.65*	0.51	.082	1.49	0.42	<.001	.183 (.075), p = .016

a. Cases with valid data used in the analysis. Baselines without one year follow-up were not analyzed; case loss is n= 142 originally assigned to intervention and n= 68 originally assigned to control condition. In addition, 16 cases originally assigned to the control group were subsequently placed in the intervention.

b. Multivariate analysis adjusted for baseline value of outcome measure, ethnicity and school to estimate 1 year program effects.

* Unadjusted difference between intervention and control group means significant (p< .05) at this time point.