PRIMER FOR MULTIVARIATE REGRESSION

Regression analysis represents a key tool for assessing the nature of a focal association, checking for the statistical significance of a focal association and, most importantly, elaborating on a bivariate association to test theoretical arguments about the relationship between two variables. This document provides a brief refresher on the purpose and meaning of bivariate regression analysis and an overview of multivariate regression analysis. Rather than focus on the mathematics of multivariate regression (which are really a straightforward extension of those for bivariate regression), this introduction will focus on the practical application of these regression tools. Specifically, it demonstrates the creation and interpretation of output from an SPSS estimation of a bivariate regression model and a few multivariate models to exemplify a typical elaboration process. Please note that the example here is meant to demonstrate the mechanics of multivariate regress as a tool for assessing and elaborating on basic associations; it, by no means, represents a complete treatment of the topic of multivariate regression.

The example uses data from the General Social Survey to explore the simple hypothesis that individual income is positively related to formal educational attainment. Note that the GSS subsample used here includes only adults who earned income from salary or wages in the previous year.

Bivariate regression model

- The first step in exploring any focal relationship is the test of whether the two variables are associated with one another in a theoretically predicted way. A simple bivariate regression model does the trick in many situations.
- In SPSS, choose the *Analyze* menu, then *Regression*, and then *Linear*. Select *r_income*, "Rs annual income in dollars" as the dependent variable and *educ*, "Highest year of school completed" as the independent variable.
- Partial output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.307[a]	.094	.094	29208.27928

a. Predictors: (Constant), HIGHEST YEAR OF SCHOOL COMPLETED

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-12340.0	3416.649		-3.612	.000
	HIGHEST YEAR OF SCHOOL COMPLETED	3390.532	243.573	.307	13.920	.000

a. Dependent Variable: Rs annual income in dollars

- **Important interpretation points:**

  - **Model fit:** R-square in the "Model Summary" box indicates the proportion of the total variance in the dependent variable explained by the independent variable, or the proportional reduction of error in predicting values of dependent variable achieved by taking into consideration values on the independent variable. Our results indicate that about 9.4% of the total variation in income is explained by differences in education. This low R^2 is not surprising given that so many other factors likely affect income as well. The "Standard Error for the Estimate" of R-Square allows us to perform a hypothesis test to determine whether the R^2 for the model is significantly different from 0 or significantly different from a model with a different set of variables. This kind of hypothesis test involves what is called an F-test.

- **Components of the regression line** ("B" column under "Unstandardized Coefficients") characterize the straight line that best fits the observed scatterplot of income by education. This least-squares regression line is the straight line for which the sum of the squared prediction errors is minimized. This line is characterized by two components:
  - The coefficient labeled "Constant" is the Y-intercept of the regression and indicates the predicted value of the dependent variable for a case with a value of 0 on the independent variable. Someone with no education is predicted to have lost $12,340 last year.
  - The slope coefficient next to the name of the independent variable indicates the substantive effect of the independent variable on the dependent variable. The slope for the "Highest year of school completed" indicates how much the predicted value of income changes for a one-unit change in education. The slope coefficient of 3390.532 indicates that for each one-year increase in education, income is predicted to increase by $3,390.53.

- **Information for inferential testing**: As part of the effort to demonstrate the validity of our theoretical expectations, we need to determine whether the observed effect of education on income occurred by chance (as a result of random sampling error) or reflects a real association that exists in the population from which the GSS sample was drawn. In other words, we want to know whether the slope coefficient in the least-squares regression line is *statistically significant*. Answering this question involves setting up a test of the following hypotheses:

  H_1: $\beta \neq 0$ (the independent and dependent variables are associated in the population)
  H_0: $\beta = 0$ (the independent and dependent variables are NOT associated in the population; the sample association that appears in the sample data reflects chance sampling error.)

  where β refers to the slope coefficient in the least squares regression line for the population.

  In testing for statistical significance we are trying to determine how likely it is to have observed a regression slope this large in the sample if, in reality, the slope in the population is equal to zero. We could test this hypothesis by calculating the standard error for the slop coefficient and converting this slope coefficient to appropriate test-score (t-score). We could then compare the obtained test score to the theoretical sampling distribution of all such possible scores to figure out the probability of obtaining a test score of this magnitude if the sample was drawn from a population in which the coefficient was actually 0 (i.e., the probability that the null hypothesis is true).

  Fortunately, SPSS has already done all the dirty work of calculating the standard error, t-score, and even the p-value for the regression coefficients. All we have to do is interpret the results. The standard error for the slope of our regression model is listed in the "Std. Error" column. The obtained t-value for each component is found in the column marked "t" and the probability (p-value) of obtaining such a t-value if the population value was actually 0 (i.e., if the null hypothesis was true) is found in the column marked "Sig." In order to answer the question about the statistical significance of our slope coefficient, we can refer to the "Sig." column. If the p-value reported there is lower than our preset maximum probability (alpha usually set at .05), then we can say that the probability that the null hypothesis is true is so low that we don't believe that it is true. In our example, the obtained t for the slope coefficient is 13.920 which has a p-value of .000. This means that there is essentially no chance that the coefficient happened by chance as the null hypothesis implies. We can confidently reject the null hypothesis in favor of the research hypothesis that the effect of education on income is *not* 0 in the population; the slope coefficient is statistically significant, supporting our theoretical expectations.

- **Standardized coefficients:** The next column of the "coefficients" box displays the "Standardized coefficients" for the effect of "Highest year of school completed" on "R's income." The standardized coefficient (called Beta or b*) expresses the impact of the independent variable in terms of standard deviation units. It tells us the number of standard deviations the dependent variable increases or decreases with a one standard deviation increase in the independent variable. In our example, the standardized coefficient of .307 indicates that income goes up by about three-tenths of a standard deviation for each increase of one standard deviation in education. The standardized coefficient is calculated by multiplying the unstandardized coefficient, b, by the ratio of the standard deviations for the independent and dependent variables. Because they express all coefficients in terms of the same units (standard deviations), standardized coefficients become especially handy in multivariate models where we want to directly compare the size of the impacts of different independent variables.

Multivariate regression model

- Demonstrating that there is a substantively large and statistically significant bivariate association between education and income is only the first step in the process of building the case for the existence of a real (causal) relationship between the two variables in the population. The next steps involve eliminating alternative explanations for the association (both sources of spuriousness and redundancy) and demonstrating that the association between the independent and dependent variables fits into a broader system of relationships involving antecedent, intervening, consequent, and/or conditioning variables in a way that is theoretically anticipated. These steps are carried out through the use of multivariate regression analysis.
- Multivariate regression techniques represent a straightforward extension of the bivariate regression analysis just reviewed. As the name implies, multivariate models simply add more predictors to the existing bivariate model. Three key differences are worth highlighting: 1) the interpretation of model fit (R-squared) must incorporate a reference to all of the variables in the model; 2) the slope coefficients must be interpreted as *partial* coefficients; and 3) we are now interested not only in the magnitude of a coefficient in a single model, but how the slope coefficient related to our focal independent variable changes from one model to another. An extension of the bivariate example above will quickly illustrate these differences.

Exclusionary strategy Part 1

- Lets start by excluding one possible source of spuriousness, respondents' age. Compared to younger people, older respondents have had more time to accumulate education and, simply because they have had more time to build a career, are likely to make more money. Since age is likely correlated to both the independent and dependent variable and is causally prior to both, it represents a potential source of spuriousness that must be controlled if we are to build the case that education has a causal impact on income.
- In SPSS, choose the *Analyze* menu, then *Regression*, and then *Linear*. As before, we select *r_income*, "Rs annual income in dollars" as the dependent variable. Now, however, our independent variables include both *educ*, "Highest year of school completed" and *age*, "Age of respondent." *Note that one of these variables, age, is really a control variable and that we are primarily interested in examining the effect of our focal independent variable, education.* Although we know the difference, SPSS does not so the program refers to all predictors, regardless of their role in the analysis, “independent variables.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.341[a]	.117	.116	28855.14823

a. Predictors: (Constant), AGE OF RESPONDENT, HIGHEST YEAR OF SCHOOL COMPLETED

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-25540.6	3887.936		-6.569	.000
	HIGHEST YEAR OF SCHOOL COMPLETED	3312.657	240.897	.300	13.751	.000
	AGE OF RESPONDENT	344.062	50.292	.149	6.841	.000

a. Dependent Variable: Rs annual income in dollars

Important interpretation points:

- **Model fit:** R-square is a general measure of the overall fit of entire model; it indicates the proportion of the total variance in the dependent variable explained by *all of the independent* (focal and control) *variables in combination*. Now that we have added age as another predictor, we can now explain 11.7% of the total variance in income. Or, to think of this another way, we can reduce our errors in predicting values of income by 11.7% if we take into consideration values on *both education and age*. Once again, SPSS provides us with the information needed to conduct an F-test of for whether the R^2 for the model is significantly different from 0 or significantly different from our earlier model containing only education as a predictor.

- **Y-intercept:** In a multivariate model, the Y-intercept ("Constant") indicates the predicted value of the dependent variable for a case with a value of 0 on ALL of the independent variables. So, a person age 0 with no education is predicted to have lost $25,540.60 last year. Of course, we should avoid placing too much emphasis on the substantive meaning on the value of this y-intercept since there is there are no members of the sample (or the population of interest) who have these characteristics. Instead, the y-intercept represents an important starting point in predicting values of income for individuals with combinations of age and education that do exist in the sample.

- **Partial slope coefficients**: Each slope coefficient in a multiple regression model reflects the impact of the given variable while *controlling for all other variables included in the model*. In other words, these partial regression coefficients show use the *net* impact of these variables on the dependent variable with the influence of all other variables in the model removed. In our results, the coefficient for education indicates that, once the influence of age is removed, income increases by an average of $3,312.66 for each additional year of education. Some people like to think of partial coefficients as the influence of a one-unit change in the independent variable *among those with similar characteristics on the control variables*. So, it is reasonable (although not exactly statistically precise) to say that among people that are of similar age, a one-year difference in education is association with an additional $3,312.66 in income. Of course, the partial slope coefficient of 344.06 for age can be interpreted in a similar way: Among people with similar levels of education, a one-year increase in age is associated with an additional $340.06 in income.

- **Attenuation of the education effect**: If education does really have a causal impact on income as our theoretical argument suggests, then we should see a statistically significant effect of education even when we remove the effects of age (and any other potential source of spuriousness or redundancy). If the original association is wiped out (completely attenuated) when we statistically control for age, then your theoretical expectation of a causal relationship between education and income would be contradicted. Thus, comparing the partial coefficient for our focal independent variable to the coefficient from the bivariate model provides an indication of how much of the original association was actually spurious, created by the influence of age on both education and income. In fact, after controlling for age, the effect of education doesn't change much (from 3390.532 in the bivariate model to 3312.657 in the multivariate model) and remains statistically significant with a .000 p-value. This adds a bit more support to the idea that there is a real (causal) relationship between education and income.

Exclusionary strategy Part 2

- The exclusionary strategy is potentially endless. Given the complexity of social phenomena, there are countless factors that might affect the dependent variable in most studies and, to the extent that they are also associated with the focal independent variable, each of these influences also represents a potential source of spuriousness or redundancy. Predicting income is no exception; we can think of a wide range of other factors that might influence income and, to the extent that they are also correlated with education, these third variables might account for the observed association between education and income. Two more fairly obvious examples are gender and race; both of these variables are may be associated with both education and income and controlling for these will help us to identify the true net effect of education on income.

- **Dummy variables**: Both gender and race are categorical variables, so their inclusion adds a new wrinkle to our analysis. Both variables must be converted into dummy variables before they can be used in our regression (or correlation) analysis. To tap the effects of gender, we can create a variable, *female*, that takes a value of 1 for women and a value of 0 for men (note that *male* is the omitted category). We can also create a series of dummy variables to characterize respondents' race. Using the GSS data, we could create dummy variables for each major racial group plus a residual category for people of other races. The dummy variable for one of these race categories is omitted as the reference group and the rest are entered into the multivariate regression model as new independent variables. Recall that education remains our focal independent variable and age, gender, and race are simply added as controls for possible sources of redundancy or spuriousness.

Dummy variable: A **dichotomous** categorical variable taking a value of 1 for those cases with a particular attribute and 0 for all others

- Example: The effect of gender can be examined using a dummy variable called "female" that takes a value of 1 for women and a value of 0 for men.

- The regression coefficient related to the dummy variable indicates the difference in the value of the dependent variable between cases with a value of 1 on the dummy variable and those with a value of 0 on the dummy variable
  - Example: In predicting income, the coefficient for "female" indicates the average difference in income between females and males.

- A series of dummy variables is often used to examine differences in the dependent variable across a larger number of categories of a variable
  - Representing a multi-category variable with a single variable is unacceptable because a "one-unit increase" is meaningless for such a variable
  - Example: region represented with a series of four dummies: North, South, East, and West
  - Note that categories may combined in order to produce stable estimates and to capture only the statistically important distinctions between categories

- In a regression model, one category (dummy) must be omitted. This omitted category is the reference category against which the other categories are compared; the regression coefficient for each included dummy represents the mean difference in the dependent variable between that category and the reference category.

- **More multivariate regression output**:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.423[a]	.179	.176	27860.76952

a. Predictors: (Constant), OTHRAC, FEMALE, ASIANPI, NATAMER, AGE OF RESPONDENT, HIGHEST YEAR OF SCHOOL COMPLETED, BLACK

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-15425.5	3923.485		-3.932	.000
	HIGHEST YEAR OF SCHOOL COMPLETED	3346.168	235.430	.303	14.213	.000
	AGE OF RESPONDENT	298.343	49.083	.129	6.078	.000
	FEMALE	-13872.9	1296.783	-.226	-10.698	.000
	BLACK	-6115.009	1850.374	-.071	-3.305	.001
	ASIANPI	-9747.498	3983.689	-.052	-2.447	.015
	NATAMER	-14088.9	6441.599	-.046	-2.187	.029
	OTHRAC	-7544.070	3640.494	-.044	-2.072	.038

a. Dependent Variable: Rs annual income in dollars

Important interpretation points:

- **Model fit:** The reported R-square indicates that we have boosted the explanatory power of our model by adding indicators of respondents' gender and race. The combination of education, age, gender, and race account for just about 18% of the total variation in respondents' income.

- **Partial slope coefficient for the dummy variables**: The key to interpreting slope coefficients for dummy variables is remembering that you are making comparisons to the category that is omitted from the model -- the reference category. For example, in estimating the effects of gender on income, we have included a variable taking a value of 1 for females and have omitted the (implicit) dummy variable for males.
  - The coefficient of -13872.9 for the variable female indicates that, on average, women in our sample earned almost $14,000 less in 2002 than did the men in our sample. As before, each slope coefficient in a multiple regression model reflects the impact of the given variable while controlling for all other variables included in the model. So, we know that the observed gender difference in income is not due to differences in education, age, or race among the women and men in the sample.
  - The coefficient for each race category is interpreted in a similar way. Here we have omitted the dummy variable for white race as our reference category (whites represent the largest group and most theoretical arguments focus on income differences between whites and other groups). So, the coefficient for each of the included categories indicates the difference between the average income of members of the group and the average income of whites in the sample. For example, the coefficient for the dummy variable *black* indicates that, on average, black members of the sample earned about $6,115 less than did white members of the sample. Once again, this racial difference is among those with similar values on the other variables in the model (i.e., net of the influence of education, age, and gender).

- **Attenuation of the education effect**: Given our focal interest in the effects of education on income, the most important finding in the output above is that the coefficient for education is not greatly attenuated when we control for two more potential sources of spuriousness, gender and race. In fact, a comparison of coefficients in the previous two models shows that the coefficient for education actually increases slightly with the addition of controls for gender and race (from 3312.66 to 3346.17). So, once again, the persistence of the education effect even after eliminating two more potential sources of spuriousness bolsters the argument that education has a real causal effect on income.

Inclusive strategy

- After you have eliminated all of the sources of redundancy and spuriousness that you can think of, it is time to move on to the inclusive strategy. Demonstrating that the focal relationship is connected to a broader network of relationships predicted by your theory helps to strengthen the support for the theoretical explanation. Fortunately, this inclusive strategy can also be accomplished through the use of multivariate regression models.
- In this example, we will test the theoretical argument that education affects income through its influence on occupational prestige. That is, we expect that high levels of education increase the chances of ending up in a prestigious occupation and that this occupational location is associated with higher pay. In this sense, occupational prestige can be thought of as a key intervening variable linking education to income. We can classify the occupations of the workers in the sample using an occupational prestige score created in 1980. The score ranges from 0 to 100 with higher scores indicating higher levels of occupational prestige. *If this theoretical argument is correct, the coefficient for education should be attenuated (reduced) when we control for the influence of occupational prestige in our multivariate regression model.* Here is the multivariate output to test this theoretical assumption:

- **More multivariate regression output**:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.471[a]	.222	.218	27127.91522

[a] Predictors: (Constant), RS OCCUPATIONAL PRESTIGE SCORE (1980), ASIANPI, FEMALE, OTHRAC, NATAMER, AGE OF RESPONDENT, BLACK, HIGHEST YEAR OF SCHOOL COMPLETED

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-20043.0	3847.403		-5.209	.000
	HIGHEST YEAR OF SCHOOL COMPLETED	2078.833	261.181	.188	7.959	.000
	AGE OF RESPONDENT	261.187	47.933	.113	5.449	.000
	FEMALE	-14248.5	1263.217	-.232	-11.280	.000
	BLACK	-5652.328	1802.281	-.066	-3.136	.002
	ASIANPI	-7899.541	3883.193	-.042	-2.034	.042
	NATAMER	-11401.7	6277.770	-.037	-1.816	.069
	OTHRAC	-6774.255	3545.549	-.040	-1.911	.056
	RS OCCUPATIONAL PRESTIGE SCORE (1980)	527.004	52.048	.239	10.125	.000

[a] Dependent Variable: Rs annual income in dollars

Important interpretation points:

- **Model fit:** According to the reported R-square statistic, we can account for about 22% of the variation in income by taking into consideration a combination of education, age, gender, race, and occupational prestige.
- **Partial slope coefficient for occupational prestige**: As expected, occupational prestige has positive effect on income; a one-unit increase in prestige is associated with an increase in income of about $527. This may not seem like much in comparison to the large coefficients for race and gender. However, we must keep in mind that these predictor variables have drastically different scales. A comparison of the standardized coefficients indicates that a increase of one standard deviation in occupational prestige produces a larger bump in income than does a one-standard-deviation increase in any other predictor
- **Attenuation of the education effect**: Once again, the change attenuation of the education coefficient is of central interest to us given that our focal relationship refers to the impact of education on income. As predicted in our theoretical arguments, the inclusion of occupational prestige in our model does result in a substantial reduction in the coefficient for education (from 3346.17 to 2078.83). This attenuation of the coefficient is consistent with the theoretical argument that education influences income, at least in part, by influencing the types of jobs that respondents end up in. Of course, the fact that the net coefficient for education remains fairly large and statistically significant indicates that there must be other mechanisms, besides occupational prestige, through which education affects income. Specifying these mechanisms from our theoretical model and testing for these mechanisms using multivariate regression techniques represents an important continuation of the inclusive strategy.