SPSS PC Version 10: Regression Analysis [1]

The following uses a set of variables from the "1995 National Survey of Family Growth" to demonstrate how to use some procedures available in SPSS PC Version 10.

Regression analysis allows us to examine the substantive impact of one or more variables on another by using the components of the equation for the "best-fitting" regression line. Once again, while the calculations of these components can be tedious by hand, they are lightning fast with SPSS.

Using SPSS for bivariate and multi-variate regression analysis:

- Choose *Analyze* then *Regression* and then *Linear*. For a bivariate regression model you just need to specify the dependent variable and the single independent variable in the dialogue box that comes up. Hit "OK."

- For example, if I wanted to look at how the educational attainment of the women in our NSFG data is affected by the educational attainment of their mothers, I would enter *momed* as my independent variable and *educ* as my dependent variable (would it make sense to switch these?). The syntax for these commands is (first setting values of 95 to missing for momed):

```
missing values momed (95).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT educ
  /METHOD=ENTER momed .
```

  Once I hit *OK*, SPSS automatically kicks out a ton of output on the regression model predicted.

- This output allows you to answer a wide variety of questions.

    - The "Model Summary" and "Anova" boxes give goodness of fit measures and measures of significance for the entire model.

    - The "Coefficients" box gives information about the independent variable(s). First, the "B" column under "Unstandardized Coefficients" in the "Coefficients" box provides the value of the Y-intercept [labeled "(Constant")] and the slope representing the effect of mothers' education on the dependent variable, the education of the respondents. This least squares regression line is the straight line for which the sum of the squared prediction errors are minimized. The Y-intercept tells us that the predicted value of education for someone whose mother had 0 years of education is 9.464 years. The slope for the *momed* variable tells us that the predicted value of respondents' education increases by about .248 years for every one-unit increase in mother's education.

[1] Prepared by Kyle Crowder of the Sociology Department of Western Washington University, and modified by Patty Glynn, University of Washington.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.375[a]	.140	.140	2.5711

a. Predictors: (Constant), MOMED Mother's Education Years

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11629.337	1	11629.337	1759.201	.000[a]
	Residual	71215.775	10773	6.611		
	Total	82845.111	10774			

a. Predictors: (Constant), MOMED Mother's Education in Years

b. Dependent Variable: EDUC Education in Years

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.464	.081		117.158	.000
	MOMED Mother's Education in Years	.284	.007	.375	41.943	.000

a. Dependent Variable: EDUC Education in Years

- The next column of the "coefficients" box displays the "Standardized coefficient" for the effect of "momed" on "Years of School Completed." 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. 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.

- Since we are dealing with sample data here, the next question we may want to answer is whether the observed effect of mothers education on education occurred by chance (as a result of random sampling error) or represents a real relationship that exists in the population from which the NSFG sample was drawn. In other words, we want to know whether the components of the least-squares regression line are *statistically significant*. Here we are trying to determine how likely it is to have received the regression components that we did if, in reality, these components are equal to zero in the population. Fortunately, SPSS has already done all the 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 Y-intercept and the slope of our regression model are listed in the column of the "Coefficients" box marked "Std. Error." The obtained t-value for each component is found in the column marked "t" and the probability of obtaining such a t-value (the P-value) if the population value was actually 0 is found in the column marked "Sig." In order to answer the question about the statistical significance of our components, we can refer to the "Sig." column. Since our p-values for both the constant and the regression slope are well below any conventional alpha level, the components of the bivariate regression equation are both statistically significant.