

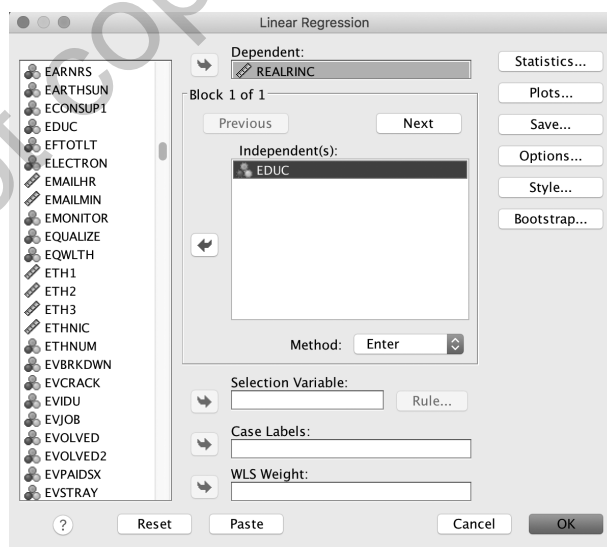
CORRELATION AND REGRESSION ANALYSIS

Regression analysis allows us to predict one variable from information we have about other variables. In this chapter, linear regression is discussed. Linear regression is a type of analysis that is performed on interval and ratio variables (labeled “scale” variables in SPSS Statistics). However, it is possible to incorporate data from variables with lower levels of measurement (i.e., nominal and ordinal variables) through the use of dummy variables. We will begin with a bivariate regression example and then add some more detail to the analysis.

BIVARIATE REGRESSION

In the case of bivariate regression, researchers are interested in predicting the value of the dependent variable, Y , from the information they have about the independent variable, X . We will use the example below, in which respondent’s occupational prestige score is predicted from number of years of education. Choose the following menus to begin the bivariate regression analysis:

Analyze → Regression → Linear . . .



The “Linear Regression” dialog box will appear. Initially, select the variables of interest and drag them into the appropriate areas for dependent and independent variables. The variable “REALRINC,” respondent’s actual annual income, should be moved to the “Dependent” area, and “EDUC,” respondent’s number of years of education, should be moved to the “Independent(s)” area. Now, simply click “OK.” The following SPSS Statistics output will be produced:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.287 ^a	.083	.082	28181.172

a. Predictors: (Constant), Highest year of school completed

In the first column of the “Model Summary” box, the output will yield Pearson’s r (in the column labeled “R”), followed in the next column by r -square (r^2). SPSS Statistics also computes an adjusted r^2 for those interested in using that value. R -square, like lambda, gamma, Kendall’s tau- b , and Somers’ d , is a PRE (proportional reduction in error) statistic that reveals the proportional reduction in error by introducing the dependent variable(s). In this case, $r^2 = .083$, which means that 8.3% of the variation in real annual income is explained by the variation in years of education. Although this percentage might seem low, consider that years of education is one factor among many (8.3% of the factors, to be exact) that contribute to income, including major field of study, schools attended, prior and continuing experience, region of the country, gender, race/ethnicity, and so on. We will examine gender (sex) later in this chapter to demonstrate multiple regression.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.164E+11	1	1.164E+11	146.50	.000 ^b
	Residual	1.294E+12	1629	794178451		
	Total	1.410E+12	1630			

a. Dependent Variable: R’s income in constant \$

b. Predictors: (Constant), Highest year of school completed

ANOVA (analysis of variance) values, including the F statistic, are given in the above table of the linear regression output.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-17734.68	3500.562		-5.066	.000
	Highest year of school completed	2933.597	242.368	.287	12.104	.000

a. Dependent Variable: R’s income in constant \$

The coefficients table reveals the actual regression coefficients for the regression equation, as well as their statistical significance. In the “Unstandardized Coefficients” columns, and in the “B” column, the coefficients are given. In this case, the *b* value for number of years of education completed is 2,933.597. The *a* value, or constant, is -17,734.68. By looking in the last column (“Sig.”), you can see that both values are statistically significant ($p = .000$). Remember, the *p* value refers to the probability that the result is due to chance, so smaller numbers are better. The standard in social sciences is usually .05; a result is deemed statistically significant if the *p* value is less than .05. We would write the regression equation describing the model computed by SPSS Statistics as follows:

$$\hat{Y} = bX + a \rightarrow \hat{Y} = \$2,933.60X * -\$17,734.68 *$$

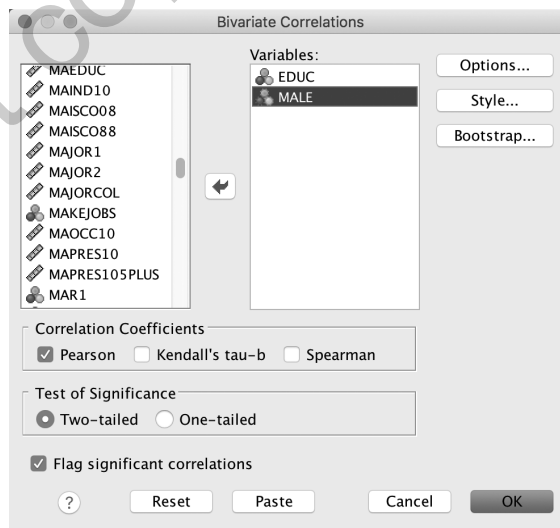
*Statistically significant at the $p \leq .05$ level.

The coefficient in the bivariate regression model above can be interpreted to mean that each additional year of education provides a \$2,933.60 predicted increase in real annual income. The constant gives the predicted real annual income when years of education is zero; however, as is often the case with a regression equation, that may be beyond the range of the data for reasonable prediction. In other words, if no one had zero or near zero years of education in the sample, the range of the data upon which the prediction was calculated did not include such, and we should be cautious in making predictions at those levels.

CORRELATION

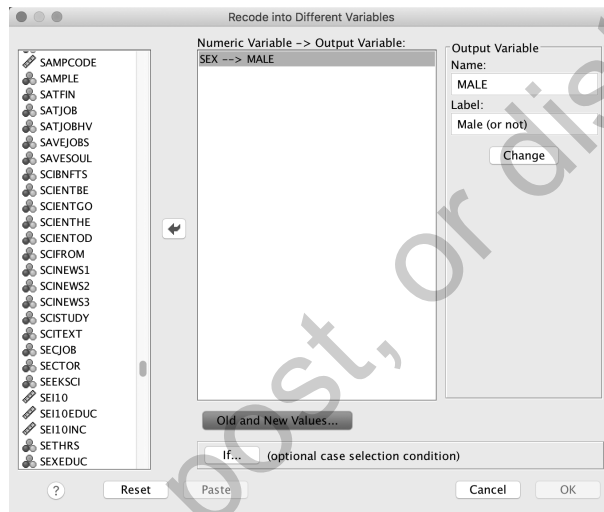
Information about correlation tells us the extent to which variables are related. Below, the Pearson method of computing correlation is requested through SPSS Statistics. To examine a basic correlation between two variables, use the following menus:

Analyze → Correlate → Bivariate . . .

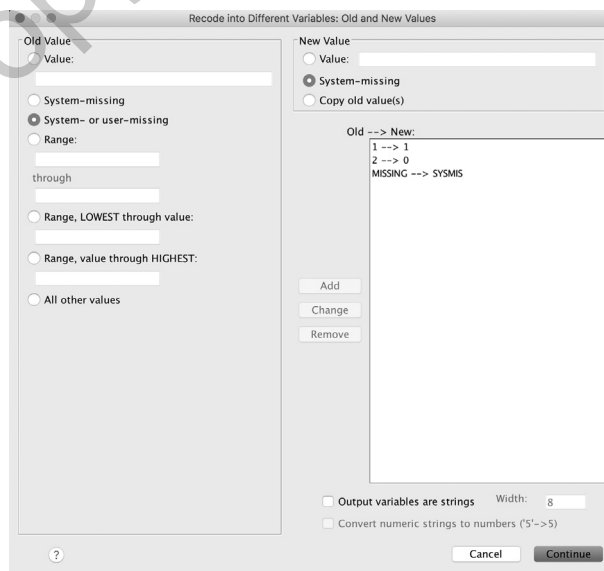


In the “Bivariate Correlations” dialog box, choose the variables you wish to examine. In the preceding case, “MALE” (a dummy variable representing sex, described in further detail below, under “Multiple Regression”) and “EDUC,” representing years of education, have been selected. For now, “MALE” is a recoded version of the sex variable, where a male respondent is coded as 1 and a female respondent is coded as 0. Thus, a “1” indicates “male” and a “0” indicates “not male,” with a proportional range in between. This allows us to treat a nominal dichotomy as an interval/ratio variable and then use it in regression and correlation analysis. Follow the following menus to create the male dummy variable:

Transform → Recode into Different Variables . . .



Select SEX, and then add the name and label, as above. Now click “Old and New Values . . .” Enter the recoding instructions, as illustrated below.



Now, click “Continue,” and then click “OK” in the first dialog box. The new variable, “MALE,” will be created. Be sure to do the appropriate fine-tuning for this new variable (e.g., eliminate decimal places, because there are only two possible values this variable can take: 0 and 1) in the Variable View window.

Returning to the correlation exercise, the output that results is shown in the following table:

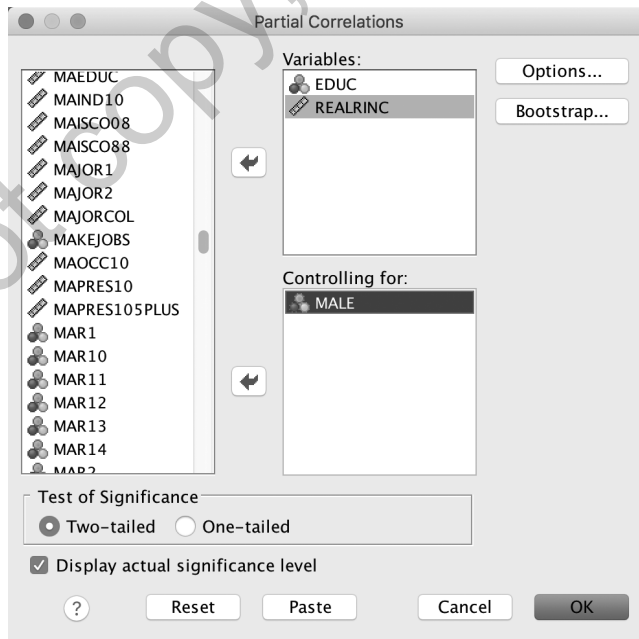
Correlations

		Highest year of school completed	Male (or not)
Highest year of school completed	Pearson Correlation	1	-.012
	Sig. (2-tailed)		.513
	N	2858	2858
Male (or not)	Pearson Correlation	-.012	1
	Sig. (2-tailed)	.513	
	N	2858	2867

Note that in the output, the correlation is an extremely small, $-.12$, which is not statistically significant ($p = .513$). This tells us that being male is not correlated with having completed a greater number of years of education.

It is also possible to produce partial correlations. Suppose you are interested in examining the correlation between occupational prestige and education. Further suppose you wish to determine the way that sex affects that correlation. Use the following menus to produce a partial correlation:

Analyze → Correlate → Partial . . .



In the “Partial Correlations” dialog box, you will be able to select the variables among which you wish to examine a correlation. You will also be able to select the control variable, around which partial correlations will be computed. In this case, years of education (“EDUC”) and occupational prestige score (“REALRINC”) have been selected for correlation analysis. The control variable is “MALE.” (It is also possible to include more than one control variable.)

SPSS Statistics provides the following output:

Correlations

Control Variables			Highest year of school completed	R's income in constant \$
Male (or not)	Highest year of school completed	Correlation	1.000	.302
		Significance (2-tailed)	.	.000
		df	0	1628
R's income in constant \$	R's income in constant \$	Correlation	.302	1.000
		Significance (2-tailed)	.000	.
		df	1628	0

Here, the correlation is noteworthy, at .302, and is statistically significant ($p = .000$). This is indicative of a relationship between education and income. Correlation information about variables is useful to have before constructing regression models. Should you want to know more, many textbooks in statistics and research methods have detailed discussions about how this information aids in regression analysis.

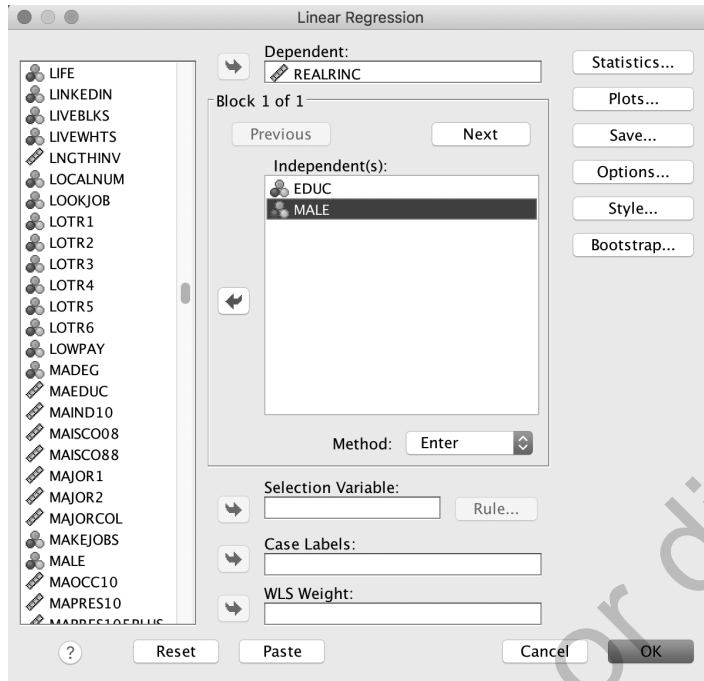
MULTIPLE REGRESSION

Now, suppose a researcher wished to include one or more additional independent variables in a bivariate regression analysis. This is very easy to do using SPSS Statistics. All you need to do is move the additional variables into the “Independent(s)” area in the “Linear Regression” dialog box, as seen below:

Analyze → Regression → Linear . . .

Because linear regression requires interval-ratio variables, one must take care when incorporating variables such as sex, race/ethnicity, religion, and the like. By creating dummy variables from the categories of these nominal variables, you can add this information to the regression equation.

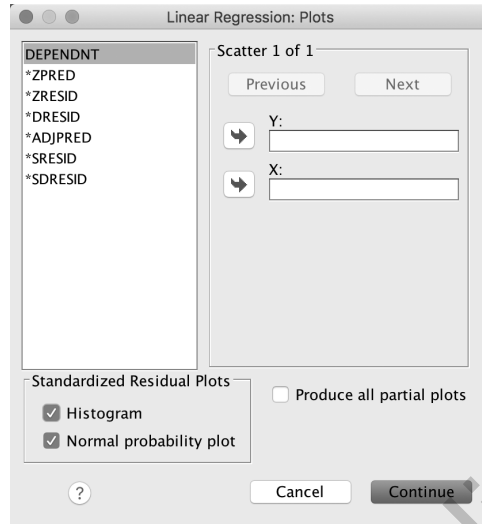
To do this, use the recode function (for more information about recoding variables, see Chapter 2, “Transforming Variables”). Create a dichotomous



variable for all but one category, the “omitted” comparison category or attribute, and insert each of those dichotomies into the “Independent(s)” area. The number of dummy variables necessary for a given variable will be equal to $K - 1$, where K is the number of categories of the variable. Dichotomies are an exception to the cumulative property of levels of measurement, which tells us that variables measured at higher levels can be treated at lower levels but *not* vice versa. Dichotomies, typically considered categorical or nominal, can be “coded” to be treated as if they are at any level of measurement.

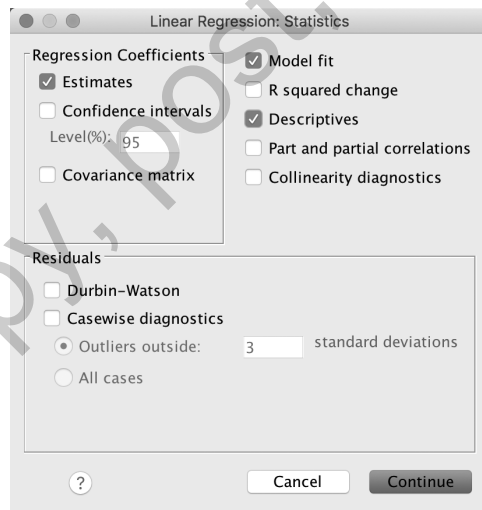
For the case of sex, we already have a dichotomy exclusive of transgender categories and other conditions, so the recoding just changes this to one variable: “MALE.” (Alternatively, you could have changed it to “FEMALE.”) The coding should be binary: 1 for affirmation of the attribute, 0 for respondents not possessing the attribute. Now, as was entered into the previous dialog box, just select the new recoded variable, “MALE,” from the variable bank on the left and drag it into the “Independent(s)” area on the right. You may need to set the variable property to scale in the Variable View tab of the Data Editor window so that SPSS Statistics will allow that variable to be included in the regression analysis. Newer versions of SPSS Statistics track variable types and often will not allow you to include variables with lower levels of measurement in analyses requiring variables with higher levels of measurement.

After recoding as necessary and dragging your variables of interest into their respective areas, click the “Plots . . .” button, and you will be shown the “Linear Regression: Plots” dialog box:



Here, you can avail yourself of a couple of useful graphics: a histogram and a normal probability plot. Click each box to request them. Then click “Continue.”

When you are returned to the “Linear Regression” dialog box, select the “Statistics . . .” button. The following dialog box will appear:



There are a number of options, including descriptive statistics, that you may select to be included in the SPSS Statistics linear regression output. For now, leave the defaults checked as shown, and click “Continue” in this box; then click “OK” when returned to the “Linear Regression” dialog box.

On next page you will find tables from the SPSS Statistics output that results. The first table reports the descriptive statistics that were requested. The next two tables give the same sort of information as before in the bivariate regression case: Pearson’s r (correlation coefficient), r^2 (PRE), and ANOVA (analysis of variance) values.

Descriptive Statistics

	Mean	Std. Deviation	N
R's income in constant \$	23785.34	29412.102	1631
Highest year of school completed	14.15	2.880	1631
Male (or not)	.49	.500	1631

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.339 ^a	.115	.114	27686.392

a. Predictors: (Constant), Male (or not), Highest year of school completed

b. Dependent Variable: R's income in constant \$

In this case, $r^2 = .115$, which means that 11.5% of the variation in respondents' real annual income ("REALRINC") is explained by the variation in the independent variables: years of education ("EDUC") and sex ("MALE").

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.621E+11	2	8.107E+10	105.765	.000 ^b
	Residual	1.248E+12	1628	766536284		
	Total	1.410E+12	1630			

a. Dependent Variable: R's income in constant \$

b. Predictors: (Constant), Male (or not), Highest year of school completed

The "Coefficients" table (on page 136), again, provides the information that can be used to construct the regression model and equation. Note that the dummy variable, "male," was not statistically significant.

$$\hat{Y} = bX_1 + bX_2 + a \rightarrow \hat{Y} = \$3,045.39X_1 + \$10,619.76X_2 - \$24,512.82$$

*Statistically significant at the $p \leq .05$ level.

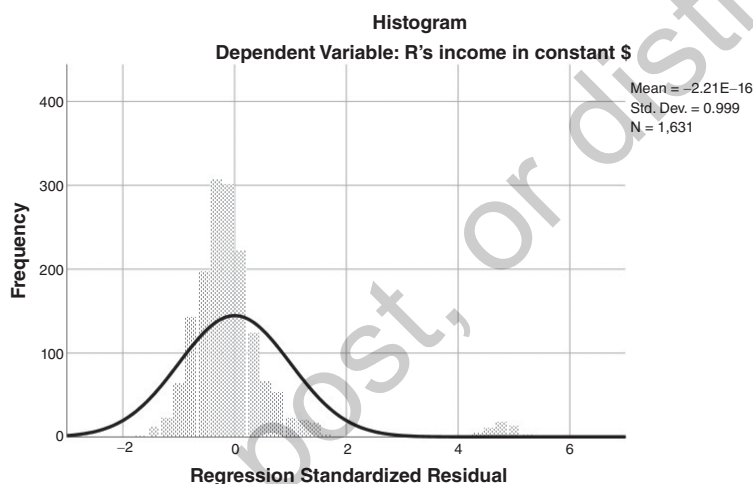
The X_1 coefficient ("EDUC," years of education) can be interpreted to mean that each additional year of education provides a \$3,045.39 predicted increase in real annual income. The X_2 coefficient ("MALE," dummy variable for gender) can be interpreted to mean that men have a predicted real annual income of \$10,619.76 more than women for this prediction model. In this case, both independent variables are statistically significant, with $p = .000$.

Coefficients^a

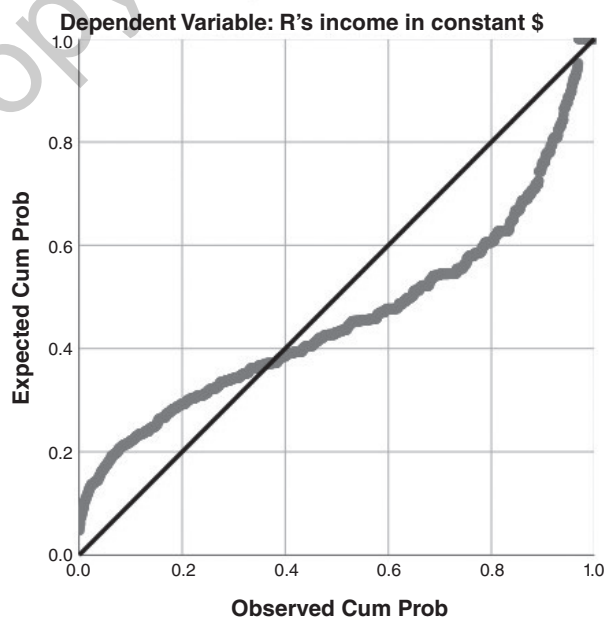
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-24512.824	3549.145		-6.907	.000
	Highest year of school completed	3045.388	238.552	.298	12.766	.000
	Male (or not)	10619.760	1373.944	.181	7.729	.000

a. Dependent Variable: R's income in constant \$

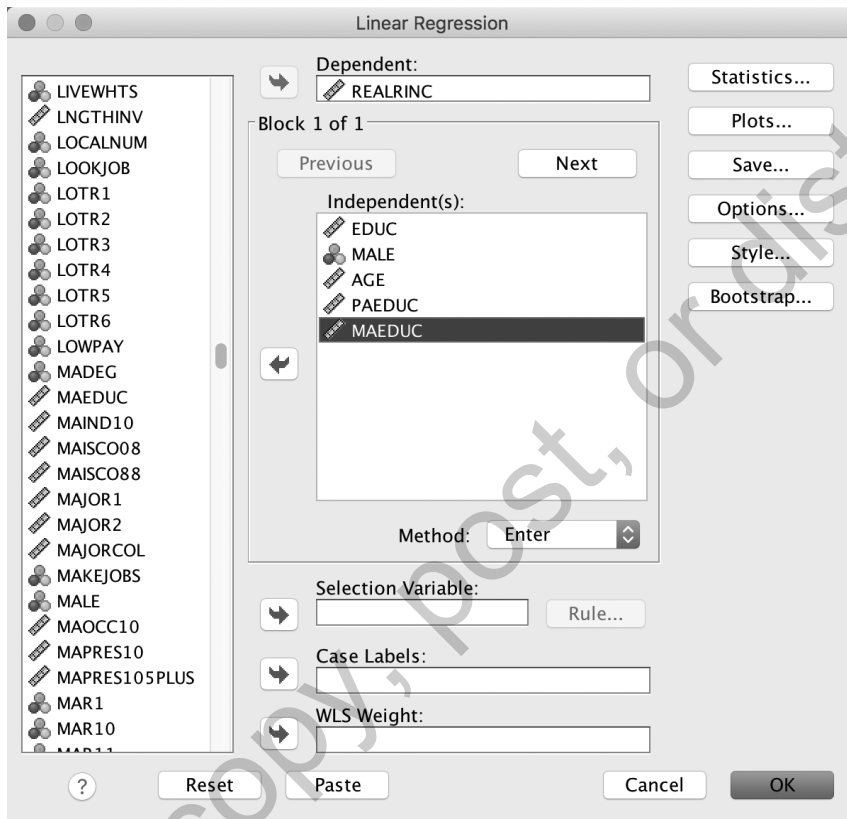
The two graphics that follow show a histogram of the regression standardized residual for the dependent variable and the observed by expected cumulative probability for the dependent variable, real annual income.



Normal P-P Plot of Regression Standardized Residual



It is possible to add additional variables to your linear regression model, such as those in the dialog box featured below. Interval-ratio variables may be included, as well as dummy variables, along with others such as interaction variables. Interaction variables may be computed using the compute function (in the “Transform” menu). More information about computing variables can be found in Chapter 2, “Transforming Variables.” The computation would consist of: Variable 1 \times Variable 2 = Interaction Variable.



 SAGE edge™

Access the full 2016 data file and the 1972–2016 Cumulative Codebook at the student study site: study.sagepub.com/wagner7e.