SPSS: Descriptive and Inferential Statistics

For Windows

August 2012

Table of Contents

This tutorial describes the use of SPSS to obtain descriptive and inferential statistics. In the first section, you will be introduced to procedures used to obtain several descriptive statistics, frequency tables, and crosstabulations. In the second section, the chi-square test of independence, independent and paired sample *t* tests, bivariate correlations, regression, and the general linear model will be covered. If you are not familiar with SPSS or need more information about how to get SPSS to read your data, you may wish to read our **SPSS for Windows: Getting Started** tutorial. This set of documents uses a sample dataset, *Employee data.sav*, that SPSS provides. It can be found in the root SPSS directory. If you installed SPSS in the default location, then this file will be located in the following location: C:\Program Files\SPSS\Employee Data.sav.

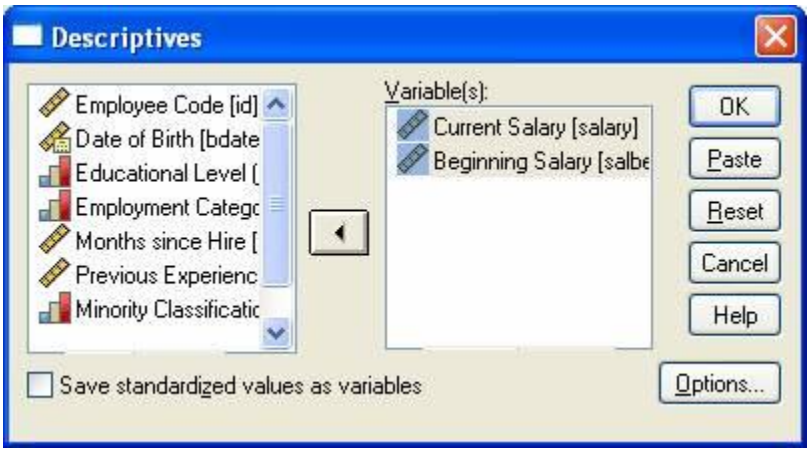
Section 1: Summarizing Data

1.1 Descriptive Statistics

A common first step in data analysis is to summarize information about variables in your dataset, such as the averages and variances of variables. Several summary or descriptive statistics are available under the *Descriptives* option available from the *Analyze* and *Descriptive Statistics* menus:

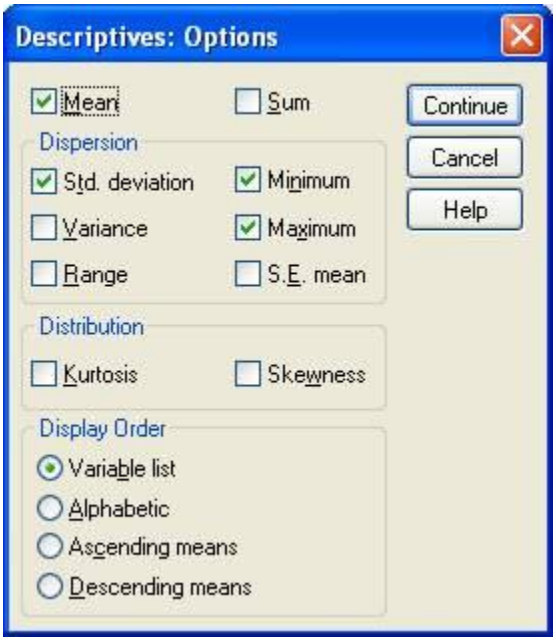
Analyze
Descriptive Statistics
Descriptives...

After selecting the *Descriptives* option, the following dialog box will appear:



This dialog box allows you to select the variables for which descriptive statistics are desired. To select variables, first click on a variable name in the box on the left side of the dialog box, then

click on the arrow button that will move those variables to the *Variable(s)* box. For example, the variables *salbegin* and *salary* have been selected in this manner in the above example. To view the available descriptive statistics, click on the button labeled **Options**. This will produce the following dialog box:



Clicking on the boxes next to the statistics' names will result in these statistics being displayed in the output for this procedure. In the above example, only the default statistics have been selected (mean, standard deviation, minimum, and maximum); however, there are several others that could be selected. After selecting all of the statistics you desire, output can be generated by first clicking on the **Continue** button in the *Options* dialog box, then clicking on the **OK** button in the *Descriptives* dialog box. The statistics that you selected will be printed in the Output Viewer. For example, the selections from the preceding example would produce the following output:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Current Salary	474	$15,750	$135,000	$34,419.57	$17,075.66
Beginning Salary	474	$9,000	$79,980	$17,016.09	$7,870.64
Valid N (listwise)	474				

The number of cases in the dataset is recorded under the column labeled *N*. Information about the range of variables is contained in the *Minimum* and *Maximum* columns. The average salary is contained in the *Mean* column. Variability can be assessed by examining the values in the *Std. Deviation* column. The more that individual data points differ from the mean, the larger the standard deviation will be. Conversely, if there is a great deal of similarity between data points, the standard deviation will be quite small. Examining differences in variability could be useful

for anticipating further analyses: in the above example, it is clear that there is much greater variability in the current salaries than beginning salaries. Because equal variances is an assumption of many inferential statistics, this information is important to a data analyst.

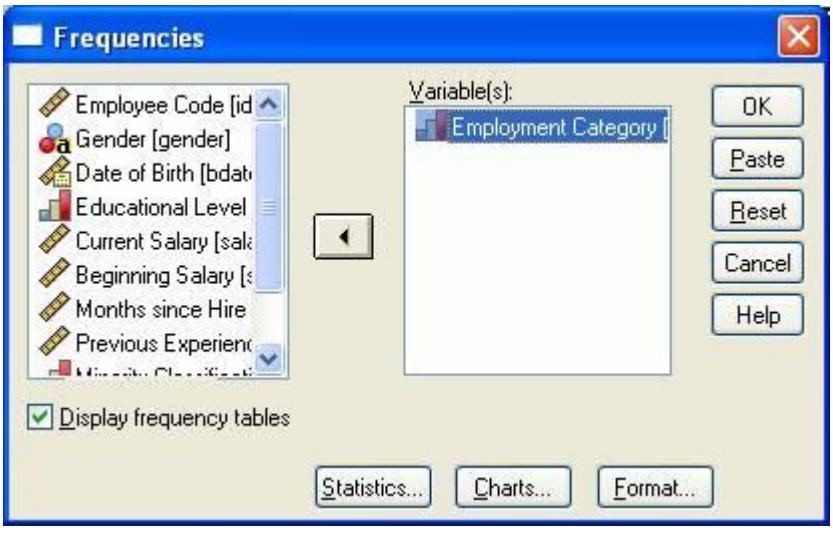
As a side note, if your distribution is "normal," almost all (96%) of your observations should fall within +/- 2 standard deviations from the mean. A starting salary of $32,757.37 is two standard deviations above the mean of $17,016.09, while a starting salary of $1,274.81 is two standard deviations below the mean; accordingly, 96% of salaries *should* fall between these values, with a few people (2%) earning salaries below $1,274.81 and a few (2%) earning salaries above $32,757.37. Given that the minimum value is $9,000 and the maximum is $79,980, however, we can see that these data may not follow the normal distribution. (For the purposes of this tutorial, we will treat salary, educational level, and a number of other variables as though they were normally-distributed continuous variables. In your own research, however, if your outcome variables are not normally distributed, you may need to pursue an alternate analysis. Feel free to direct your questions on this topic to us at stats@its.utexas.edu).

Frequencies

While the descriptive statistics procedure described above is useful for summarizing data with an underlying continuous distribution, the *Descriptives* procedure will not prove helpful for interpreting categorical data. Instead, it is more useful to investigate the numbers of cases that fall into various categories. The *Frequencies* option allows you to obtain the number of people within each employment category in the dataset. The *Frequencies* procedure is found under the *Analyze* menu:

Analyze
Descriptives Statistics
Frequencies...

Selecting this menu item produces the following dialog box:

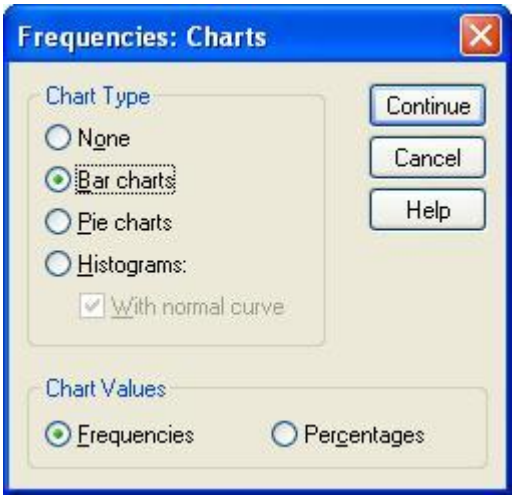


Select variables by clicking on them in the left box, then clicking the arrow in between the two boxes. Frequencies will be obtained for all of the variables in the box labeled *Variable(s)*. This is the only step necessary for obtaining frequency tables; however, there are several other descriptive statistics available, many of which are described in the preceding section. The example in the above dialog box would produce the following output:

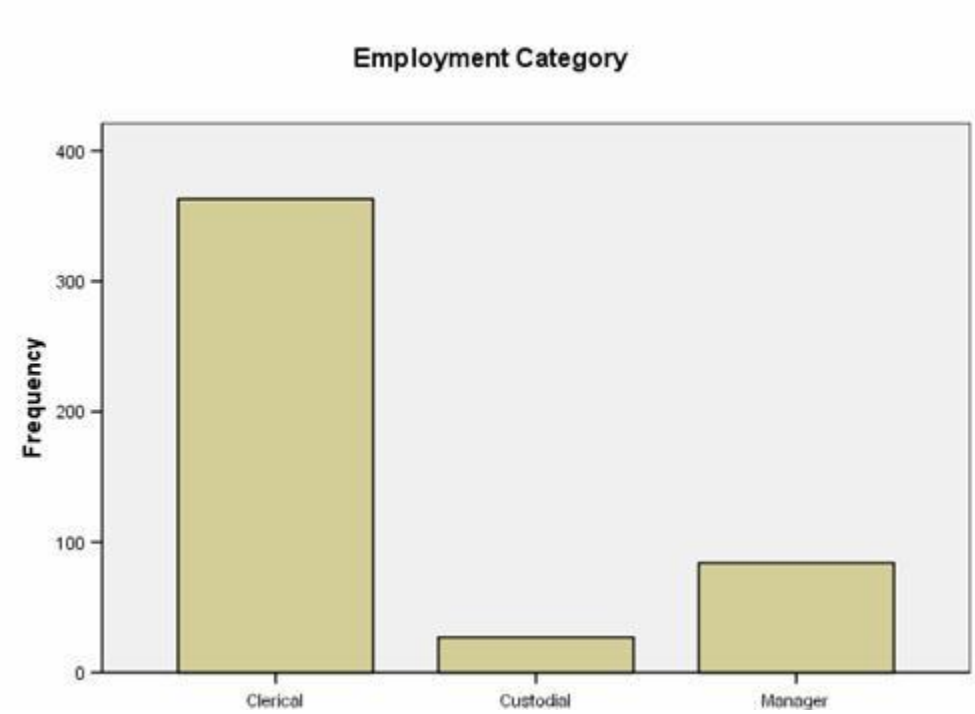
Employment Category

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Clerical	363	76.6	76.6	76.6
	Custodial	27	5.7	5.7	82.3
	Manager	84	17.7	17.7	100.0
	Total	474	100.0	100.0	

Going back to the Frequencies dialog box, you may click on the **Statistics** button to request additional descriptive statistics. Clicking on the **Charts** button produces the following box which allows you to graphically examine their data in several different formats:



Each of the available options provides a visual display of the data. For example, clicking on the *Bar charts* button produces the following output:



If you have continuous data (such as salary) you can also use the *Histograms* option and its suboption, *With normal curve*, to allow you to assess whether your data are normally distributed, which is an assumption of several inferential statistics. (You can also use the *Explore* procedure, available from the *Descriptive Statistics* menu, to obtain the *Kolmogorov-Smirnov test*, which is a hypothesis test to determine if your data are normally distributed.)

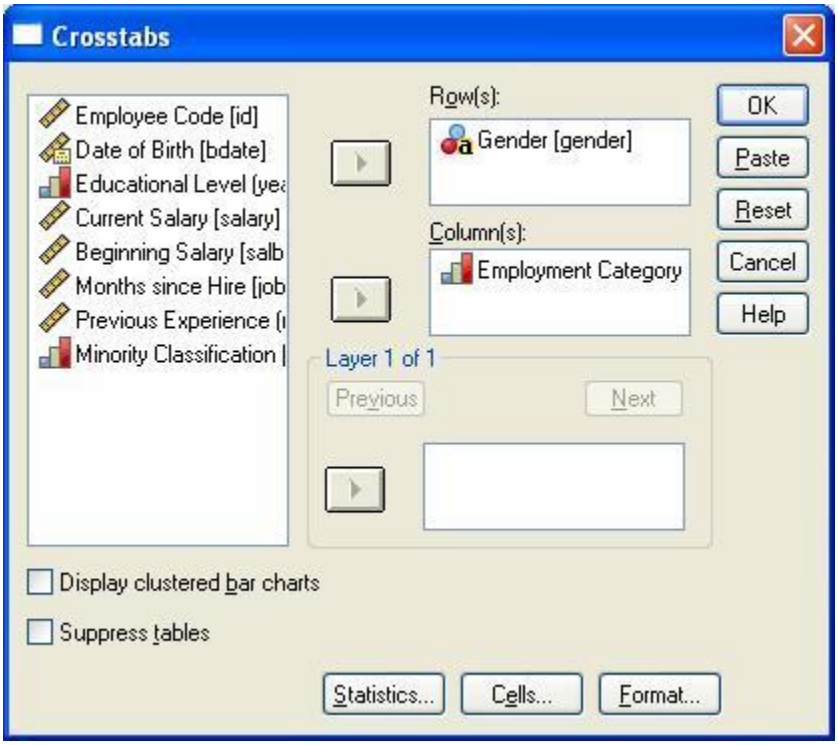
Crosstabulation

While frequencies show the numbers of cases in each level of a categorical variable, they do not give information about the relationship between categorical variables. For example, frequencies can give you the number of men and women in a company AND the number of people in each employment category, but not the number of men and women IN each employment category. The *Crosstabs* procedure is useful for investigating this type of information because it can provide information about the intersection of two variables. The *Crosstabs* procedure is found in the *Analyze* menu:

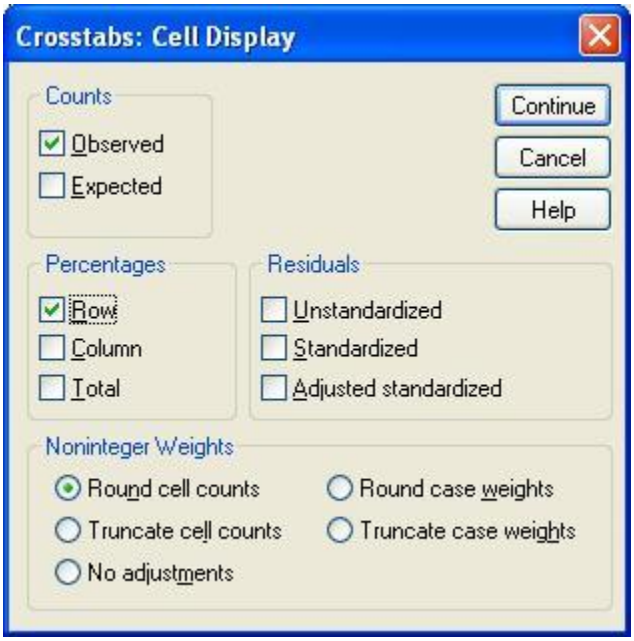
Analyze
Descriptive Statistics
Crosstabs…

After selecting *Crosstabs* from the menu, the dialog box shown above will appear on your monitor. The box on the left side of the dialog box contains a list of all of the variables in the working dataset. If theory suggests that one variable may cause the other, then the causal

variable would typically be placed in the Row, while the outcome variable would be placed in the Column. For example, selecting the variable *gender* for the rows of the table and *jobcat* for the columns would produce a crosstabulation of gender by job category.



The options available by selecting the **Statistics** and **Cells** buttons provide you with several additional output features. Selecting the **Cells** button will produce a menu that allows you to add additional values to your table; it is often most informative to select the Row percentages.



The combination of the two dialog boxes shown above will produce the following output table:

Gender * Employment Category Crosstabulation

			Employment Category			
			Clerical	Custodial	Manager	Total
Gender	Female	Count	206	0	10	216
		% within Gender	95.4%	.0%	4.6%	100.0%
	Male	Count	157	27	74	258
		% within Gender	60.9%	10.5%	28.7%	100.0%
Total		Count	363	27	84	474
		% within Gender	76.6%	5.7%	17.7%	100.0%

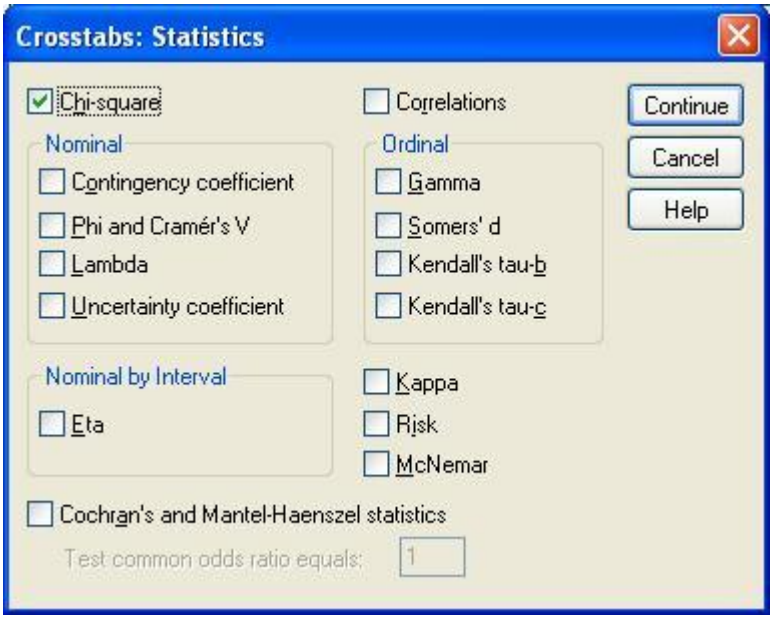
This table shows that 95.4% of females are clerical workers, while only 60.9% of males are clerical workers. It also seems that men are much more likely to be custodians (5.7%) or managers (28.7%) than are women. However, we do not yet know whether this apparent difference is statistically significant.

Section 2: Inferential Statistics

2.1 Chi-Square Test

In the section above, it appeared that there were some differences between men and women in terms of their distribution among the three employment categories. Conducting a *Chi-square test of independence* would tell us if the observed pattern is statistically different from the pattern expected due to chance.

The Chi-square test of independence can be obtained through the *Crosstabs* dialog boxes that were used above to get a crosstabulation of the data. After opening the *Crosstabs* dialog box as described in the preceding section, click the **Statistics** button to get the following dialog box:



By clicking on the box labeled *Chi-Square*, you will obtain the Chi-square test of independence for the variables you have crosstabulated. This will produce the following table in the Output Viewer:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79.277[a]	2	.000
Likelihood Ratio	95.463	2	.000
N of Valid Cases	474		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.30.

Inspecting the table in the previous section, it appears that the two variables, gender and employment category, are related to each other in some way. For example, if gender and employment classification were unrelated, we would expect 17.7% of women to be in the manager classification as opposed to the observed percentage, 4.6%. The output above provides a statistical hypothesis test for the hypothesis that gender and employment category are independent of each other. The large Chi-Square statistic (79.28) and its small significance level ($p < .001$) indicate that it is very unlikely that these variables are independent of each other. Thus, you can conclude that there is a relationship between a person's gender and their employment classification.

2.2 *T* tests

The *t* test is a useful technique for comparing mean values of two sets of numbers. The comparison will provide you with a statistic for evaluating whether the difference between two means is statistically significant. *T* tests can be used either to compare two independent groups (independent-samples *t* test) or to compare observations from two measurement occasions for the same group (paired-samples *t* test). To conduct a *t* test, your outcome data should be a sample drawn from a continuous underlying distribution. If you are using the *t* test to compare two groups, the groups should be randomly drawn from normally distributed and independent populations. For example, if you were comparing clerical and managerial salaries, the *independent populations* are clerks and managers, which are two nonoverlapping groups. If you have more than two groups or more than two variables in a single group that you want to compare, you should use one of the General Linear Model procedures in SPSS, which are described below.

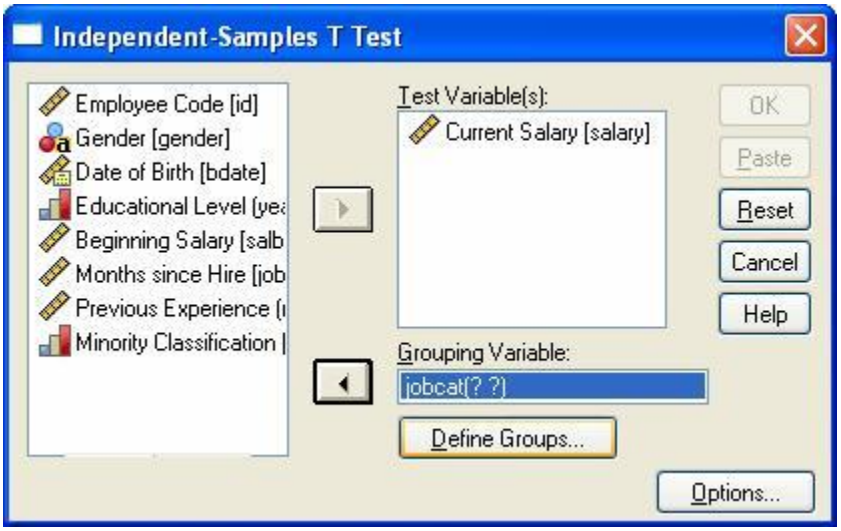
There are three types of *t* tests; the options are all located under the *Analyze* menu item:

Analyze
Compare Means
One-Sample *T* test...
Independent-Samples *T* test...
Paired-Samples *T* test...

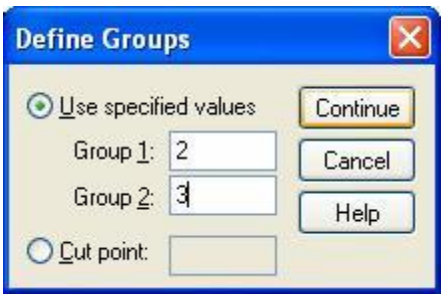
While each of these *t* tests compares mean values of two sets of numbers, they are designed for distinctly different situations:

- The *one-sample t test* is used compare a single sample with a population value. For example, a test could be conducted to compare the average salary of managers within a company with a value that was known to represent the national average for managers.
- The *independent-sample t test* is used to compare two groups' scores on the same variable. For example, it could be used to compare the salaries of clerks and managers to evaluate whether there is a difference in their salaries.
- The *paired-sample t test* is used to compare the means of two variables within a single group. For example, it could be used to see if there is a statistically significant difference between starting salaries and current salaries among the custodial staff in an organization.

To conduct an independent sample *t* test, first select the **Independent-Samples *T* test** option to produce the following dialog box:



To select variables for the analysis, first highlight them by clicking on them in the box on the left. Then move them into the appropriate box on the right by clicking on the arrow button in the center of the box. Your independent variable should go in the *Grouping Variable* box, which is a variable that defines which groups are being compared. For example, because employment categories are being compared in this analysis, the *jobcat* variable is selected. However, because *jobcat* has more than two levels, you will need to click on **Define Groups** to specify the two levels of *jobcat* that you want to compare. This will produce another dialog box as is shown below:



Here, the groups to be compared are limited to the groups with the values 2 and 3, which represent the clerical and managerial groups. After selecting the groups to be compared, click the **Continue** button, and then click the **OK** button in the main dialog box. The above choices will produce the following output:

Group Statistics

	Employment Category	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	Custodial	27	$30,938.89	$2,114.62	$406.96
	Manager	84	$63,977.80	$18,244.78	$1,990.67

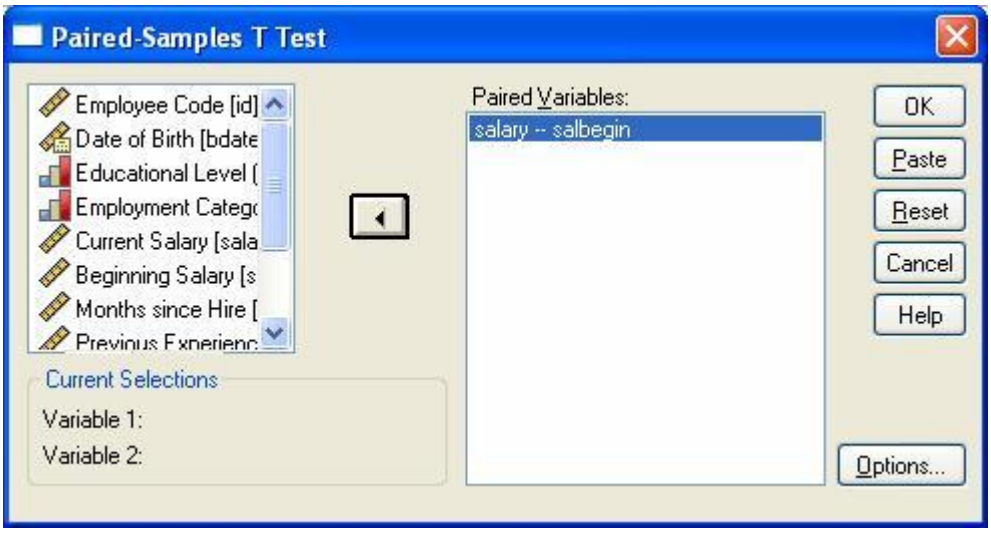
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Current Salary	Equal variances assumed	29.23	.000	-9.36	109.00	.000	-$33,039	$3,530	-$40,034	-26044
	Equal variances not assumed			-16.26	89.58	.000	-$33,039	$2,032	-$37,076	-29002

The first output table, labeled *Group Statistics*, displays descriptive statistics. The second output table, labeled *Independent Samples Test*, contains the statistics that are critical to evaluating the current research question. This table contains two sets of analyses: the first assumes equal variances and the second does not. To assess whether you should use the statistics for equal or unequal variances, use the significance level associated with the value under the heading, *Levene's Test for Equality of Variances*. It tests the hypothesis that the variances of the two groups are equal. A small value (<.05) in the column labeled *Sig.* indicates that this hypothesis is false and that the groups do indeed have unequal variances. In the above case, the value <.05 in that column indicates that the variance of the two groups, clerks and managers, is not equal. Thus, you should use the t-test statistics in the row labeled *Equal variances not assumed*.

The SPSS output reports a *t statistic* and *degrees of freedom* for all *t* test procedures. Every unique value of the *t* statistic and its associated degrees of freedom have a significance value. In the above example in which the hypothesis is that clerks and managers do not differ in their salaries, the *t* statistic under the assumption of unequal variances has a value of -16.26, and the degrees of freedom has a value of 89.58 with an associated significance level of .000. The significance level tells us that the probability that there is no difference between clerical and managerial salaries is very small: specifically, less than one time in a thousand would we obtain a mean difference of $33,038 or larger between these groups if there were really no differences in their salaries.

To obtain a paired-samples *t* test, select **Paired-Samples *T* test** and the following dialog box will appear:



The above example illustrates a *t* test between the variables *salbegin* and *salary*, which represent employees' beginning salary and their current salary. To set up a paired-samples *t* test, click on the two variables that you want to compare. The variable names will appear in the section of the box labeled *Current Selections*. When these variable names appear there, click the arrow in the middle of the dialog box and they will appear in the *Paired Variables* box. Clicking the **OK** button with the above variables selected will produce output for the paired-samples *t* test. The following output is an example of the statistics you would obtain from the above example.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Current Salary - Beginning Salary	$17,403	$10,815	$497	$16,427	$18,380	35.04	473	.000

As with the independent samples *t* test, there is a *t* statistic and degrees of freedom that has a significance level associated with it. The *t* test in this example tests the hypothesis that there is no difference in clerks' beginning and current salaries. The *t* statistic (35.04) and its associated significance level ($p < .001$) indicate that this is not the case. In fact, the observed mean difference of $17,403.48 between beginning and current salaries would occur fewer than once in a thousand times if there really were no difference between clerks' beginning and current salaries.

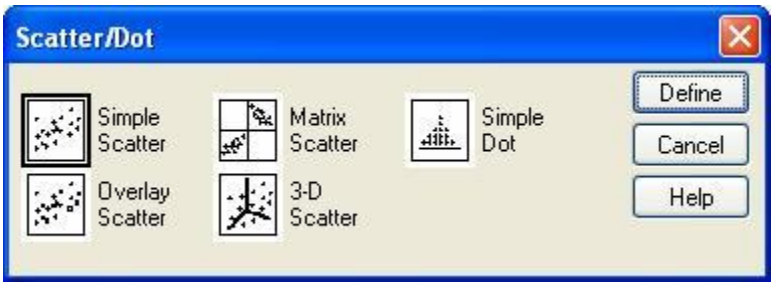
2.3 Correlation

Correlation is one of the most common forms of data analysis because it underlies many other analyses. C*orrelations* measure the linear relationship between two variables. A correlation coefficient has a value ranging from -1 to 1. Values that are closer to the absolute value of 1 indicate that there is a strong relationship between the variables being correlated, whereas values closer to 0 indicate that there is little or no linear relationship. The sign of a correlation coefficient describes the type of relationship between the variables being correlated. A positive correlation coefficient indicates that there is a positive linear relationship between the variables: as one variable increases in value, so does the other. An example of two variables that are likely to be positively correlated are the number of days a student attended class and test grades, because as the number of classes attended increases in value, so do test grades. A negative value indicates a negative linear relationship between variables: as one variable increases in value, the other variable decreases in value. The number of days students miss class and their test scores are likely to be negatively correlated because as the number of days of missed classed increases, test scores typically decrease.

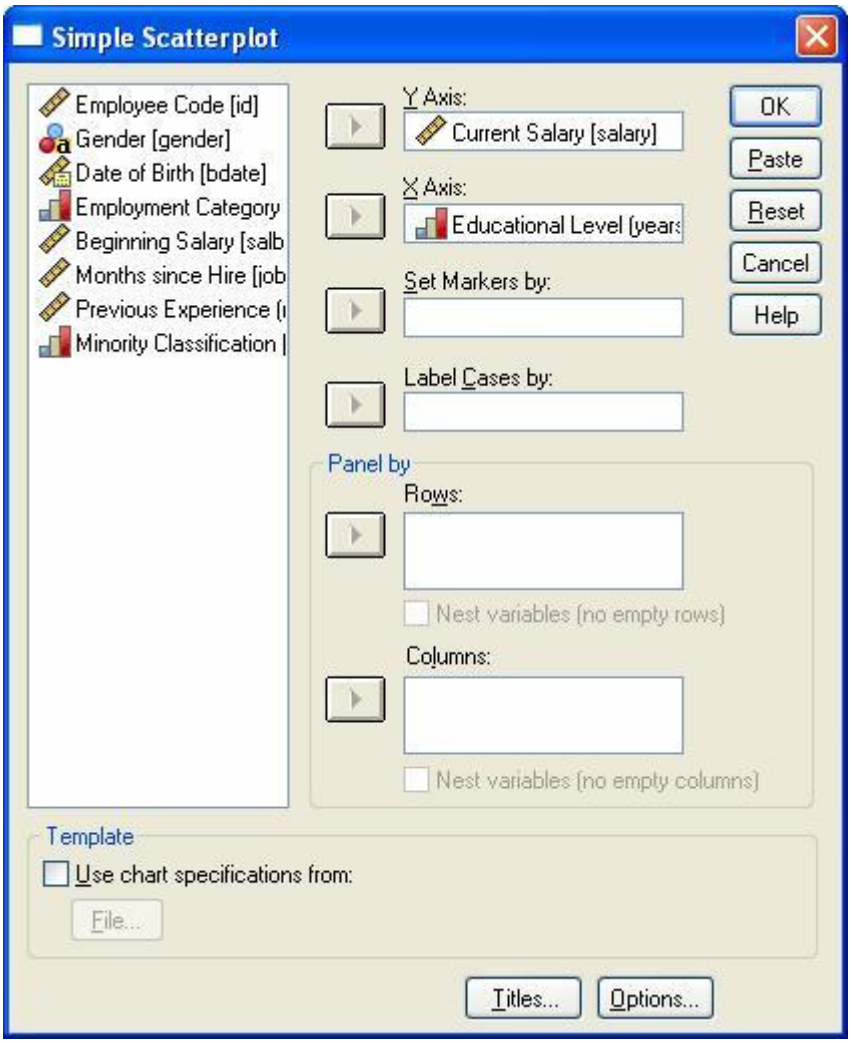
Prior to conducting a correlation analysis, it is advisable to plot the two variables to visually inspect the relationship between them. To produce scatter plots, select the following menu option:

Graphs
Scatter/Dot...

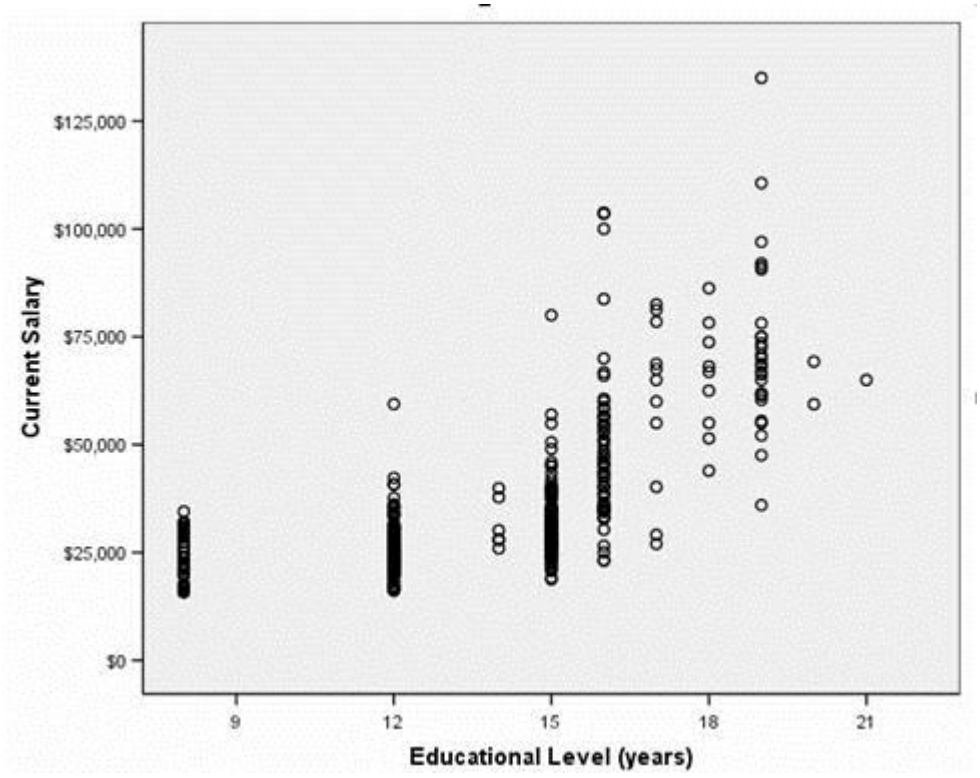
This will produce the following dialog box:



Although several plot options are listed, only the Simple Scatter method is discussed here. To create a scatter plot of current salary by education, select the *Simple Scatter* option and then click the **Define** button to produce the following dialog box:



If one variable can be theoretically conceptualized as causing the other, then the causal variable would typically be placed on the X axis, and the outcome variable on the Y axis. Click on the current salary variable to select it from the list on the left side of the dialog box and then click the right arrow next to the *Y Axis* box to move it over to the right. Next, click on the education variable and then click on the right arrow next to the *X Axis* box to move it over to this box. Finally, click the **OK** button to create the output shown below:

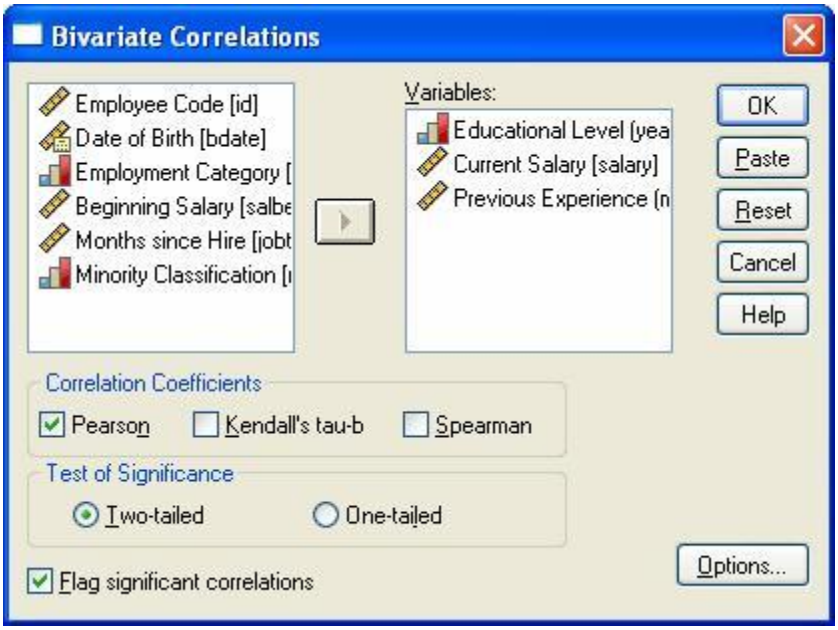


This plot indicates that there is a positive, fairly linear relationship between current salary and education level. In order to test whether this apparent relationship is statistically significant, we could run a correlation.

To obtain a correlation in SPSS, start at the *Analyze* menu. Select the *Correlate* option from this menu. By selecting this menu item, you will see that there are three options for correlating variables: (1) *Bivariate*, (2) *Partial*, and (3) *Distances.* This document will cover only *bivariate correlations*, which is used in situations where you are interested only in the relationship between two variables. To obtain a bivariate correlation, choose the following menu option:

Analyze
Correlate
Bivariate...

This will produce the following dialog box:



To obtain correlations, first click on the variable names in the variable list on the left side of the dialog box. Next, click on the arrow between the two white boxes, which will move the selected variables into the *Variables* box. Each variable listed in the *Variables* box will be correlated with every other variable in the box. For example, with the above selections, we would obtain correlations between *Education Level* and *Current Salary*, between *Education Level* and *Previous Experience*, and between *Current Salary* and *Previous Experience*. We will maintain the default options shown in the above dialog box in this example. The first option to consider is the type of correlation coefficient. Pearson's is appropriate for continuous data as noted in the above example, whereas the other two correlation coefficients, Kendall's tau-b and Spearman's, are designed for ranked data. The choice between a one- and two-tailed significance test in the *Test of Significance* box should be determined by whether the hypothesis you are testing is making a prediction about the direction of effect between the two variables: if you are making a prediction that there is a negative or positive relationship between the variables, then the one-tailed test is appropriate; if you are not making a directional prediction, you should use the two-tailed test. The selections in the above dialog box will produce the following output:

Correlations

		Educational Level (years)	Current Salary	Previous Experience (months)
Educational Level (years)	Pearson Correlation	1.000	.661**	-.252**
	Sig. (2-tailed)	.	.000	.000
	N	474	474	474
Current Salary	Pearson Correlation	.661**	1.000	-.097*
	Sig. (2-tailed)	.000	.	.034
	N	474	474	474
Previous Experience (months)	Pearson Correlation	-.252**	-.097*	1.000
	Sig. (2-tailed)	.000	.034	.
	N	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

This output gives us a correlation matrix for the three correlations requested in the above dialog box. Note that despite there being nine cells in the above matrix, there are only three correlation coefficients of interest: (1) the correlation between current salary and educational level, the correlation between previous experience and educational level, and the correlation between current salary and previous experience. The reason only three of the nine correlations are of interest is because the diagonal consists of correlations of each variable with itself, always resulting in a value of 1.00 and the values on each side of the diagonal replicate the values on the opposite side of the diagonal. For example, the three unique correlation coefficients show there is a positive correlation between employees' number of years of education and their current salary. This positive correlation coefficient (.661) indicates that there is a statistically significant ($p <$ .001) linear relationship between these two variables such that the more education a person has, the larger that person's salary is. Also observe that there is a statistically significant ($p < .001$) negative correlation coefficient (-.252) for the association between education level and previous experience, indicating that the linear relationship between these two variables is one in which the values of one variable decrease as the other increases. The third correlation coefficient (-.097) also indicates a negative association between employee's current salaries and their previous work experience, although this correlation is fairly weak.

2.4 Regression

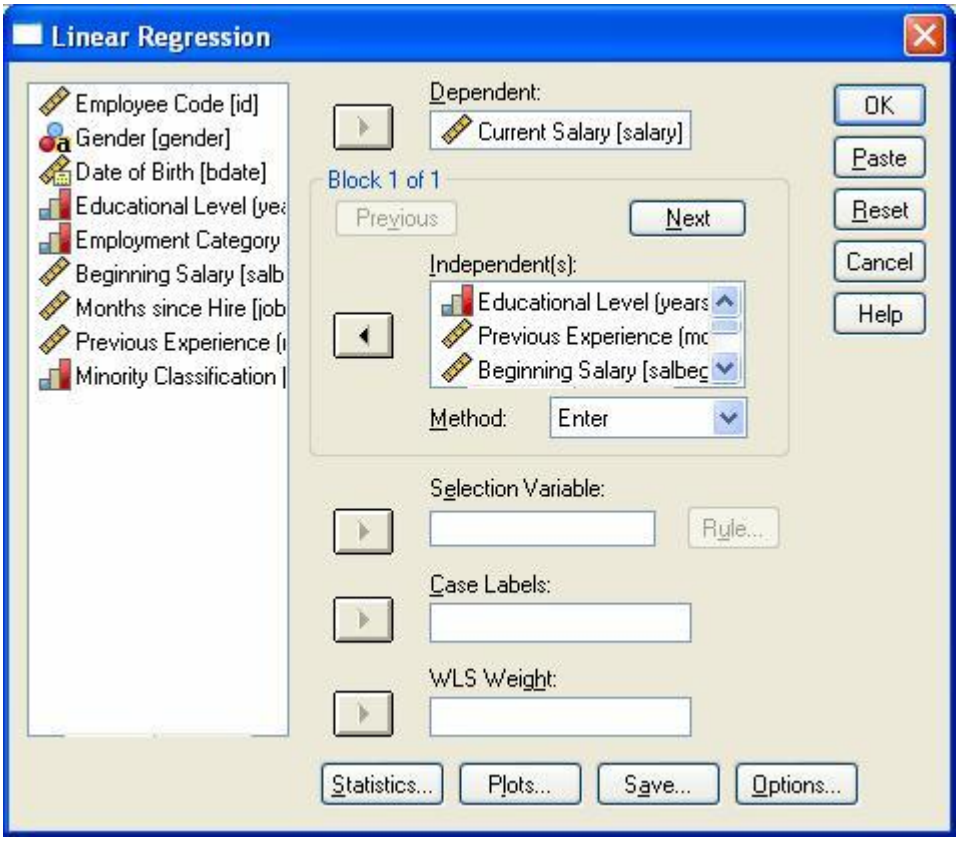
Regression is a technique that can be used to investigate the effect of one or more predictor variables on an outcome variable. Regression allows you to make statements about how well one or more independent variables will predict the value of a dependent variable. For example, if you were interested in investigating which variables in the employee database were good predictors of employees' current salaries, you could create a regression equation that would use several of the variables in the dataset to predict employees' salaries. By doing this you will be able to make

statements about whether variables such as employees' number of years of education, their starting salary, or their number of months on the job are good predictors of their current salaries.

To conduct a regression analysis, select the following from the *Analyze* menu:

Analyze
Regression
Linear...

This will produce the following dialog box:



This dialog box illustrates an example regression equation. As with other analyses, you select variables from the box on the left by clicking on them, then moving them to the boxes on the right by clicking the arrow next to the box where you want to enter a particular variable. Here, employees' current salary has been entered as the dependent variable. In the *Independent(s)* box, four predictor variables have been entered: educational level, previous experience, beginning salary, and months since hire.

NOTE: Before you run a regression model, you should consider the method that you use for selecting or rejecting variables in that model. The box labeled *Method* allows you to select from one of five methods: *Enter*, *Remove*, *Forward*, *Backward*, and *Stepwise*. Unfortunately, we cannot offer a comprehensive discussion of the characteristics of each of these methods here, but

you have several options regarding the method you use to remove and retain predictor variables in your regression equation. In this example, we will use the SPSS default method, *Enter*, which is a standard approach in regression models. If you have questions about which method is most appropriate for your data analysis, consult a regression text book, the SPSS help facilities, or contact a consultant.

The following output assumes that only the default options have been requested. If you have selected options from the *Statistics*, *Plots*, or *Options* boxes, then you will have more output than is shown below and some of your tables may contain additional statistics not shown here.

The first table in the output, shown below, includes information about the quantity of variance that is explained by your predictor variables. The first statistic, *R*, is the multiple correlation coefficient between all of the predictor variables and the dependent variable. In this model, the value is .90, which indicates that there is a great deal of variance shared by the independent variables and the dependent variables. The next value, *R* Square, is simply the squared value of *R*. This is frequently used to describe the goodness-of-fit or the amount of variance explained by a given set of predictor variables. In this example, the value is .81, which indicates that 81% of the variance in the dependent variable is explained by the independent variables in the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.900[a]	.810	.809	$7,465.14

a. Predictors: (Constant), Months since Hire, Previous Experience (months), Beginning Salary, Educational Level (years)

The second table in the output is an ANOVA table that describes the overall variance accounted for in the model. The *F* statistic represents a test of the null hypothesis that the expected values of the regression coefficients are equal to each other and that they equal zero. Put another way, this *F* statistic tests whether the R square proportion of variance in the dependent variable accounted for by the predictors is zero. If the null hypothesis were true, then that would indicate that there is not a regression relationship between the dependent variable and the predictor variables. But, instead, it appears that the four predictor variables in the present example are not all equal to each other and could be used to predict the dependent variable, current salary, as is indicated by a large *F* value and a small significance level.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	111779919524.3	4	27944979881.07	501.450	.000[a]
	Residual	26136575912.07	469	55728306.849		
	Total	137916495436.3	473			

a. Predictors: (Constant), Educational Level (years), Months since Hire, Previous Experience (months), Beginning Salary

b. Dependent Variable: Current Salary

The third table in standard regression output provides information about the effects of individual predictor variables. Generally, there are two types of information in the *Coefficients* table: coefficients and significance tests. The unstandardized coefficients indicate the increase in the value of the dependent variable for each unit increase in the predictor variable. For example, the unstandardized coefficient for *Educational Level* in the example is 669.91, which indicates to us that for each year of education, a person's predicted salary will increase by $669.91. A well-known problem with the interpretation of unstandardized coefficients is that their values are dependent on the scale of the variable for which they were calculated, which makes it difficult to assess the relative influence of independent variables through a comparison of unstandardized coefficients. For example, comparing the unstandardized coefficient of *Education Level*, 669.91, with the unstandardized coefficient of the variable *Beginning Salary*, 1.77, it could appear that Educational Level is a greater predictor of a person's current salary than is *Beginning Salary*. We can see that this is deceiving, however, if we examine the standardized coefficients, or *Beta coefficients*. Beta coefficients are based on data expressed in standardized, or *z* score form. Thus, all variables have a mean of zero and a standard deviation of one and are thus expressed in the same units of measurement. Examining the Beta coefficients for *Education Level* and *Beginning Salary*, we can see that when these two variables are expressed in the same scale, *Beginning Salary* is more obviously the better predictor of *Current Salary*.

In addition to the coefficients, the table also provides a significance test for each of the independent variables in the model. The significance test evaluates the null hypothesis that the unstandardized regression coefficient for the predictor is zero when all other predictors' coefficients are fixed to zero. This test is presented as a *t* statistic. For example, examining the *t* statistic for the variable, *Months Since Hire*, you can see that it is associated with a significance value of .000, indicating that the null hypothesis, that states that this variable's regression coefficient is zero when all other predictor coefficients are fixed to zero, can be rejected.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-16149.7	3255.470		-4.961	.000
	Months since Hire	161.486	34.246	.095	4.715	.000
	Previous Experience (months)	-17.303	3.528	-.106	-4.904	.000
	Beginning Salary	1.768	.059	.815	30.111	.000
	Educational Level (years)	669.914	165.596	.113	4.045	.000

a. Dependent Variable: Current Salary

2.5 General Linear Model

The majority of procedures used for conducting analysis of variance (ANOVA) in SPSS can be found under the *General Linear Model* (GLM) menu item in the *Analyze* menu. Analysis of variance can be used in many situations to determine whether there are differences between groups on the basis of one or more outcome variables or if a continuous variable is a good predictor of one or more dependent variables. There are three varieties of the general linear model available in SPSS: univariate, multivariate, and repeated measures. The *univariate general linear model* is used in situations where you only have a single dependent variable, but may have several independent variables that can be fixed between-subjects factors, random between-subjects factors, or covariates. The *multivariate general linear model* is used in situations where there is more than one dependent variable and independent variables are either fixed between-subjects factors or covariates. The *repeated measures general linear model is* used in situations where you have more than one measurement occasion for a dependent variable and have fixed between-subjects factors or covariates as independent variables. Because it is beyond the scope of this document to cover all three varieties of the general linear model in detail, we will focus on the univariate version of the general linear model with some attention given to topics that are unique to the repeated measures general linear model. Several features of the univariate general linear model are useful for understanding other varieties of the model that are provided in SPSS: understanding the univariate model will prove useful for understanding other GLM options.

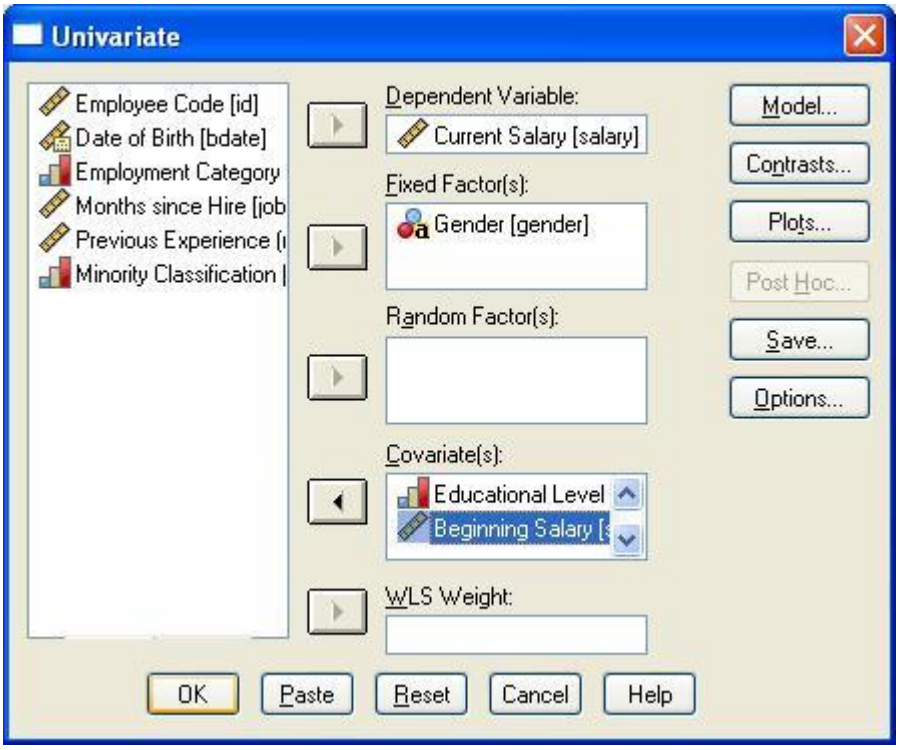
The univariate general linear model is used to compare differences between group means and estimating the effect of covariates on a single dependent variable. For example, you may want to see if there are differences between men and women's salaries in a sample of employee data. To do this, you would want to demonstrate that the average salary is significantly different between men and women. However, in doing such an analysis, you are likely aware that there are other factors that could affect a person's salary that need to be controlled for in such an analysis. For example, educational background and starting salary are some such variables. By including these

variables in our analysis, you will be able to evaluate the differences between men and women's salaries while controlling for the influence of these other variables.

To specify a univariate general linear model in SPSS, go to the analyze menu and select univariate from the general linear model menu:

Analyze
General Linear Model
Univariate...

This will produce the following dialog box:



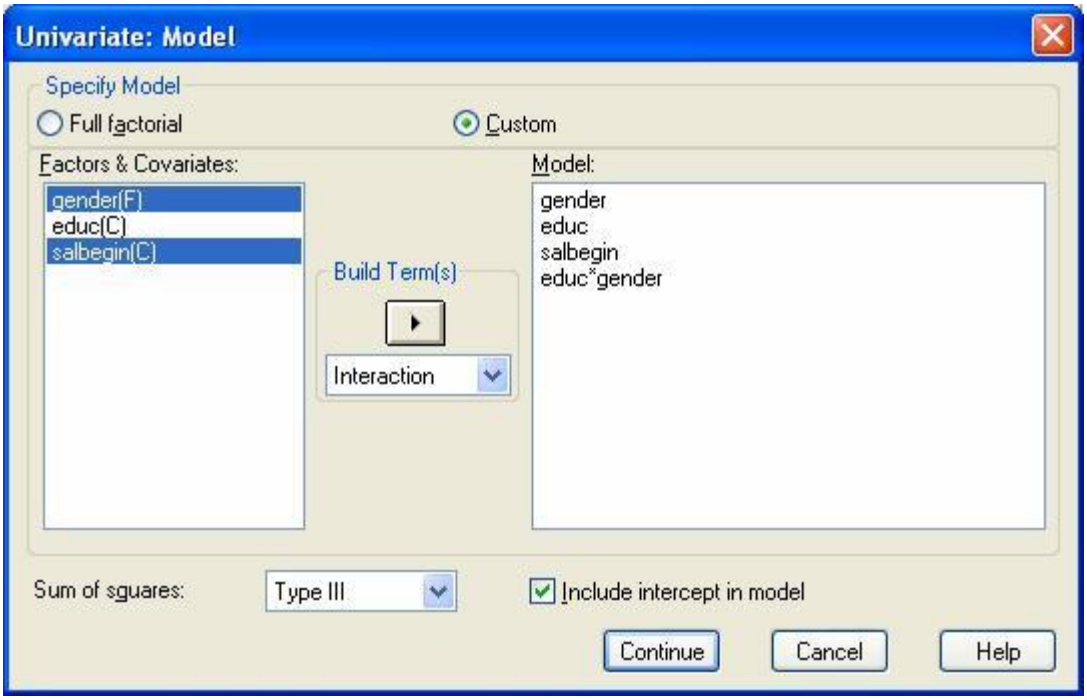
The above box demonstrates a model with multiple types of independent variables. The variable gender has been designated as a *fixed factor* because it contains all of the levels of interest.

In contrast, *random variables* are variables that represent a random sample of the possible levels that could be sampled. There are not any true random variables in our dataset; therefore, this input box has been left blank here. However, you could imagine a situation similar to the above example where you sampled data from multiple corporations for our employee database. In that case, you would have introduced a random variable into the model--the corporation to which an employee belongs. Corporation is a random factor because you would only be sampling a few of the many possible corporations to which you would want to generalize your results. (It should be noted, however, that the GLM method for handling random factors is inferior to the method used in the Mixed Models procedure.)

The next input box contains the covariates in your model. A *covariate* is a quantitative independent variable. Covariates are often entered in models to reduce error variance: by removing the effects of the relationship between the covariate and the dependent variable, you can often get a better estimate of the amount of variance that is being accounted for by the factors in the model. Covariates can also be used to measure the linear association between the covariate and a dependent variable, as is done in regression models. In this situation, a linear relationship indicates that the dependent variable increases or decreases in value as the covariate increases or decreases in value.

The box labeled *WLS Weight* can contain a variable that is used to weight other variables in a weighted least-squares analysis. This procedure is infrequently used however, and is not discussed in any detail here.

The default model for the SPSS univariate GLM will include main effects for all independent variables and will provide interaction terms for all possible combinations of fixed and random factors. You may not want this default model, or you may want to create interaction terms between your covariates and some of the factors. In fact, if you intend to conduct an analysis of covariance, you should test for interactions between covariates and factors. Doing so will determine whether you have met the *homogeneity of regression slopes* assumption, which states that the regression slopes for all groups in your analysis are equal. This assumption is important because the means for each group are adjusted by averaging the slopes for each group so that group differences in the covariate are removed from the dependent variable. Thus, it is assumed that the relationship between the covariate and the dependent variable is the same at all levels of the independent variables. To make changes in the default model, click on the **Model** button, which will produce the following dialog box:



The first step for modifying the default model is to click on the button labeled *Custom,* to activate the grayed out areas of the dialog box. At this point, you can begin to move variables in the *Factors & Covariates* box into the *Model* box. First, move all of the main effects into the *Model* box. The quickest way to do that is to double-click on their names in the *Factors & Covariates* box. After entering all of the main effects, you can begin building interaction terms. To build the interactions, click on the arrow facing downwards in the *Build Term(s)* section and select interaction, as shown in the figure above. After you have selected the interaction, you can click on the names of the variables with which you would like to build an interaction, then click on the arrow facing right under the *Build Term(s)* heading. In the above example, the *educ*gender* term has already been created. The *salbegin*gender* and *salbegin*educ* terms can be created by highlighting two terms at a time as shown above, then clicking on the right-facing arrow. Some of the other options in the *Build Terms* list that you may find useful are the *All n-way* options. For example if you highlighted all three variables in the Factors & Covariates box, you could create all of the three possible 2-way interactions by selecting the *All 2-way* option from the *Build Terms(s)* drop-down menu, then clicking the right-facing arrow.

If you are testing the homogeneity of regression slopes assumption, you should examine your group by covariate interactions, as well as any covariate by covariate interactions. In order to meet the ANCOVA assumption, these interactions should not be significant. Examining the output from the example above, we expect to see nonsignificant effects for the *gender*educ* and the *gender*salbegin* interaction effects:

Tests of Between-Subjects Effects

Dependent Variable: Current Salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	109713955893[a]	6	18285659315.6	302.788	.000
Intercept	1153099.423	1	1153099.423	.019	.890
GENDER	1035163.021	1	1035163.021	.017	.896
SALBEGIN	435965029.184	1	435965029.184	7.219	.007
EDUC	80521627.420	1	80521627.420	1.333	.249
GENDER * SALBEGIN	39870280.811	1	39870280.811	.660	.417
GENDER * EDUC	48780222.556	1	48780222.556	.808	.369
SALBEGIN * EDUC	90179603.501	1	90179603.501	1.493	.222
Error	28202539543.0	467	60390876.966		
Total	699467436925	474			
Corrected Total	137916495436	473			

a. R Squared = .796 (Adjusted R Squared = .793)

Examining the group by covariate effects, you can see that both were nonsignificant. The *gender*salbegin* effect has a small *F* statistic (.660) and a large significance value (.417), the *educ*salbegin* effect also has a small *F* statistic (1.808) and large significance value (.369), and the *salbegin*educ* effect also has a small *F* statistic (1.493) and large significance level (.222).

Because all of these significance levels are greater than .05, the homogeneity of regression assumption has been met and you can proceed with the ANCOVA.

Knowing that the model does not violate the homogeneity of regression slopes assumption, you can remove the interaction terms from the model by returning to the *GLM Univariate* dialog box, clicking the **Model** button, and selecting *Full Factorial*. This will return the model to its default form in which there are no interactions with covariates. After you have done this, click **OK** in the *GLM Univariate* dialog box to produce the following output:

Tests of Between-Subjects Effects

Dependent Variable: Current Salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	109520439289[a]	3	36506813096.4	604.246	.000
Intercept	842156178.459	1	842156178.459	13.939	.000
EDUC	2287985765.92	1	2287985765.92	37.870	.000
SALBEGIN	42057263689.0	1	42057263689.0	696.115	.000
GENDER	234049991.723	1	234049991.723	3.874	.050
Error	28396056147.3	470	60417140.739		
Total	699467436925	474			
Corrected Total	137916495436	473			

a. R Squared = .794 (Adjusted R Squared = .793)

The default output for the univariate general linear model contains all main effects and interactions between fixed factors. The above output contains no interactions because gender is the only fixed factor. Each factor, covariate, or other source of variance is listed in the left column. For each source of variance, there are several test statistics. To evaluate the influence of each independent variable, look at the *F* statistic and its associated significance level. Examining the first covariate, education level, the *F* statistic (37.87) and its associated significance level (.000) indicate that it has a significant linear relationship with the dependent variable. The second covariate, *salbegin*, also has a significant *F* statistic (696.12) as can be seen from its associated significance level (.000). In both cases, this indicates that the values of the dependent variable, *salary*, increase as the values of education level and beginning salaries increase. The next source of variance, gender, provides us with a test of the null hypothesis that there are no differences between gender groups (i.e., that there are not differences between men and women's salaries) when education and beginning salary are controlled. This test provides a small *F* statistic (3.87) and a significance level that is just barely statistically significant ($p = .05$). In the above model containing education level and beginning salaries as covariates, we can say that there is a statistically significant difference between men and women's salaries. To further interpret this effect, go back to the *Univariate* dialog box and click on the *Options* button. Choose the variable gender and click it over to the *Display Means For* box, then choose *Continue* and *OK*. This will present the estimated means for each group, taking into account the covariates educational level and beginning salary, as shown below. Note that after setting both

groups equal to the mean educational and beginning salary level, men in this company earn approximately $1,590 more than women in current salary. It is possible that the inclusion of other covariates could explain away the remaining difference between the two groups' salaries.

Gender

Dependent Variable: Current Salary

Gender	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Female	33552.223[a]	567.147	32437.765	34666.681
Male	35145.717[a]	513.397	34136.879	36154.554

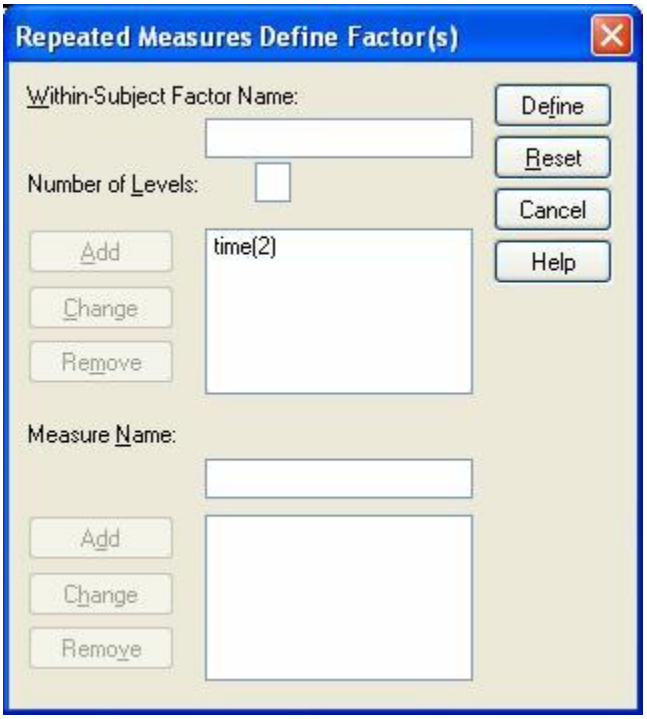
a. Covariates appearing in the model are evaluated at the following values: Educational Level (years) = 13.49, Beginning Salary = $17,016.09.

The repeated measures version of the general linear model has many similarities to the univariate model described above. However, the key difference between the models is that there are multiple measurement occasions of the dependent variable in repeated measures models, whereas the univariate model only permits a single dependent variable. You could conduct a similar model with repeated measurements by using beginning salaries and current salaries as the repeated measurement occasions.

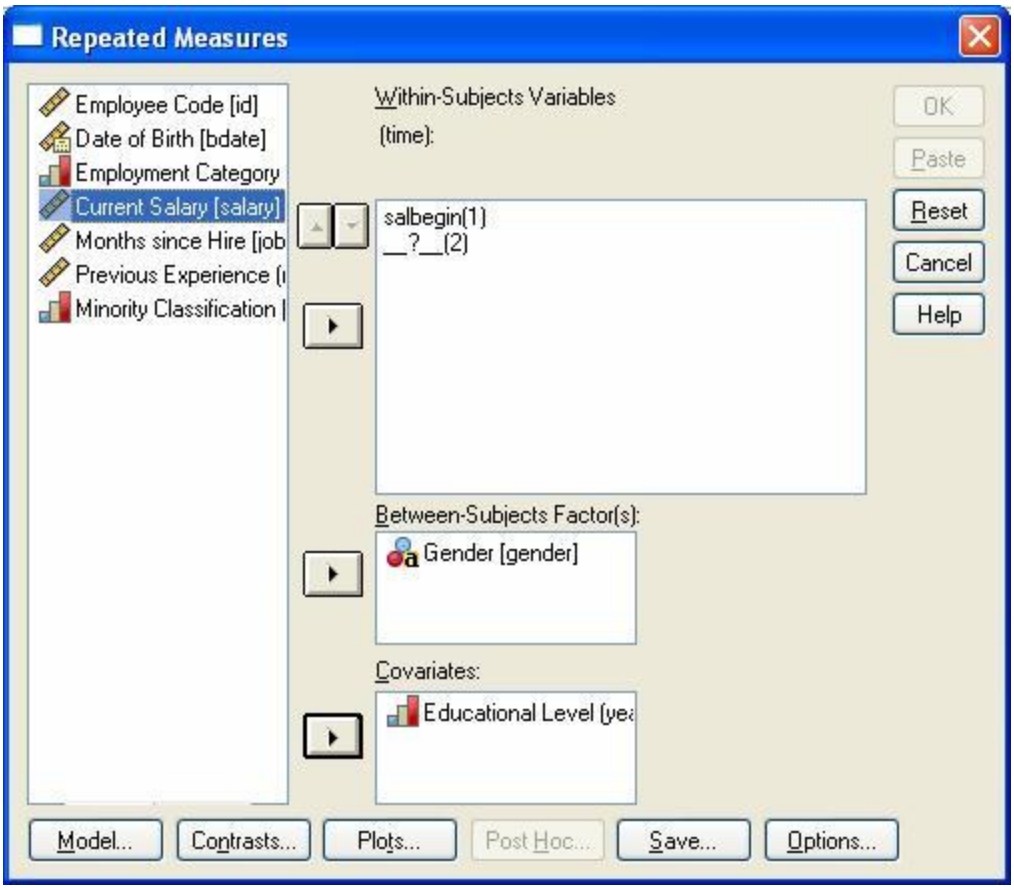
To conduct this analysis, you should select the *Repeated Measures* option from the *General Linear Model* submenu of the *Analyze* menu:

Analyze
General Linear Model
Repeated Measures...

Selecting this option will produce the following dialog box:



This dialog box is used for defining the repeated measures, or within-subjects, dependent variables. You first give the within-subject factor a name in the box labeled *Within-Subject Factor Name*. This name should be something that describes the dependent variables you are grouping together. For example, in this dialog box, the salaries that are being analyzed differ in terms of the time in which the salary data were collected, so the within-subject factor was given the name *time*. Next, specify the number of levels, or number of measurement occasions, in the box labeled *Number of Levels*. This is the number of times the dependent variable was measured. Thus, in the present example, there are two measurement occasions for salary because you are measuring beginning salaries and current salaries. After you have filled in the *Within-Subject Factor Name* and the *Number of Levels* input boxes, click the **Add** button to transfer the information in the input boxes into the box below. Repeat this process until you have specified all of your within-subject factors. Then, click on the **Define** button, and the following dialog box will appear:



When this box initially appears, you will see a slot for each level of the within-subject factor variables that you specified in the previous dialog box. These slots are labeled numerically for each level of the within-subject factor but do not contain variable names. You still need to specify which variable fills each slot of the within-subject factors. To do this, click the appropriate variable name in the list on the left side of the dialog box. Next, click on the arrow pointing towards the *Within-Subject Variables* dialog box to move the variable name from the list to the top slot in the within-subjects box. This process has been completed for *salbegin*, the first level of the *salaries* within-subject factor. The same process should be repeated for *salary*, the variable representing an employee's current salary.

After you have completed the specifications for the within-subjects factors, you can define your independent variables. Between-subject factors, or fixed factors, should be moved into the box labeled *Between-Subjects Factors(s)* by first clicking on the variable name in the variable list, then clicking on the arrow to the left of the *Between-Subjects Factor(s)* box. In this example, gender has been selected as a between-subjects factor. Covariates, or continuous predictor variables, can be moved into the *Covariates* box. Above, *educ*, the variable representing employee's number of years of education, has been specified as a covariate.

This will produce several output tables, but we will focus here on the tables describing between-subject and within-subject effects. However, these tables for univariate analysis of variance may not always be the appropriate. The univariate tests have an additional assumption: the

assumption of sphericity. If this assumption is violated, you should use the multivariate output or adjust your results using one of the correction factors in the SPSS output. For a more detailed discussion of this topic, see the usage note, *Repeated Measures ANOVA Using SPSS MANOVA* in the section, "Within-Subjects Tests: The Univariate versus the Multivariate Approach." This usage note can be found at http://www.utexas.edu/cc/docs/stat38.html.

The following output contains the statistics for the effects in the model specified in the above dialog boxes:

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
time	Sphericity Assumed	697223511	1	697223511.2	18.915	.000
	Greenhouse-Geisser	697223511	1.000	697223511.2	18.915	.000
	Huynh-Feldt	697223511	1.000	697223511.2	18.915	.000
	Lower-bound	697223511	1.000	697223511.2	18.915	.000
time * educ	Sphericity Assumed	6345201343	1	6345201343	172.136	.000
	Greenhouse-Geisser	6345201343	1.000	6345201343	172.136	.000
	Huynh-Feldt	6345201343	1.000	6345201343	172.136	.000
	Lower-bound	6345201343	1.000	6345201343	172.136	.000
time * gender	Sphericity Assumed	924057793	1	924057792.9	25.068	.000
	Greenhouse-Geisser	924057793	1.000	924057792.9	25.068	.000
	Huynh-Feldt	924057793	1.000	924057792.9	25.068	.000
	Lower-bound	924057793	1.000	924057792.9	25.068	.000
Error(time)	Sphericity Assumed	1.736E+010	471	36861538.99		
	Greenhouse-Geisser	1.736E+010	471.000	36861538.99		
	Huynh-Feldt	1.736E+010	471.000	36861538.99		
	Lower-bound	1.736E+010	471.000	36861538.99		

This table contains information about the within-subject factor, *time*, and its interactions with the independent variables. The main effect for salaries is a test of the null hypothesis that all levels of within-subjects factor are equal, or, more specifically, it is a test of the hypothesis that beginning and current salaries are equal. The *F* statistic (18.915) and its associated significance level ($p < .001$) indicate that you can reject this hypothesis as false. In other words, it appears that there is a statistically significant difference between beginning salaries and current salaries. After you have tested this hypothesis, you can then investigate whether the increase in salaries is the same across all values or levels of the other independent variables that are included in the model. The first interaction term in the table tests the hypothesis that the increase in salaries is constant, regardless of educational background. The *F* statistic (172.10) and its associated significance level ($p < .001$) allow us to reject this hypothesis as well. The knowledge that this

interaction is significant indicates that it is worthwhile to examine characteristics of the interaction, which we will do using plots in a later example. Finally, the second interaction tests whether salary increase differs by gender. The *F* statistic (25.07) and its significance level ($p < .001$) indicate that the increase in salaries does vary by gender.

Tests of Between-Subjects Effects

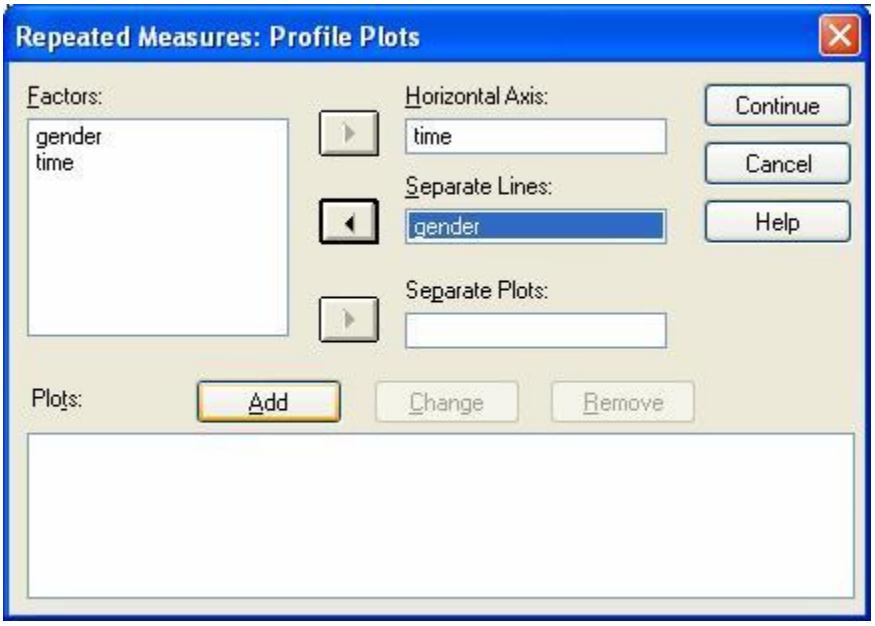
Measure: MEASURE_1

Transformed Variable: Average

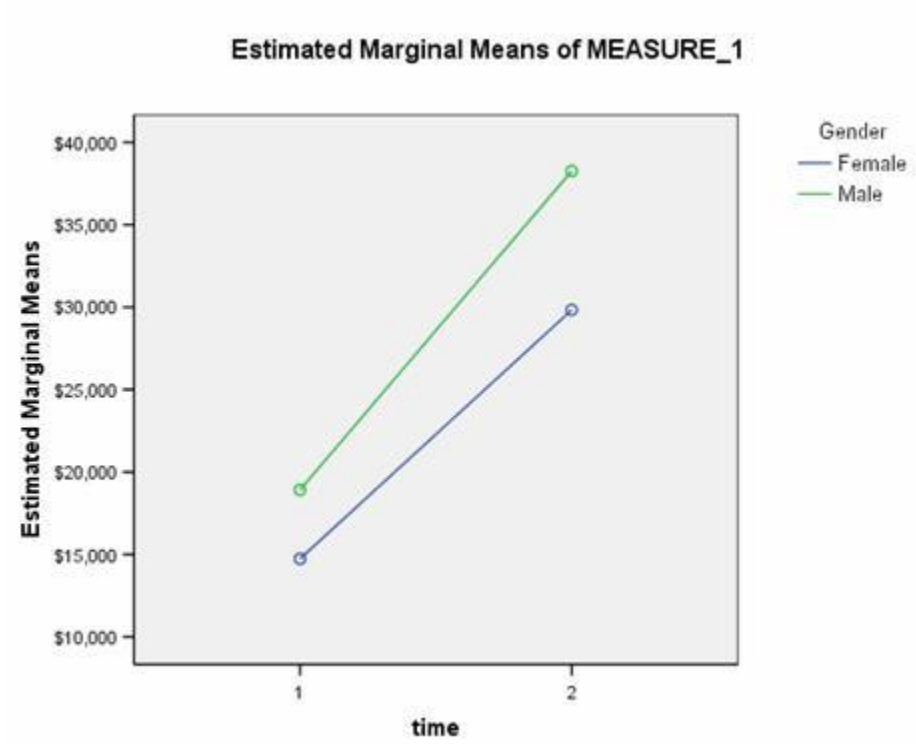
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1992569475	1	1992569475	13.631	.000
educ	4.063E+010	1	4.063E+010	277.955	.000
gender	8155460238	1	8155460238	55.791	.000
Error	6.885E+010	471	146178494.3		

The output for the repeated measures general linear model also provides statistics for between-subject effects. In this example, the model contains two between-subjects factors: employees' education level and their gender. Education level was entered as a covariate in the model, and therefore the statistics associated with it are a measure of the linear relationship between education level and salaries. In contrast, the statistics for the between-subjects factor, gender, represents a comparison between groups across all levels of the within-subjects factors. Specifically, it is a comparison between males and females on differences between their beginning and current salaries. In the above example, both education level and gender are statistically significant. The *F* statistic (277.96) and significance level ($p < .001$) associated with education level allows us to reject the null hypothesis that there is not a linear relationship between education and salaries. By rejecting the null hypothesis, you can conclude that there is a positive linear relationship between the two variables, indicating that as number of years of education increases, salaries do as well. The *F* statistic (55.79) for gender and its associated significance level ($p < .001$) represent a test of the null hypothesis that there are no group differences in salaries. The significant *F* statistic indicates that you can reject this null hypothesis and conclude that there is a statistically significant difference between men and women's salaries.

In this analysis, we saw that there was a significant interaction for Time*Gender. It is important to investigate the properties of your interactions through graphical displays, mean comparisons, or statistical tests, because a significant interaction can take on many forms. For this analysis, we will add a plot for the Time*Gender interaction by returning to the *Repeated Measures* dialog box and clicking on the *Plots* button.



Typically, within-subjects factors such as *time* are placed in the *Horizontal Axis* box, while between-subjects factors are placed in the *Separate Lines* box. After placing *time* and *gender* in the appropriate box, click on *Add* to place the interaction plot into the list in the bottom pane, then choose *Continue* and *OK*. This yields the plot below.



Here, the interaction reflects the fact that male salaries tend to increase more sharply over time than do female salaries. Note that the plot uses the estimated marginal means for salary, which

are the estimated means when the covariate (in this case, education) is set at its average (13.49 years).

Section 3: Some Further Resources

For more information on SPSS, try the following resources:

- To make a free statistical appointment with the Statistical Support group, visit http://ssc.utexas.edu/consulting/free-consulting.
- Visit our list of answers to frequently asked SPSS questions at: http://ssc.utexas.edu/software/faqs/spss
- Go to the SPSS homepage, www.spss.com. Their site includes a variety of helpful resources, including the SPSS Answer Net for answers to frequently asked technical questions, and an FTP site where you can find SPSS macros and an online version of the SPSS Algorithms manual. The support website is available at support.spss.com.