

Simple Linear Regression, Scatterplots, and Bivariate Correlation

This section covers procedures for testing the association between two continuous variables using the SPSS **Regression** and **Correlate** analyses. Specifically, we demonstrate procedures for running Simple Linear Regression, Producing Scatterplots, and running Bivariate Correlations, which report the Pearson's r statistic. These analyses will allow us to identify the strength (Pearson's r) and direction (the sign of r and b) of the relationship between our variables of interest and make predictions regarding the value of the outcome variable (Y) for known values of a predictor (X).

For the following examples, we

have recreated the data set based on cartoon 8.1 (Santa Deals with the reindeer) found in Table 8.1. Again, the Independent variable is the number of times in a month that the reindeer complain to Santa. The dependent variable is the size of the herd for a given month.

THE FAR SIDE

By GARY LARSON



And I've only one thing to say about all these complaints I've been hearing about . . . venison!

<i>Month</i>	<i>Number of Complaints Received Per Month (X)</i>	<i>Current Size of Herd (Y)</i>
January	2	25
February	1	24
March	3	24
April	4	23
May	3	22
June	5	20
July	4	20
August	7	18
September	8	17
October	10	14
November	11	12
December	14	9

Figure 8.1 Variable View of Reindeer Data

	Name	Type	Width	Decimals	Label	Values
1	complain	N	8	2	Number of Complaints Received	None
2	herdsize	N	8	2	Current Size of the Herd	None

Figure 8.2 Data View of Reindeer Data

	complain	herdsize	var	var	var
1	2.00	25.00			
2	1.00	24.00			
3	3.00	24.00			
4	4.00	23.00			
5	3.00	22.00			
6	5.00	20.00			
7	4.00	20.00			
8	7.00	18.00			
9	8.00	17.00			
10	10.00	14.00			
11	11.00	12.00			
12	14.00	9.00			

Setting Up the Data

Figure 8.1 presents the variable view of the SPSS data editor where we have defined two variables (both continuous). The first variable represents the frequency with which the reindeer complain during the months we sampled. We have given it the variable name **complain** and given it the variable label “Number of Complaints Received per Month.” The second variable represents the herd size during each month that we sampled. We have given it the variable name **herdsize** and the variable label “Current Size of the Herd.”

Figure 8.2 presents the data view of the SPSS data editor. Here, we have entered the complaint and herd size data for

the 12 months we have sampled. Remember that the columns represent each of the different variables and the rows represent each observation, which in this case is each cow. For example, during the first month, 2 reindeer complained and the herd size was 25. Similarly, during the 12th month, 14 reindeer complained and herd size consisted of 9 reindeer.

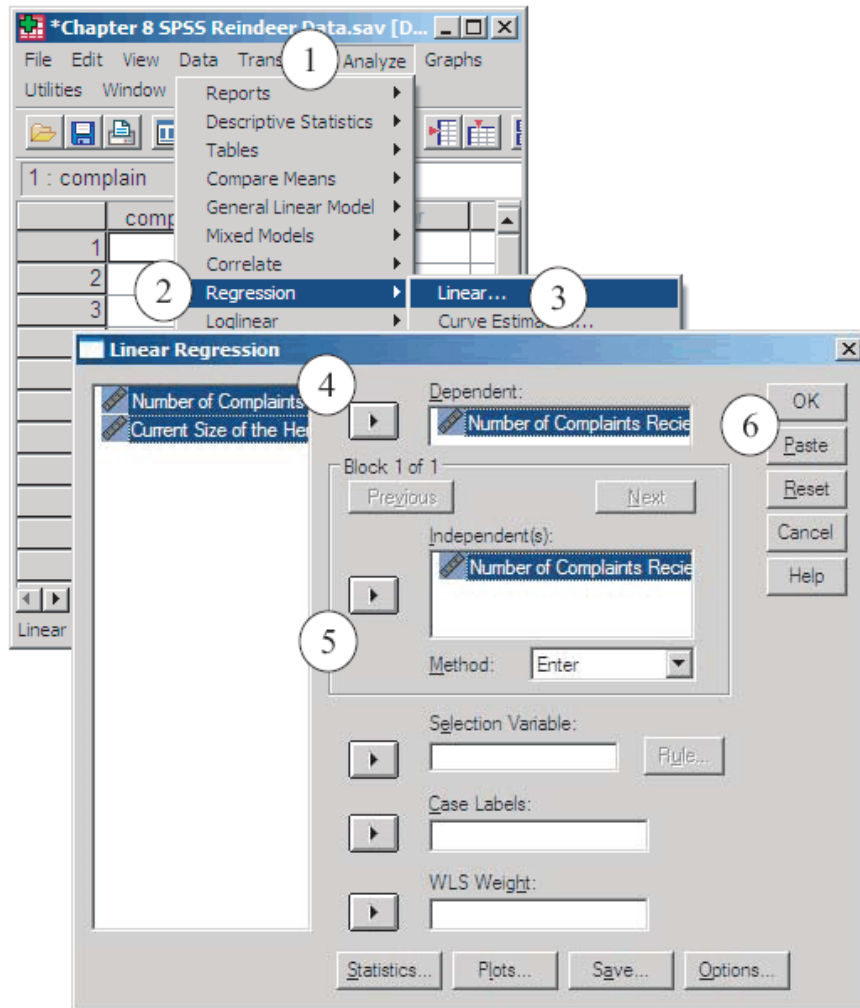
Simple Linear Regression

Simple Linear Regression allows us to determine the direction and of the association between two variables and to identify the least squares regression line that best fits the data. In conjunction with the regression equation ($Y = a + bX$), this information can be used to make predictions about the value of Y for known values of X. For example, we can make predictions about the number of reindeer that will be turned into venison if they complain a certain number of times. Further, the SPSS simple regression analysis will tell us whether a significant amount of the variance in one variable is accounted for (predicted) by another variable. That is, these analyses will tell us whether the relationship between reindeer complaints and the size of the herd is a significant relationship (not likely to have occurred by chance alone).

Running the Analyses

Simple Linear Regression (See Figure 8.3): From the **Analyze** (1) pull down menu, select **Regression** (2), then select **Linear...** (3) from the side menu. In the **Linear Regression** dialogue box, enter the variable **herdsize** in the **Dependent:** field by left-clicking on the variable

Figure 8.3 SPSS: Running Simple Linear Regression



and left-clicking on the boxed arrow (4) pointing to the **Dependent:** field. Next, enter the variable **complain** in the **Independent(s):** field by left-clicking on the variable and left-clicking on the boxed arrow (5) pointing to the **Independent(s):** field.

Finally, double check your variables and either select **OK** (6) to run, or **Paste** to create syntax to run at a later time.

If you selected the paste option from the procedure above, you should have generated the following syntax:

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT herdsiz
/METHOD=ENTER complain .
```

To run the analyses using the syntax, while in the Syntax Editor, select **All** from the **Run** pull-down menu.

Figure 8.4 Output for Regression Analyses

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	26.551	.537		49.406	.000
	Number of Complaints Recieved per Month	-1.258	.075	-.983	-16.696	.000

a. Dependent Variable: Current Size of the Herd

Reading the Simple Linear Regression Output

The Linear Regression Output is presented in Figure 8.4. This output consists of four parts: Variables Entered/Removed, Model Summary, Anova, and Coefficients. For our purposes we really only need to concern ourselves with the Coefficients output. Interpretation of the other parts of the output is more fully described in Chapter 12 on Multiple Regression. The first row of this output, labeled (**constant**), reports the Y-intercept (a) in the first column, labeled B. The second column of this row, provides us with a Standard Error (SE) of the Y-intercept. Like the standard error of the mean found in Chapter 8 (confidence intervals and 9 (t -tests), this SE is an estimate of how much the Y-intercept for sample potentially differs from the Y-intercept found in the population. The last two columns of this row report a t value and the associated significance level, which tests whether the Y intercept is significantly different from zero. The t value is obtained by dividing the Y-intercept by the Standard Error of the Y-intercept. In this case, the Y-intercept for our data is 26.551, the SE of the Y-intercept is .537, and is significantly different from zero ($t = 49.406$) at a significance level of less than .001.

The second row of this output, labeled with variable label. presents the slope of the regression line (b) in first column, labeled B. In our example, b is -1.258. That is for every

complaint that the reindeer lodge, the herd size decreases by 1.258 reindeer. The second column of this row presents the Standard Error of b (often referred to as the SE of the Regression Coefficient). Again, this SE is an estimate of how much the true b found in the population potentially differs from b found in our sample. Skipping now to the last two column of the second row, SPSS reports a t statistic that allows us to determine whether relationship between X and Y is significant. Specifically, this t -test tells us whether the slope of the regression equation is significantly different from zero and is obtained by dividing b by the SE of b . In this example b is -1.258, SE of b is .075, and t is -16.696, which is significant at least at the .001 level. Thus, we can conclude that the number of complaints reindeer make is significantly associated with a decrease in the number of reindeer in the herd.

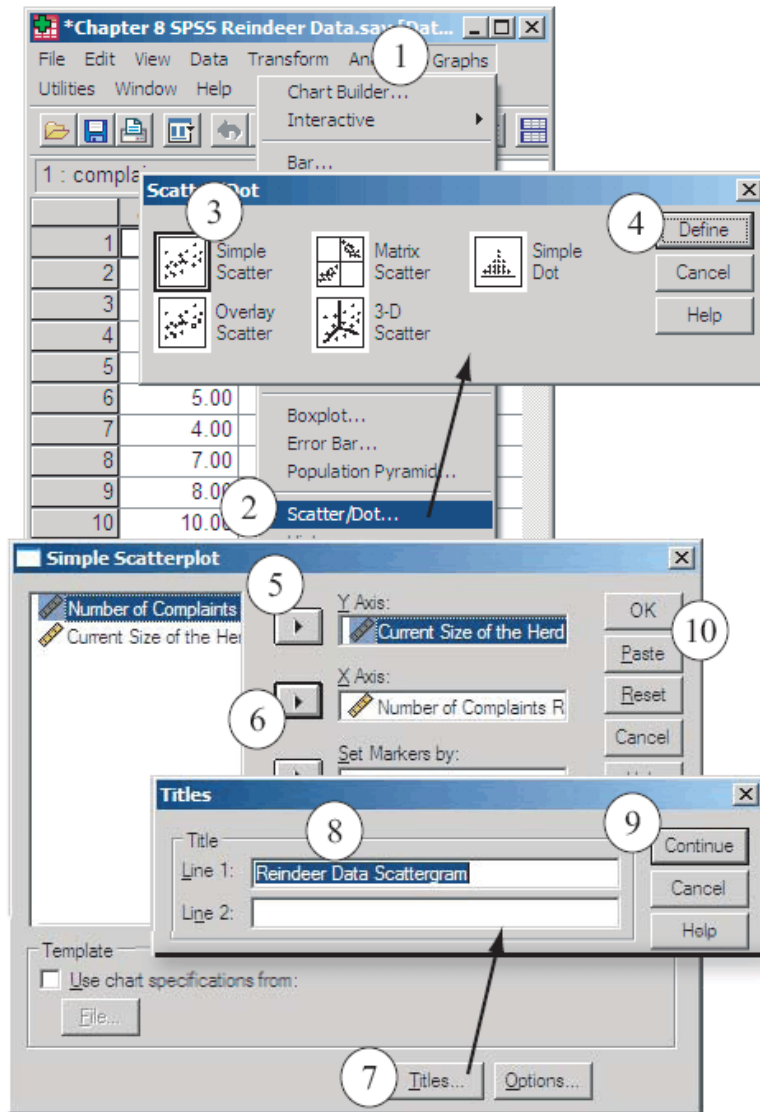
The third column of this output reports a statistic that we have not previously discussed, the Standardized Regression Coefficient Beta. This coefficient is the slope of the regression line for X and Y after both variables have been converted to Z scores. For all practical purposes, in simple regression the Standardized Beta is the same as the Pearson's r coefficient and tells us the strength and direction of the relationship between X and Y. The Standardized Beta for our reindeer data is -.983, indicating that there is strong negative linear relationship between complaint and herd size.

Obtaining the Scatterplot

SPSS gives us the ability to generate a scatterplot (scattergram) for X and Y using the **Graphs** options. Also, we can use graph options to plot the least squares regression line within

the scatterplot. In the following example, we present the steps for obtaining the scatterplot for our reindeer data. Here the X axis (abscissa) will represent the number of complaints and the Y axis (ordinate) will represent the size of the herd.

Figure 8.5 Obtaining Scatterplots



Scatterplot Steps (See

Figure 8.5): From the **Graphs** (1) pull down menu, select **Scatter...** (2). In the **Scatterplot** dialogue box, select the **Simple** option (3) and left-click **Define** (4). In the **Simple Scatterplot** dialogue box, enter the **herdsize** variable in the **Y Axis:** field by left-clicking the **herdsize** variable and left-clicking the boxed arrow (5) pointing to the **Y Axis:** field. Next, enter the **complain** variable in the **X Axis:** field by left-clicking the **complain** variable and left-clicking the boxed arrow (6) pointing to the **X Axis:**

field. To add a descriptive title to the scatterplot, left-click the **Titles...** button (7). In the **Titles** dialogue box enter the major heading desired in the **Line 1:** field (8). In this case we have entered the title “Reindeer data scattergram.” Click **Continue** (9) to return to the **Simple**

Scatterplot dialogue box. Finally, double check your variables and either select **OK** (9) to run, or **Paste** to create syntax to run at a later time.

If you selected the paste option from the procedure above, you should have generated the following syntax:

```
GRAPH
/SCATTERPLOT(BIVAR)=complain WITH herdsizes
/MISSING=LISTWISE
/TITLE= 'Reindeer Data Scattergram'.
```

To run the analyses using the syntax, while in the Syntax Editor, select **All** from the **Run** pull-down menu.

Figure 8.6 Output for Scatterplot

Graph

[DataSet1] C:\Docs\jeff\StatsBook\SPSS GUIDE\Chapter 8 support files\Chapter 8 SPSS Reindeer Data.sav

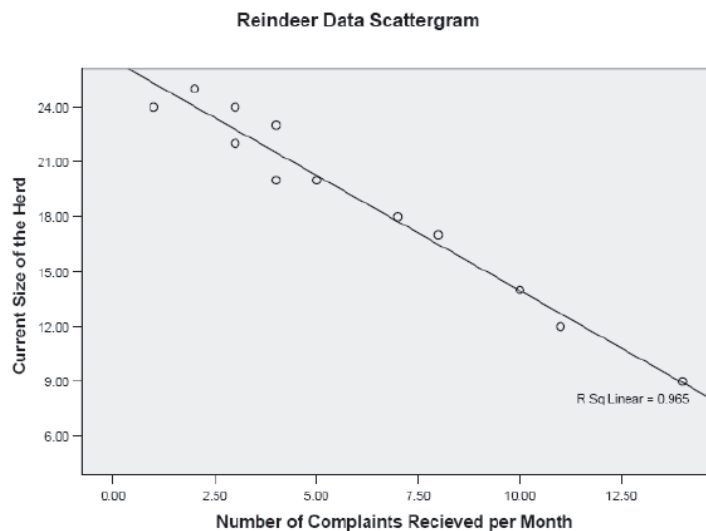
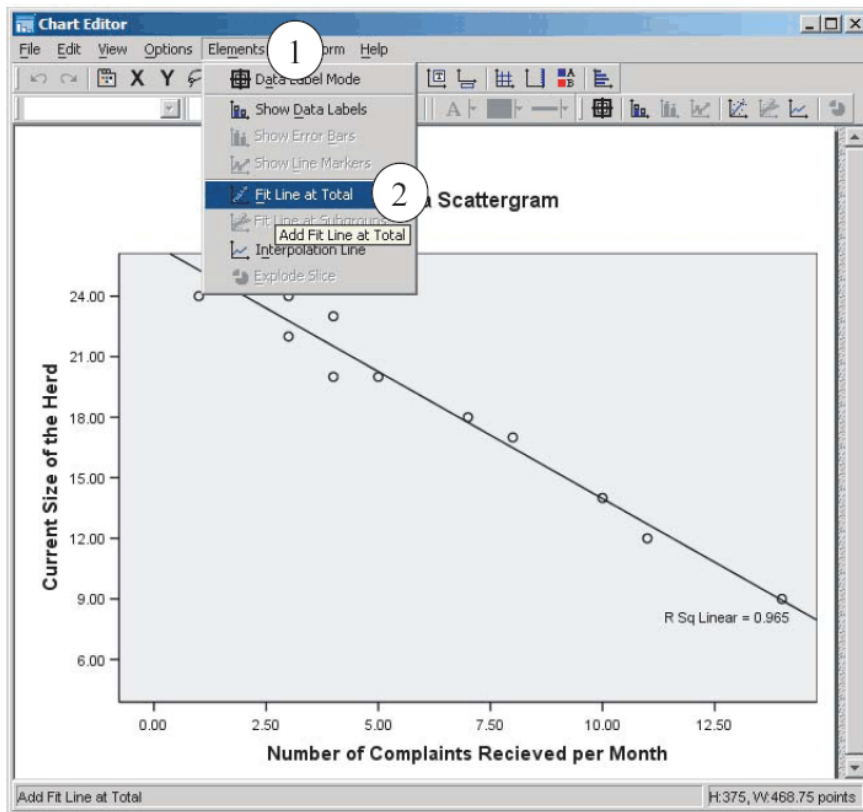


Figure 8.6 presents the scatterplot that we requested and includes the least squares regression line that best fits our data. However, the scatterplot produced using the steps above did not originally include the regression line. We must add the regression line while in the output navigator. The steps are described below.

Adding the Regression Line to a Scatterplot (see Figure 8.7): First, in the Output

Figure 8.7 Adding the Regression Line to the Scatterplot Output



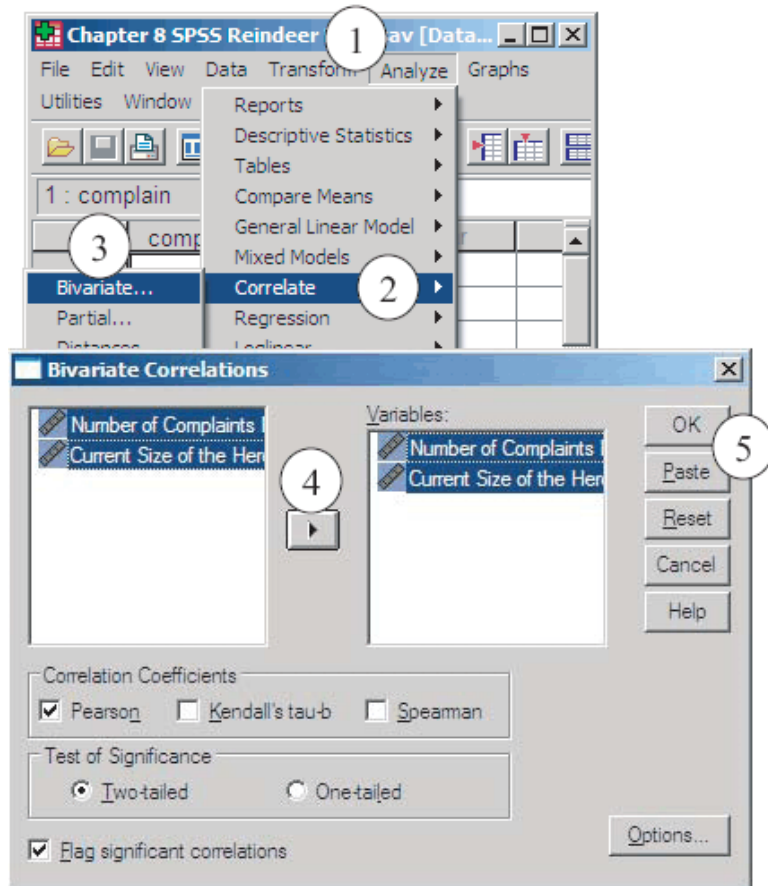
Navigator double left-click on the scatterplot to open the SPSS Chart Editor. In the Chart Editor, from the **Elements** (1) pull down menu, select **Fit Line at Total** (2). To return to the Output Navigator, close the Chart Editor by left clicking the close button in the upper right hand corner of the Chart Editor (5).

Correlation

Pearson's r can be obtained using the **Bivariate Correlation** procedures. Again, this statistic allows us to determine the strength, direction and significance of the association between two variables.

Running the Analyses

Figure 8.8 Running Correlation Analyses



Bivariate Correlation (See Figure 8.8): From the **Analyze** (1) pull down menu, select **Correlate** (2), then select **Bivariate...** (3) from the side menu. In the **Bivariate Correlations** dialogue box, enter the variables **herdsize** and **complain** in the **Variables:** field by either double left-clicking on each variable or by left-clicking on each variable and left-clicking on the boxed arrow (4) pointing to the **Variables:** field. Finally,

double check your variables and either select **OK** (5) to run, or **Paste** to create syntax to run at a later time.

If you selected the paste option from the procedure above, you should have generated the following syntax:

```
CORRELATIONS
/VARIABLES=complain herdsize
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

To run the analyses using the syntax, while in the Syntax Editor, select **All** from the **Run** pull-down menu.

Reading the Output

Figure 8.9 Output for Correlation Analyses

Correlations

[DataSet1] C:\Docs\SPSS GUIDE\Chapter 8 support files\Chapter 8 SPSS Reindeer Data.sav

Correlations

		Number of Complaints Recieved per Month	Current Size of the Herd
Number of Complaints Recieved per Month	Pearson Correlation	A 1	-.983**
	Sig. (2-tailed)		.000
	N	12	12
Current Size of the Herd	Pearson Correlation	B -.983**	A 1
	Sig. (2-tailed)		.000
	N	12	12

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

The Correlation output is presented in Figure 8.9. In our current example, the results are organized in a 2 x 2 matrix, where column 1 and column 2 represent our **complain** and **herdsize** variables, respectively, and row 1 and row 2 represent our **complain** and **herdsize** variables, respectively. Each cell of this matrix presents the Pearson's r correlation between the variables, the significance levels for each correlation, and the number of subjects represented by each correlation from which the degrees of freedom can be obtained (for Pearson's r $df = n - 2$). The cells forming the diagonal of this matrix, row 1 column 1 and row 2 column 2, represent the each variable's correlation with itself. In Figure 10.18 we have labeled the cells forming the diagonal with "A"s. These correlations are rather meaningless and therefore no significance levels are provided. The variables above the diagonal (the cell to the right of the table) and below the diagonal (to the left of the table, labeled B) are redundant. You should notice that the correlations in each of these cells are identical. Thus, even though there are four cells reported in

this matrix, we only have one correlation statistic to interpret. In our example, the correlation between the number of complaints reindeer make to Santa and the size of the reindeer herd $-.983$, and has a significance level of at least $.001$, with 10 degrees of freedom. We can conclude that these two variables are significantly and strongly negatively correlated. That is, the more the reindeer complain, the smaller the size of the herd gets.