IBM SPSS Statistics 23
Part 3: Regression Analysis

Winter 2016, Version 1

Table of Contents

Introduction **2**
Downloading the Data Files 2
Simple Regression **2**
Scatter Plot 2
Predicting Values of Dependent Variables 5
Predicting This Year's Sales with the Simple Regression Model 6
Multiple Regression **8**
Predicting Values of Dependent Variables 9
Predicting This Year's Sales with the Multiple Regression Model 10
Data Transformation **11**
Computing 12
Polynomial Regression **13**
Regression Analysis 14
Analyzing the Results 14
Chart Editing **15**
Adding a Line to the Scatter Plot 15
Manipulating the Scales on the X and Y Axes 17
Adding a Title to the Chart 19
Adding Color to the Chart 19
Applying a Background Color 20

Introduction

SPSS stands for **Statistical Package for the Social Sciences**. This program can be used to analyze data collected from surveys, tests, observations, etc. It can perform a variety of data analyses and presentation functions, including statistical analysis and graphical presentation of data. Among its features are modules for statistical data analysis. These include (1) descriptive statistics such as frequencies, central tendency, plots, charts, and lists; and (2) sophisticated inferential and multivariate statistical procedures such as analysis of variance (ANOVA), factor analysis, cluster analysis, and categorical data analysis. IBM SPSS Statistics 23 is well-suited for survey research, though by no means is it limited to just this topic of exploration.

This handout provides basic instructions on how to answer research questions and test hypotheses using linear regression (a technique which examines the relationship between a dependent variable and a set of independent variables). The value of the dependent variable (e.g., salesperson's total annual sales) can be predicted based on its relationship to the independent variables used in the analysis (e.g., age, education, and years of experience). The two research questions proposed in this handout are as follows:

- How much money will each salesperson make this year?
- Who will qualify for a $1,000 bonus?

Downloading the Data Files

This handout includes sample data files that can be used to follow along the steps. If you plan to use the data files, download the following ZIP file to your computer and extract the files. It is recommended to save the data files on your desktop for easy access.

- IBM SPSS Statistics 23 Part 3 Data Files

Simple Regression

Simple regression estimates how the value of one **dependent variable (Y)** can be predicted based on the value of one **independent variable (X)**. The linear equation for simple regression is as follows:

$$Y = aX + b$$

Simple regression can answer the following research question:

Research Question # 1

Based on last year's sales, how much money will each salesperson make this year?

Scatter Plot

A **scatter plot** displays the nature of the relationship between two variables. Before performing a regression analysis, it is recommended to run a scatter plot to determine if there is a **linear relationship** between the variables. If there is no linear relationship (i.e., points on a graph are not clustered in a straight line), then a simple regression would not be the appropriate analysis to use for this data set.

To run a scatter plot:

1. Start **IBM SPSS Statistics 23**, and then open the **Regression.sav** file.
2. Click the **Graphs** menu, point to **Legacy Dialogs**, and then click **Scatter/Dot** (see Figure 1).

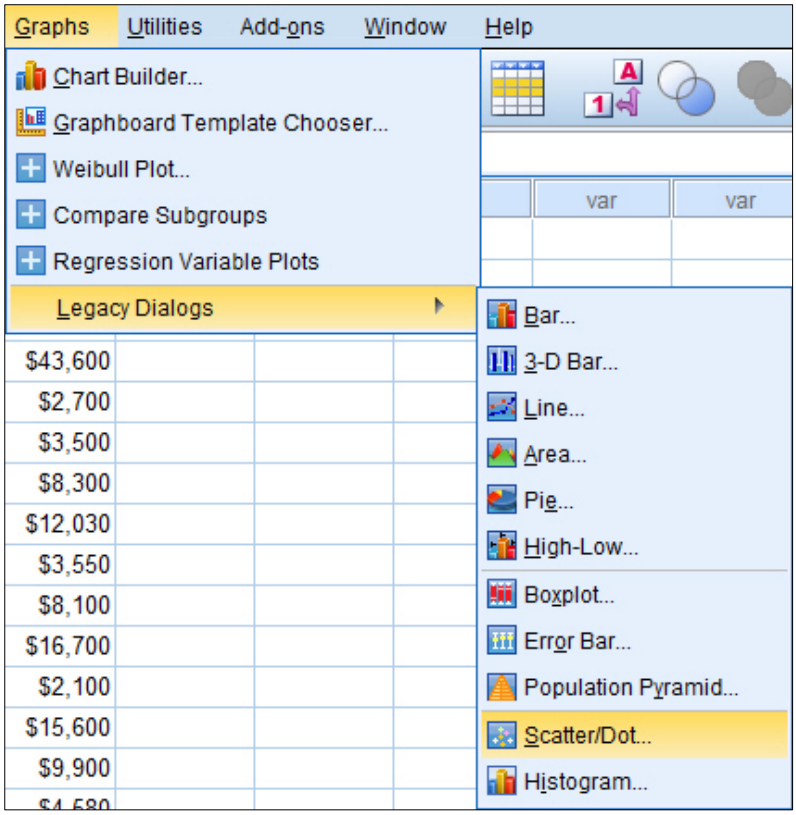

Figure 1 – Scatter/Dot Selected on the Graphs Menu

3. In the **Scatter/Dot** dialog box, make sure that the **Simple Scatter** option is selected, and then click the **Define** button (see Figure 2).

   NOTE: The **Simple Scatter** plot is used to estimate the relationship between two variables.

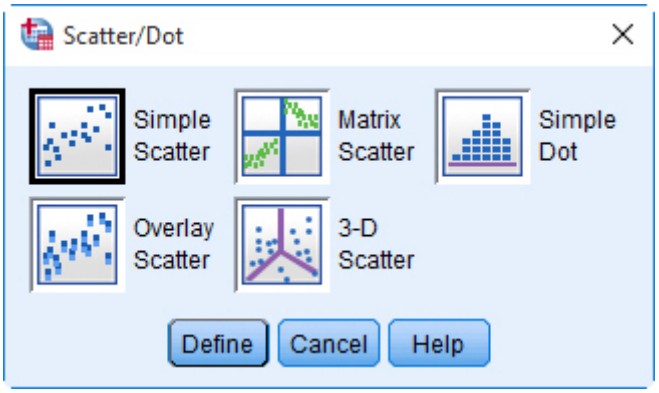

Figure 2 – Scatter/Dot Dialog Box

4. In the **Simple Scatterplot** dialog box, select the **Last year sales** variable in the left box, and then click the transfer arrow button to move it to the **Y Axis** box (see Figure 3).
5. Select the **Years of experience** variable in the left box, and then click the transfer arrow button to move it to the **X Axis** box.

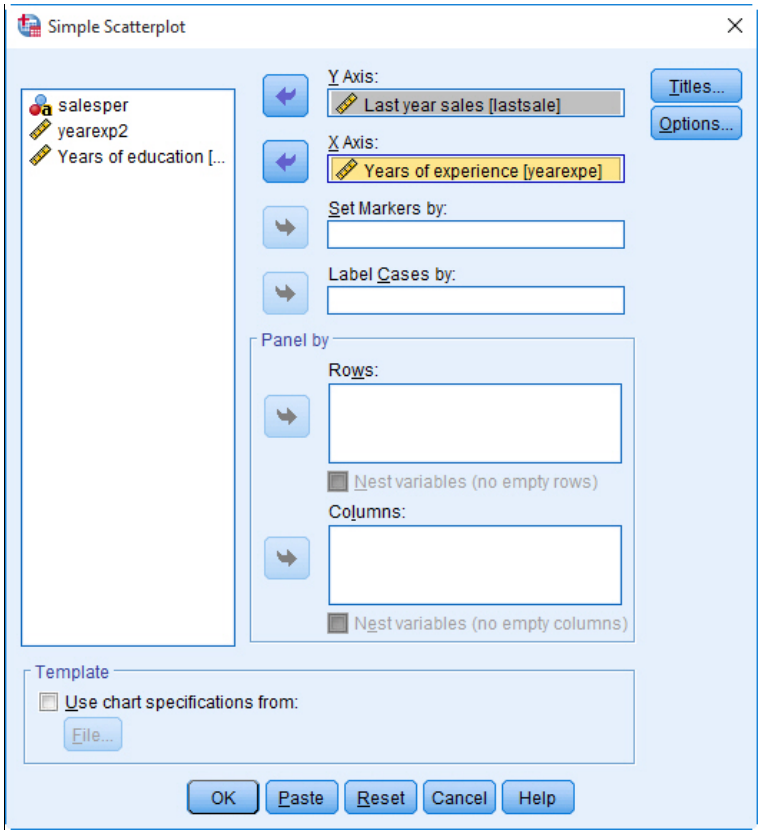

Figure 3 – Simple Scatterplot Dialog Box

6. Click the **OK** button. The **Output Viewer** window opens and displays a scatter plot of the variables (see Figure 4).

   NOTE: The scatter plot in Figure 4 indicates that a linear relationship exists between the variables **Last year sales** and **Years of experience**. The next step is to find a line that best accommodates the pattern of points in this scatter plot. The steps for enhancing the graph appearance are covered in the last section of this handout.

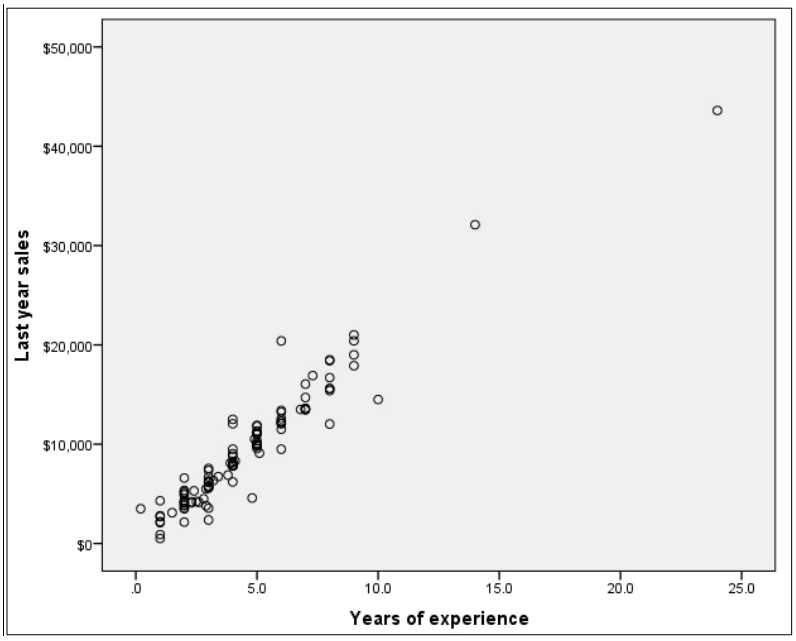

Figure 4 – Scatter Plot

Predicting Values of Dependent Variables

Judging from the scatter plot above, a linear relationship seems to exist between the two variables. Therefore, a simple regression analysis can be used to calculate an equation that will help predict this year's sales.

To run a simple regression analysis:

1. Switch to the **Data Editor** window.
2. Click the **Analyze** menu, point to **Regression**, and then click **Linear**.
3. In the **Linear Regression** dialog box, select the **Last year sales** variable in the left box, and then click the transfer arrow button to move it to the **Dependent** box (see Figure 5).
4. Select the **Years of experience** variable in the left box, and then click the transfer arrow button to move it to the **Independent(s)** box.
5. Click the **OK** button.

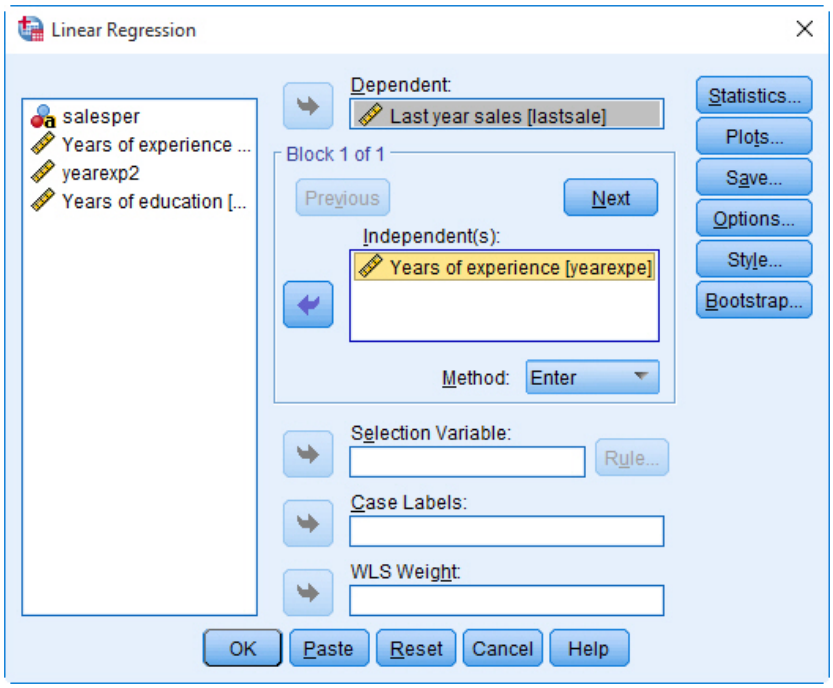

Figure 5 – Linear Regression Dialog Box

The following tables in the Output Viewer window present the results of a simple regression. **R Square (.918)** indicates that this model accounts for almost 92% of the total variation in the data (see Figure 6).

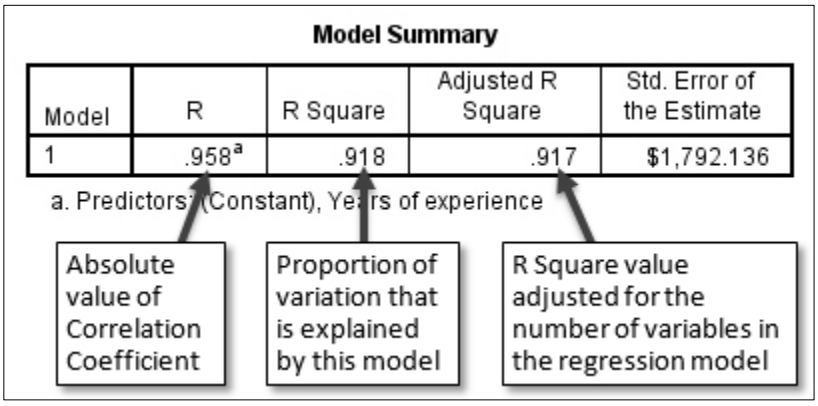

Figure 6 – Model Summary Output

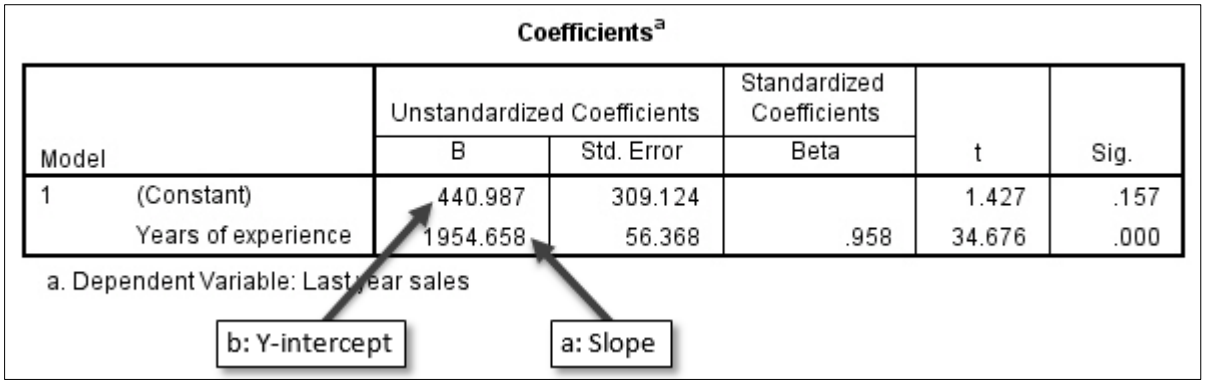

Figure 7 – Coefficients Output

The slope and the y-intercept as seen in Figure 7 should be substituted in the following linear equation to predict this year's sales: **Y = aX + b**. In this case, the values of **a**, **b**, **X**, and **Y** will be as follows:

a = 1954.658
b = 440.987
X = Years of experience (values of independent variable)
Y = Last year sales (values of dependent variable)

Predicting This Year's Sales with the Simple Regression Model

To predict this year's sales for each salesperson, substitute the values of **a** and **b** in the following linear equation:

Y = aX + b

Last year sales = (a * yearexpe) + b
This year sales = (1954.658 * yearexp2) + 440.987
a = 1954.658
b = 440.987
X = Years of experience [yearexp2]
Y = This year sales

NOTE: The new independent variable **yearexp2** is used instead of **yearexpe** in order to predict this year's sales.

To predict this year's sales using the computing function:

1. Switch to the **Data Editor** window.
2. Click the **Transform** menu, and then click **Compute Variable**.
3. In the **Compute Variable** dialog box, type **Simple** in the **Target Variable** box (see Figure 8).
4. In the **Numeric Expression** box, enter the following equation by typing or selecting from the dialog box keypad:

1954.658 * yearexp2 + 440.987

NOTE: It is recommended to select the **yearexp2** variable directly from the **Variable** box on the left side of the **Compute Variable** dialog box to prevent typing mistakes.

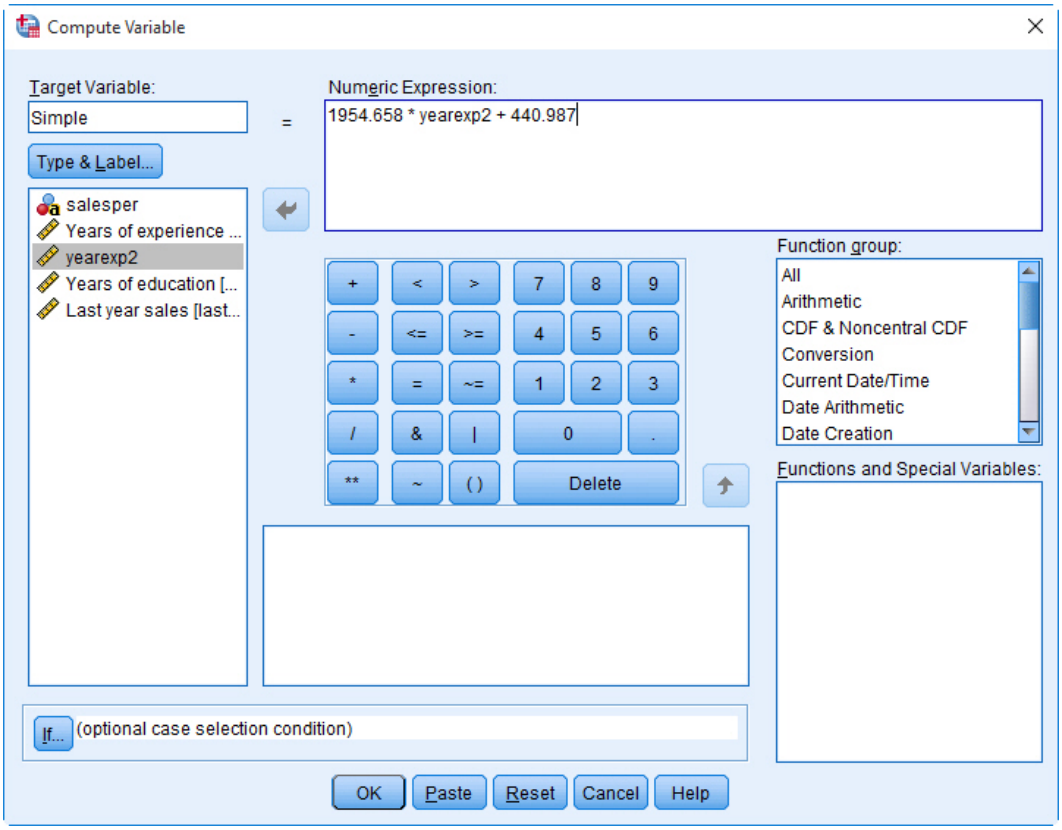

Figure 8 – Compute Variable Dialog Box

5. Click the **OK** button. The results are displayed in the **Simple** column in **Data View** (see Figure 9).

	salesper	yearexpe	yearexp2	educatio	lastsale	Simple
1	Achillis	2.0	3.0	18	$6,590	6304.96
2	Aeoa	2.5	3.5	14	$4,230	7282.29
3	Agamemno	8.0	9.0	14	$15,400	18032.91
4	Ana	24.0	25.0	16	$43,600	49307.44
5	Anticlea	1.0	2.0	16	$2,700	4350.30
6	Antipus	.2	1.2	20	$3,500	2786.58
7	Antony	4.1	5.1	16	$8,300	10409.74
8	Aphrodit	8.0	9.0	9	$12,030	18032.91

Figure 9 – Simple Regression Results

To change the data type for the Simple variable:

1. Click the **Variable View** tab in the lower-left corner of the **Data Editor** window (see Figure 10).

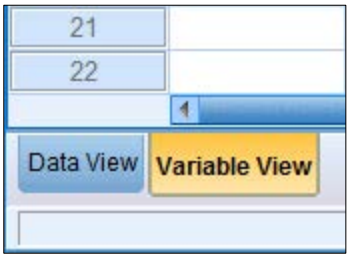

Figure 10 – Variable View Tab

2. Locate the **Simple** variable in row 6, click in the next cell under the **Type** column, and then click the **Ellipses** button that appears.
3. In the **Variable Type** dialog box, select the **Dollar** option button, select the **$###,###,###** format (12 digit width with 0 decimal places), and then click the **OK** button (see Figure 11).

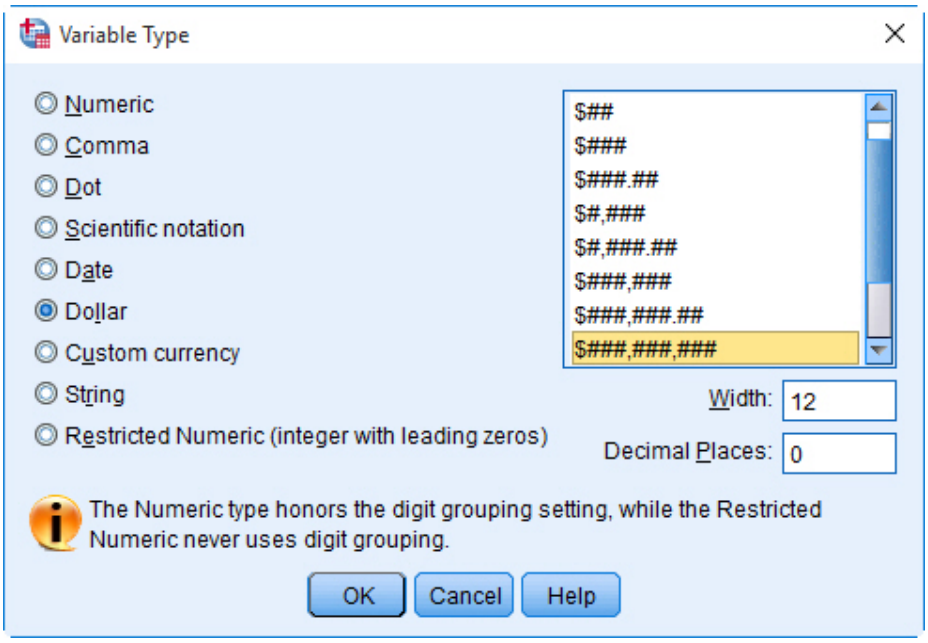

Figure 11 – Variable Type Dialog Box

4. Click the **Data View** tab in the lower-left corner of the **Data Editor** window.

   NOTE: The prediction of this year's sales for each salesperson is computed under the new variable named **Simple** (see Figure 12).

	salesper	yearexpe	yearexp2	educatio	lastsale	Simple
1	Achillis	2.0	3.0	18	$6,590	$6,305
2	Aeoa	2.5	3.5	14	$4,230	$7,282
3	Agamemno	8.0	9.0	14	$15,400	$18,033
4	Ana	24.0	25.0	16	$43,600	$49,307
5	Anticlea	1.0	2.0	16	$2,700	$4,350
6	Antipus	.2	1.2	20	$3,500	$2,787
7	Antony	4.1	5.1	16	$8,300	$10,410

Figure 12 – Simple Regression Prediction

Multiple Regression

Multiple regression estimates the coefficients of the linear equation when there is more than one independent variable that best predicts the value of the dependent variable. For example, a salesperson's total annual sales (the dependent variable) can be predicted based on independent variables such as age, education, and years of experience. The linear equation for multiple regression is as follows:

$$Z = aX + bY + c$$

Predicting Values of Dependent Variables

The previous section demonstrated how to predict this year's sales (the dependent variable) based on one independent variable (number of years of experience) by using simple regression analysis. Similarly, this year's sales (the dependent variable) can be predicted from more than one independent variable (such as **Years of experience** and **Years of education**) by using multiple regression analysis.

To run a multiple regression analysis:

1. Click the **Analyze** menu, point to **Regression**, and then click **Linear**. The **Linear Regression** dialog box opens (see Figure 13).

   NOTE: If there are variables in the **Dependent** or **Independent(s)** boxes, click the **Reset** button before performing steps 2 and 3 below.

2. Select the **Last year sales** variable in the left box, and then click the transfer arrow button to move it to the **Dependent** box.
3. Select the **Years of experience** and **Years of education** variables in the left box, and then click the transfer arrow button to move them to the **Independent(s)** box.

   NOTE: You can select multiple variables by clicking the first variable, holding down the **Ctrl** key, and then clicking each of the other variables.

4. Click the **OK** button. The **Output Viewer** window opens.

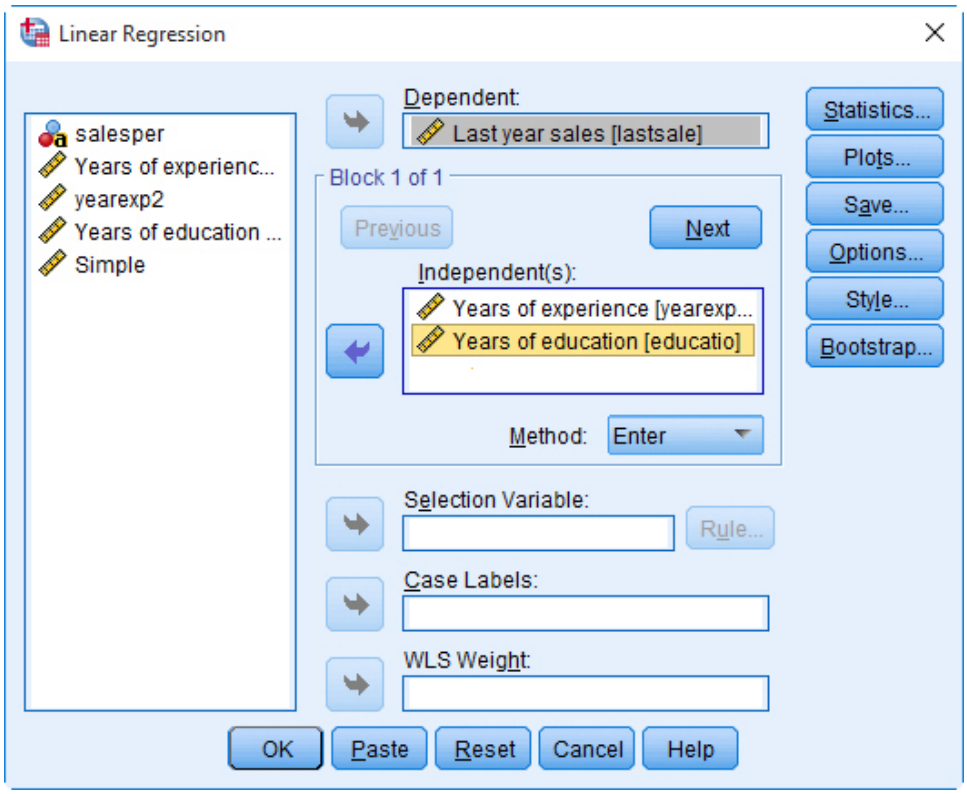

Figure 13 – Linear Regression Dialog Box

NOTE: The table should look similar to Figure 14. **R Square** = **.976** indicates that this model can predict this year's sales almost 98% correctly.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.988[a]	.976	.975	$980.236

a. Predictors: (Constant), Years of education, Years of experience

Figure 14 – Model Summary Output for Multiple Regression

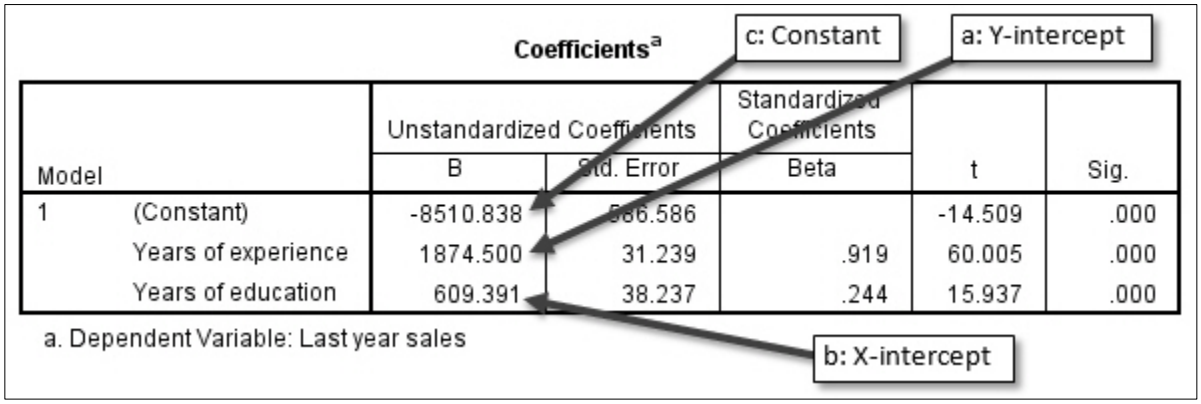

Figure 15 – Multiple Regression Output

To predict this year's sales, substitute the values for the slopes and y-intercept displayed in the Output Viewer window (see Figure 15) in the following linear equation: **Z = aX+ bY + c**.

In this case, the values of **a**, **b**, **X**, and **Y** will be as follows:

a = 1874.5
b = 609.391
c = (-8510.838)
X = Years of experience (independent variable)
Y = Years of education (independent variable)
Z = This year sales (dependent variable)

As indicated in the output table, the coefficient for **Years of experience** is **1874.5** and the coefficient for **Years of education** is **609.391**.

Predicting This Year's Sales with the Multiple Regression Model

To predict this year's sales for each salesperson, substitute the values of **a**, **b**, and **c** in the following linear equation: **Z = aX + bY + c**.

This year sales = 1874.5 * Years of experience + 609.391 * Years of education + (-8510.838)

To predict this year's sales by multiple regression analysis:

1. Switch to the **Data Editor** window.
2. Click the **Transform** menu, and then click **Compute Variable**.
3. In the **Compute Variable** dialog box, click the **Reset** button (see Figure 16).
4. In the **Target Variable** box, type **Multiple**.
5. In the **Numeric Expression** box, enter the following equation by typing or selecting from the dialog box keypad:

1874.5 * yearexp2 + 609.391 * educatio - 8510.838

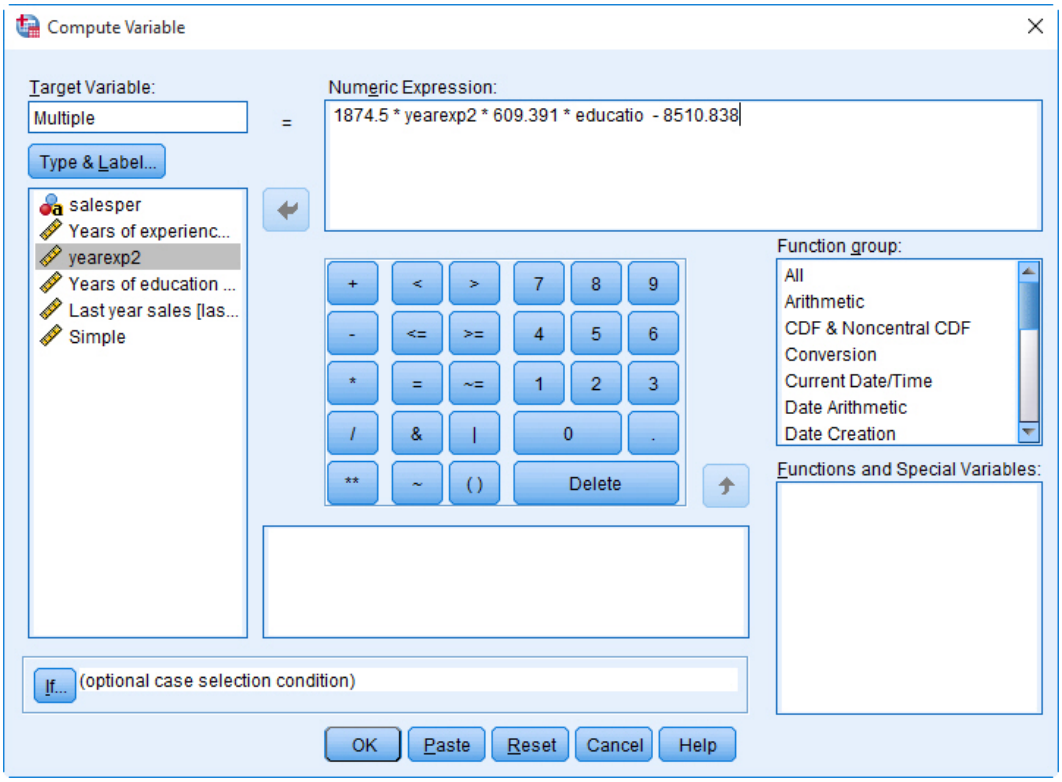

Figure 16 – Compute Variable Dialog Box

6. Click the **OK** button. The **Multiple** column in **Data View** displays the results (see Figure 17).

   NOTE: The sales prediction for each salesperson using two independent variables is listed under the new variable named **Multiple**.

	salesper	yearexpe	yearexp2	educatio	lastsale	Simple	Multiple
1	Achillis	2.0	3.0	18	$6,590	$6,305	8081.70
2	Aeoa	2.5	3.5	14	$4,230	$7,282	6581.39
3	Agamemno	8.0	9.0	14	$15,400	$18,033	16891.14
4	Ana	24.0	25.0	16	$43,600	$49,307	48101.92
5	Anticlea	1.0	2.0	16	$2,700	$4,350	4988.42
6	Antipus	.2	1.2	20	$3,500	$2,787	5926.38
7	Antony	4.1	5.1	16	$8,300	$10,410	10799.37
8	Aphrodit	8.0	9.0	9	$12,030	$18,033	13844.18

Figure 17 – Multiple Regression Results

Data Transformation

Situations may arise when data transformation is useful. Most data transformations can be performed with the **Compute** command. Using this command, the data file can be manipulated to fit various statistical performances.

Research Question # 2

Who will earn a $1,000 bonus?

Computing

Each salesperson's yearly sales were predicted using multiple regression analysis. The salespeople who made $2,000 or more than their predicted values will receive a $1,000 bonus. Use the **Compute** command to compare the values of this year's actual sales with the predictions from multiple regression analysis computed in the previous lesson to find eligible salespeople. The first step in predicting who will receive a bonus is to calculate the difference between this year's actual sales and the prediction of this year's sales from the multiple regression analysis.

To predict who qualifies for the bonus:

1. Open the **Bonus.sav** file.
2. Click the **Transform** menu, and then click **Compute Variable**.
3. In the **Compute Variable** dialog box, type **bonus** in the **Target Variable** box, type **1000** in the **Numeric Expression** box, and then click the **If** button (see Figure 18).

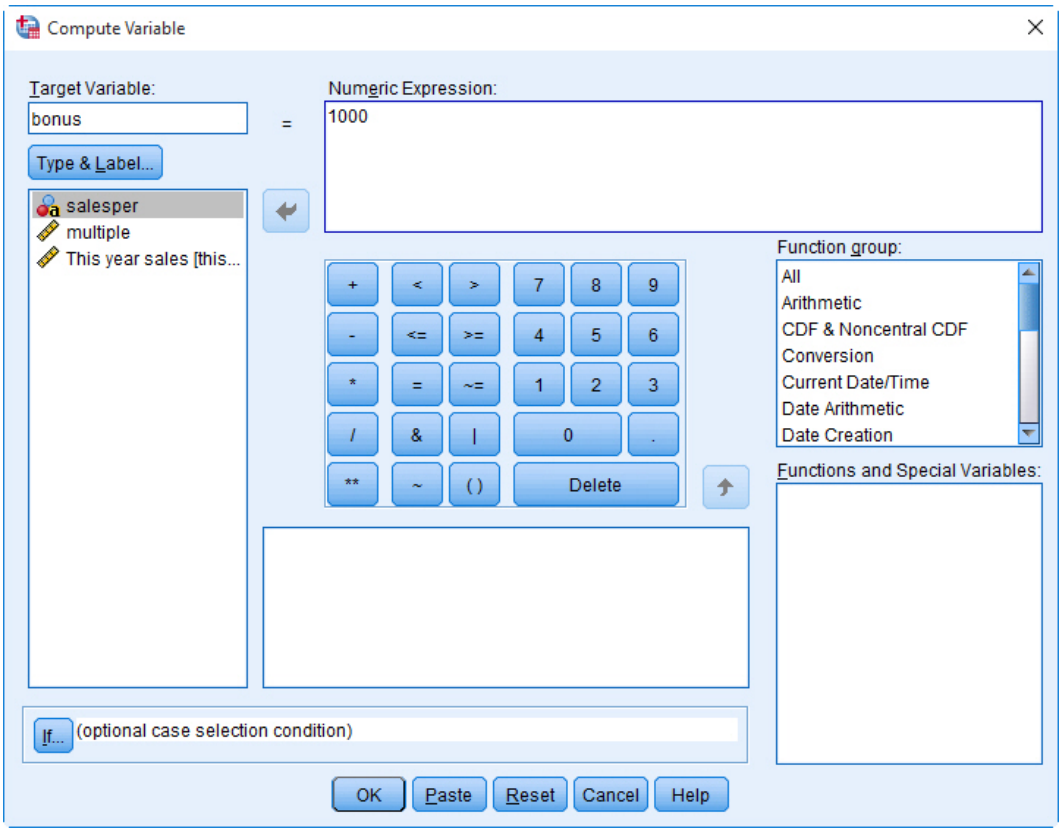

Figure 18 – Compute Variable Dialog Box

4. In the **Compute Variable: If Cases** dialog box, select the **Include if case satisfies condition** option button (see Figure 19).
5. Enter the following expression by typing or selecting from the dialog box keypad:

   thissale - multiple >= 2000

   NOTE: It is recommended to select the variables and the **>=** sign directly from the **Variable** box and keypad provided in the dialog box to prevent mistakes.

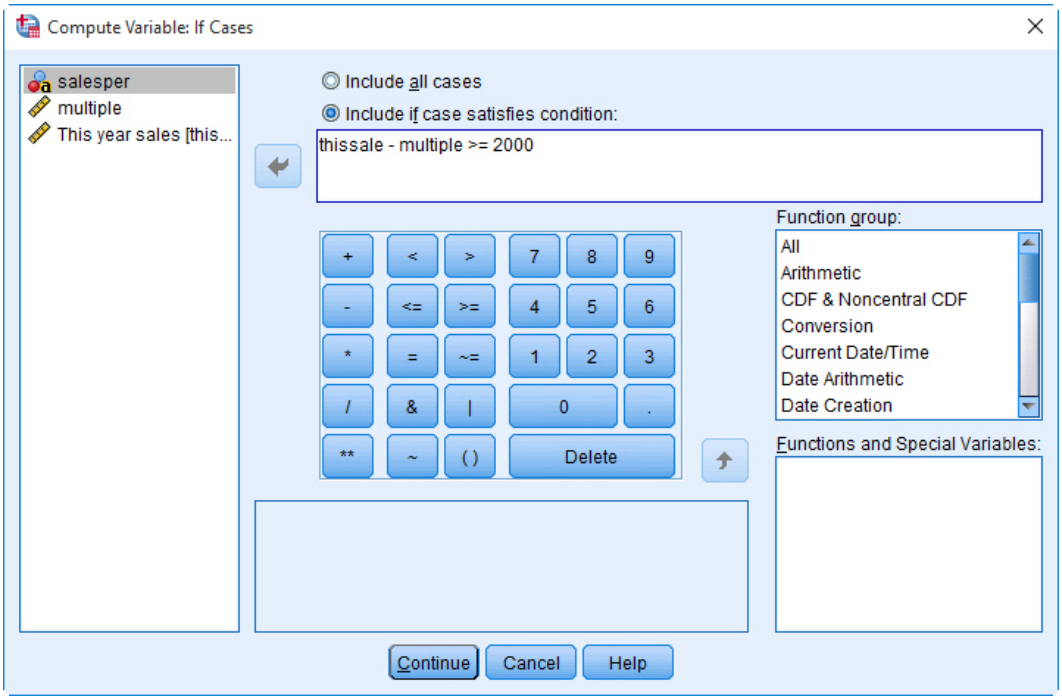

Figure 19 – Compute Variable: If Cases Dialog Box

6. Click the **Continue** button, and then click the **OK** button.

   NOTE: Salespersons **Ivett** (#44) and **Jason** (#49) are two of the sales personnel who qualify for the $1,000 bonus because they made $2,000 over their predicted sales from the last lesson (see Figure 20).

44	Ivett	$1,332	$8,000	1000.00
45	Jacky	$19,329	$8,000	.
46	Jacy	$14,361	$8,000	.
47	Jane	$16,235	$8,000	.
48	Janet	$9,393	$8,000	.
49	Jason	$4,988	$8,000	1000.00

Figure 20 – Bonus Results

Polynomial Regression

This type of regression involves fitting a dependent variable (Y_i) to a polynomial function of a single independent variable (X_i). The regression model is as follows (see Table 1 for the meaning of the variables):

$$Y_i = a + b_1X_i + b_2X_i^2 + b_3X_i^3 + \ldots + b_kX_i^k + e_i$$

Table 1 – Breakdown of the Variables

Variable	Meaning
a	Constant
$\mathbf{b_k}$	The coefficient for the independent variable to the k'th power
$\mathbf{e_i}$	Random error term

Regression Analysis

To look at the growth relationship between weight and age:

1. Open the **Growth.sav** file.
2. Click the **Analyze** menu, point to **Regression**, and then click **Curve Estimation**.
3. In the **Curve Estimation** dialog box, transfer the **wght** variable to the **Dependent(s)** box and the **age** variable to the **Independent Variable** box (see Figure 21).

   NOTE: The dependent variable **wght** is predicted using the independent variable **age**.

4. Deselect the **Plot models** check box.
5. Select the **Display ANOVA table** check box.
6. Under **Models**, deselect the **Linear** check box and select the **Cubic** check box.
7. Click the **OK** button.

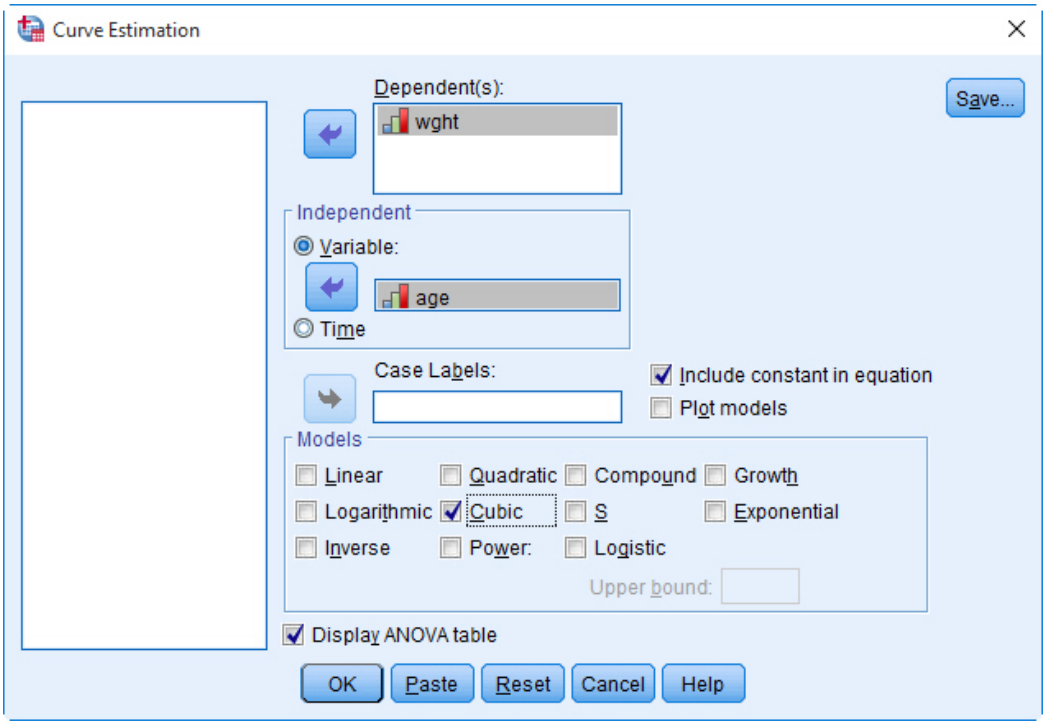

Figure 21 – Curve Estimation Dialog Box

Analyzing the Results

This cubic model has an R Square value of 99.567% (see Figure 22). The F-ratio indicates a highly significant fit. The best fitting cubic polynomial is given by the following equation:

$$Y_i = 0.052 - 0.017\ X_i + 0.010\ X_i^2 - 0.001\ X_i^3 + e_i$$

(where Y_i is weight and X_i is age)

Polynomial regression can find the line of best fit for polynomials consisting of two or more variables. If X is the dependent variable, use the **Transform** and **Compute** options of the **Data Editor** (as discussed earlier in this lesson) to create new variables X2 = X*X, X3 = X*X2, X4 = X*X3, etc., then use these new variables (X, X2, X3, X4, etc.) as a set of independent variables for the regression analysis.

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.998	.996	.995	.005

The independent variable is age.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	.099	3	.033	1150.294	.000
Residual	.000	15	.000		
Total	.100	18			

The independent variable is age.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
age	-.017	.005	-.660	-3.403	.004
age ** 2	.010	.001	4.055	8.861	.000
age ** 3	-.001	.000	-2.487	-8.857	.000
(Constant)	.052	.006		8.500	.000

Figure 22 – Polynomial Regression Summary Results

Chart Editing

During the final stage of research, enhancing the appearance of charts and figures can help viewers understand what may seem to be confusing statistics. The following steps explain some useful methods for enhancing a chart's appearance.

Adding a Line to the Scatter Plot

Adding a straight line to fit the scattered pattern of a data chart can help emphasize the linear relationship between the data.

To add a line to the scatter plot:

1. Click the **Graphs** menu, point to **Legacy Dialogs**, and then click **Scatter/Dot**.
2. In the **Scatter/Dot** dialog box, select the **Simple Scatter** option, and then click the **Define** button.
3. In the **Simple Scatterplot** dialog box, transfer the **age** variable to the **X Axis** box and the **wght** variable to the **Y Axis** box, and then click the **OK** button. A chart appears in the **Output Viewer** window.
4. Double-click the chart to modify it. The **Chart Editor** window and the **Properties** dialog box open.
5. In the **Chart Editor** window, right-click a chart marker, and then click **Add Fit Line at Total** on the shortcut menu (see Figure 23).
6. In the **Properties** dialog box, on the **Fit Line** tab, select the **Cubic** option button under **Fit Method**, deselect the **Attach label to line** check box, and then click the **Apply** button (see Figure 24).
7. Close the **Chart Editor** window.

NOTE: Notice that the **Add Fit Line at Total** does not automatically capture the way the data curves, but the **Cubic** method is almost a perfect fit (see Figure 25). Make sure to select the best fit line for the data. Once a best fit line is applied, it will stay on the graph. Selecting **Add Fit Line at Total** again will add a new fit line to the graph. To edit the fit line that has already been applied, select the fit line that is already on the graph and change the properties.

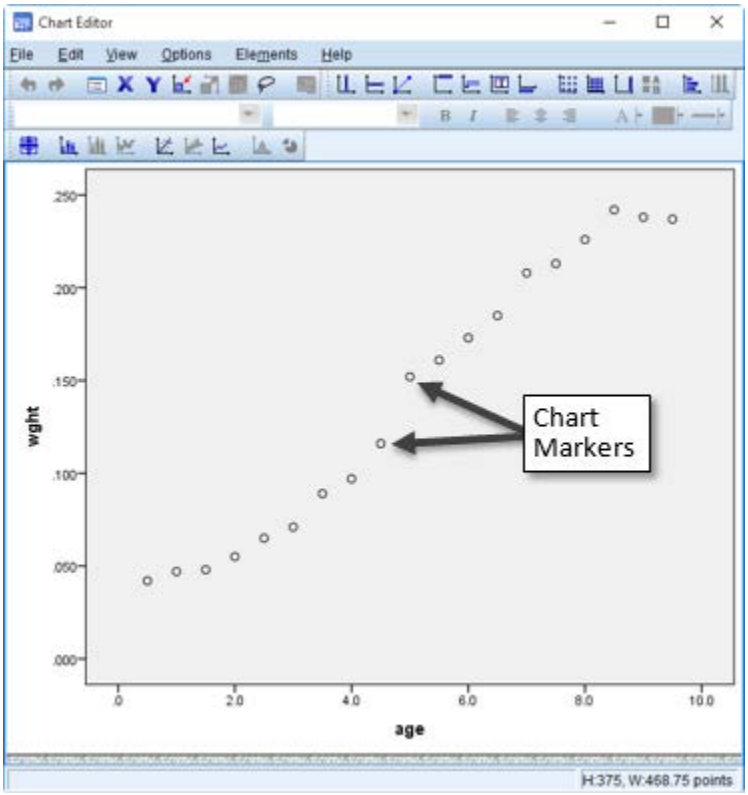

Figure 23 – Chart Editor Window

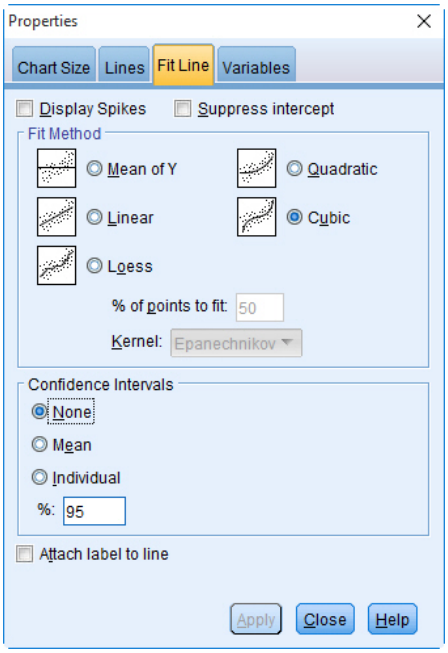

Figure 24 – Fit Line Tab of the Properties Dialog Box

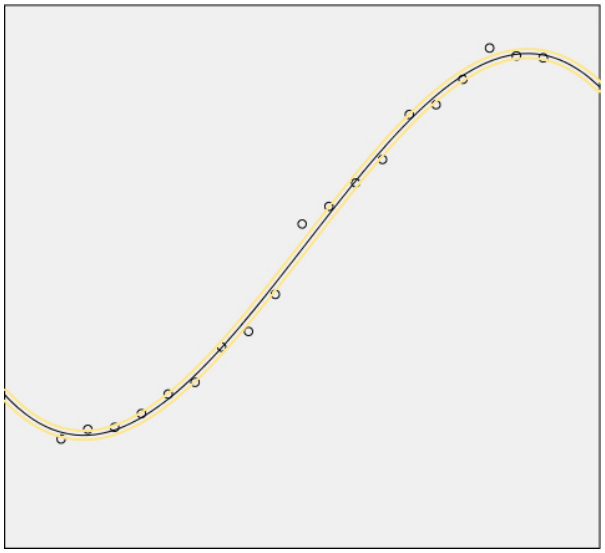

Figure 25 – Fit Line Added to the Scatter Plot

Manipulating the Scales on the X and Y Axes

Adjust the **X axis** and **Y axis** to enhance the overall appearance and readability of a chart. Various elements of the axes can be manipulated (such as **scale**, **ticks and grids**, **number format**, and **axis label**).

To manipulate the scales on the X and Y axes:

1. If necessary, open the **Regression.sav** file, and then run a scatter plot with **Last year sales** assigned to the **Y Axis** and **Years of experience** assigned to the **X Axis**.
2. In the **Output Viewer** window, double-click the chart.
3. In the **Chart Editor** window, right-click a chart marker, and then click **Add Fit Line at Total** on the shortcut menu.
4. In the **Properties** dialog box, on the **Fit Line** tab, deselect the **Attach label to line** check box, and then click the **Apply** button.
5. To select and manipulate the X axis, click the **Select the X axis** button on the **Standard** toolbar.
6. In the **Properties** dialog box, on the **Scale** tab, change the value in the **Lower margin (%)** box to **0**, and then click the **Apply** button (see Figure 26).
7. In the **Properties** dialog, select the **Labels & Ticks** tab (see Figure 27).
8. In the **Major Ticks** section, select the **Display ticks** check box, select **Inside** from the **Style** list, and then click the **Apply** button.

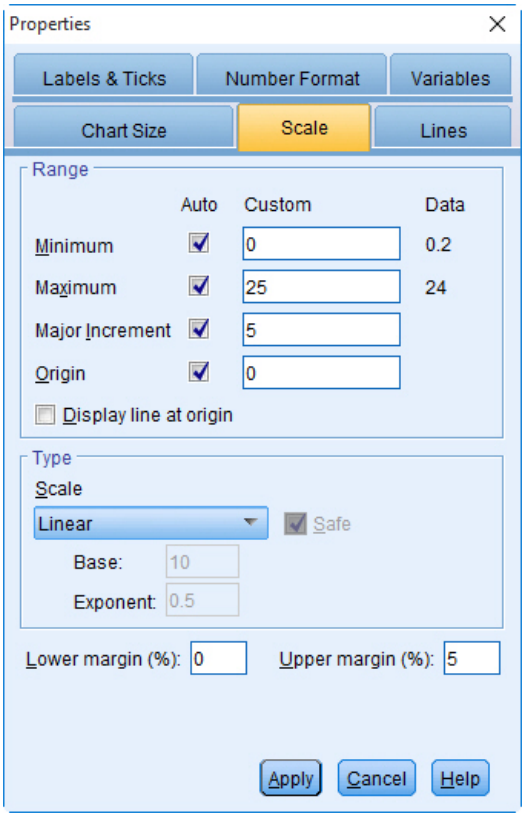

Figure 26 – Scale Tab of the Properties Dialog Box (X Axis)

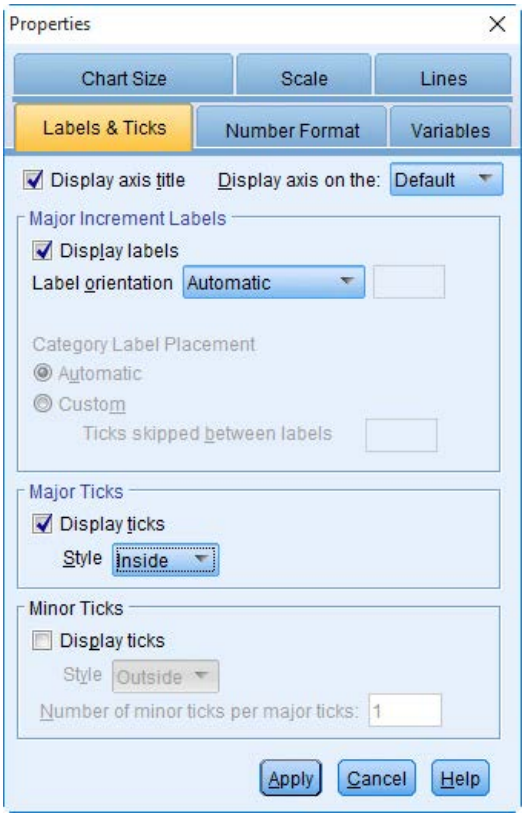

Figure 27 – Labels & Ticks Tab of the Properties Dialog Box (X Axis)

9. Click the **Show Grid Lines** button on the **Standard** toolbar. The **Properties** dialog box displays the **Grid Lines** tab (see Figure 28).
10. Select the **Major ticks only** option button, click the **Apply** button, and then click the **Close** button.

11. To select and manipulate the Y axis, click the **Select the Y axis** button  on the **Standard** toolbar.
12. In the **Properties** dialog box, on the **Scale** tab, change the value in the **Lower margin (%)** box to **0**, click the **Apply** button, and then click the **Close** button (see Figure 29).

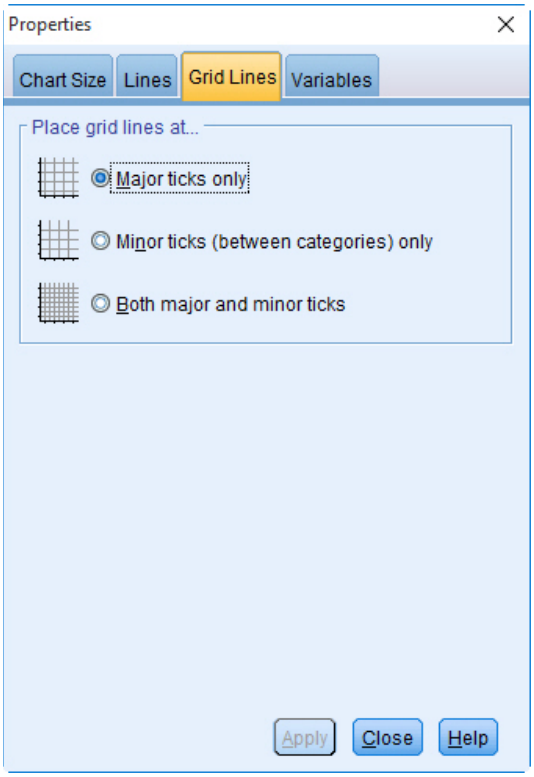

Figure 28 – Grid Lines Tab of the Properties Dialog Box (X Axis)

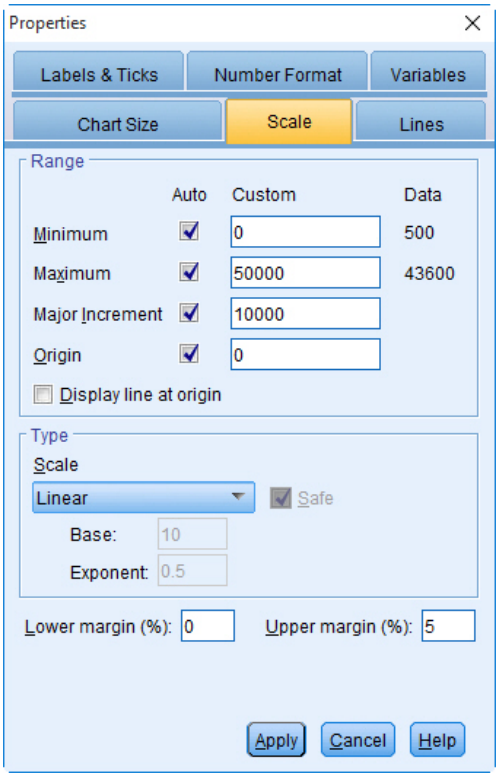

Figure 29 – Scale Tab of the Properties Dialog Box (Y Axis)

NOTE: Below is the chart before and after manipulating the X and Y axes (see Figure 30 and Figure 31).

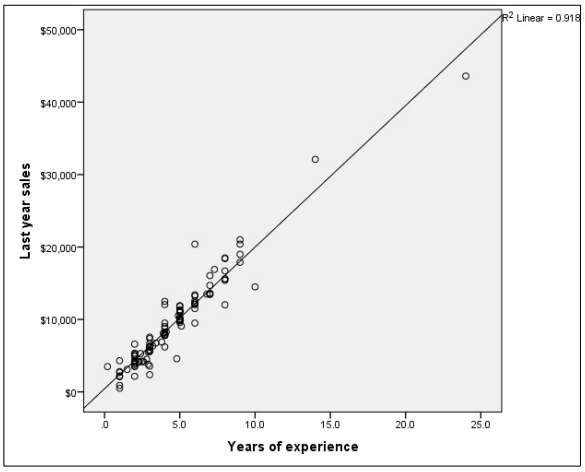

Figure 30 – Chart Before Manipulating the X and Y Axes

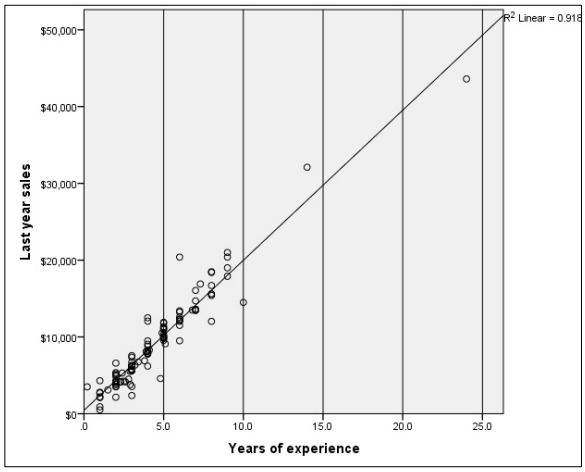

Figure 31 – Chart After Manipulating the X and Y Axes

Adding a Title to the Chart

Adding a title to a chart is a simple process that enhances the chart's appearance.

To add a title to the chart:

1. In the **Chart Editor** window, click the **Insert a title** button on the **Standard** toolbar. A text box with the word **Title** is inserted above the chart and the **Properties** dialog box opens.
2. Delete the placeholder text in the text box, and then type **Relationship Between Last Year Sales and Years of Experience**.
3. To enter a line break, click where you want to break the line, and then press **Shift+Enter**.
4. To format the title, click the border of the text box to select it.
5. In the **Properties** dialog box, select the **Text Style** tab, select the desired font size and style in the **Font** section, select the desired color from the color palette, click the **Apply** button, and then click the **Close** button (see Figure 32).
6. If necessary, move or resize the text box. The changes are applied to the chart (see Figure 33).

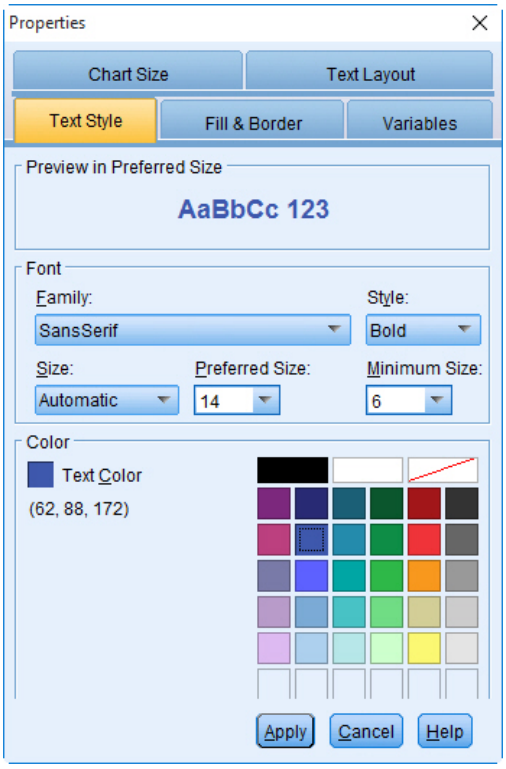

Figure 32 – Text Style Tab of the Properties Dialog Box

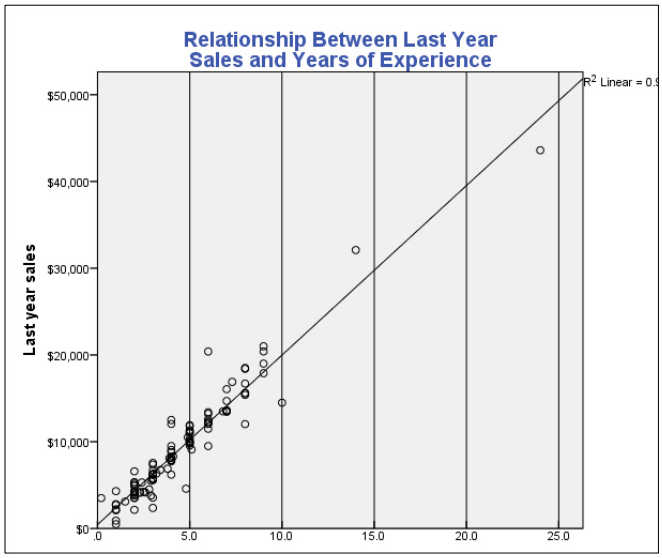

Figure 33 – Title Added at the Top of the Chart

Adding Color to the Chart

All chart elements can be colored differently to add emphasis or to distinguish between elements.

To add color to the chart:

1. In the **Chart Editor** window, select the chart element to change or add color to (such as the chart markers).
2. Click the **Show Properties Window** button on the **Standard** toolbar.
3. In the **Properties** dialog box, select the **Marker** tab (see Figure 34).

4. To change the marker color, select the desired color from the color palette.
5. To change the marker type, click the **Type** arrow in the **Marker** section and select the desired symbol from the list.
6. View the changes in the **Preview** section, click the **Apply** button, and then click the **Close** button. The changes are applied to the chart (see Figure 35).

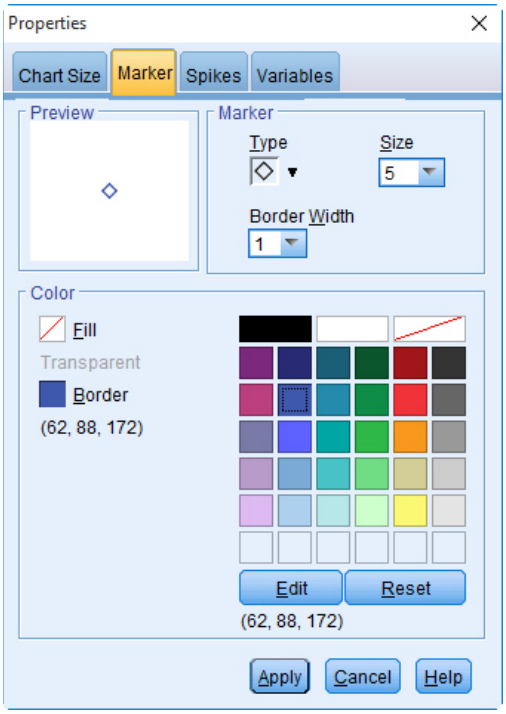

Figure 34 – Marker Tab of the Properties Dialog Box

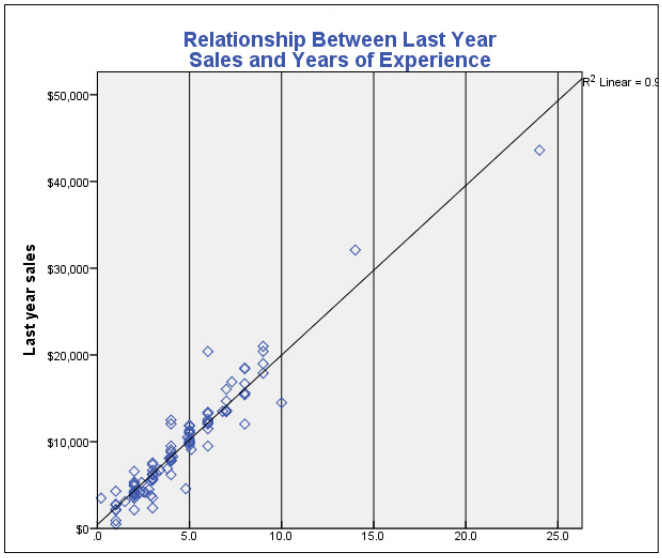

Figure 35 – Chart After Changing the Marker Type and Color

Applying a Background Color

You can change the chart's background color to make it stand out from other chart elements.

To apply a background color:

1. Click in the background area of the chart to select it.
2. Click the **Show Properties Window** button on the **Standard** toolbar.
3. In the **Properties** dialog box, on the **Fill & Border** tab, select the **Fill** swatch, and then select the desired color from the color palette (see Figure 36).

   NOTE: You can also apply a background pattern by clicking the **Pattern** arrow and selecting the desired pattern from the list.
4. Click the **Apply** button, and then click the **Close** button. The changes are applied to the chart (see Figure 37).

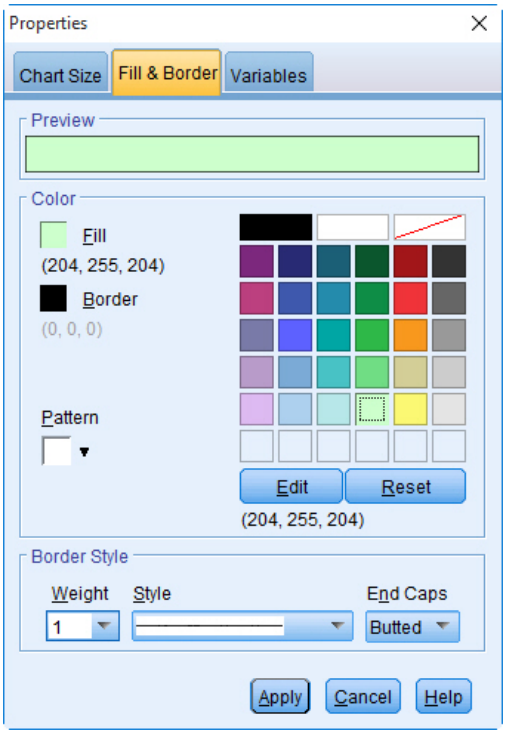

Figure 36 – Fill & Border Tab of the Properties Dialog Box

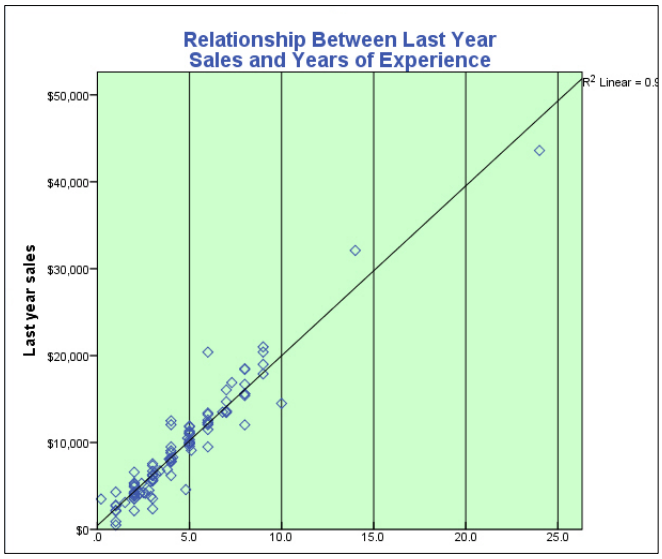

Figure 37 – Chart After Applying a Background Color