Chapter 11
Simple Linear Regression

11.1 Introduction

Chapters 11 and 12 in *Statistics* introduce the topic of regression analysis to the reader. Chapter 11 serves as the introduction of the general concepts of simple linear regression. Simple Linear Regression is how the text introduces the theories and concepts of mathematical modeling to the reader. These topics are then expanded in Chapter 12 of the text.

We will take a similar approach to regression as does the text. We will use Chapter 11 to introduce you to the methods **XLSTAT** offers to work with regression analysis. We will see how **XLSTAT** can be used to calculate both the correlation and the linear modeling ideas that are presented in the text. We will use the chapter examples that are given in the text to illustrate these methods.

Please note the simple linear regression material is covered in Chapter 9 in *A First Course in Statistics*. The following examples from *Statistics* are solved with **XLSTAT** in this chapter:

Excel Companion **Exercise**	**Page**	**Statistics Example**	**Excel File Name**
11.1	152	Example 11.1	STIMULUS
11.2	155	Example 11.3	STIMULUS
11.3	157	Example 11.5	STIMULUS
11.3	159	Examples 11.6/11.7	STIMULUS

11.2 Fitting the Model: The Least Squares Approach

Regression analysis is all about the relationship between variables. Chapters 11 and 12 spend time developing the mathematical modeling of one variable using the values of other related variables. The simplest form of this modeling idea is the linear relationship between two variables. This idea, known as correlation, is studied in Chapter 10 of *Statistics*. We examine how **XLSTAT** calculates correlations below.

Exercise 11.1: Use Example 11.1 found in the *Statistics* text.

Problem: Suppose an experiment involving five subjects is conducted to determine the relationship between the percentage of a certain drug in the bloodstream and the length of time it takes to react to a stimulus. The results are shown in Table 11.1 and are saved in the STIMULUS data file.

Table 11.1

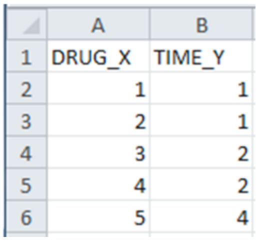
Subject	Percentage, x, of Drug	Reaction Time, y
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Consider the straight-line model, $E(y) = \beta_0 + \beta_1 x$, where y = reaction time (in seconds) and x = percent of drug received. Use the method of least square to estimate the values of β_0 and β_1.

Solution:

We solve Exercise 11.1 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **STIMULUS** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 11.1.

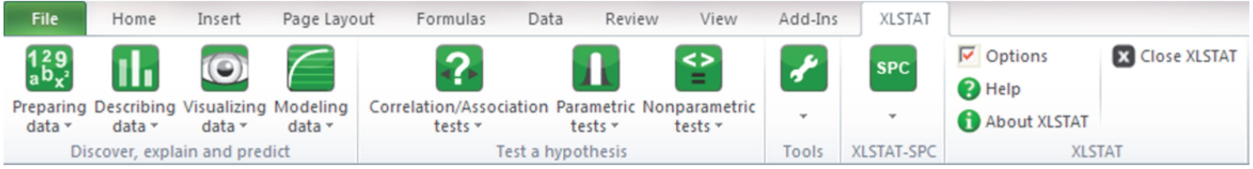
Figure 11.1



	A	B
1	DRUG_X	TIME_Y
2	1	1
3	2	1
4	3	2
5	4	2
6	5	4

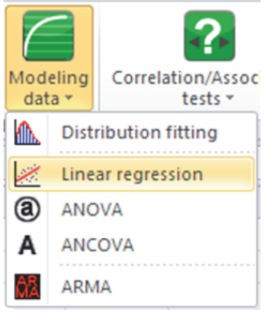
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 11.2.

Figure 11.2



To find the least squares regression model, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 11.3.

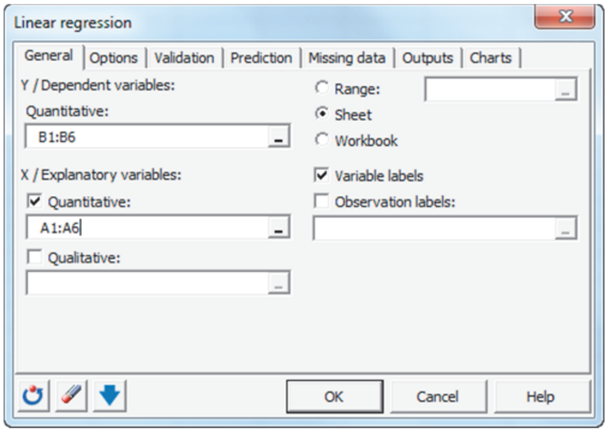
Figure 11.3



This opens the **Linear regression** menu shown in Figures 11.4-11.5. We need to first specify the location of the data that is to be analyzed. In our data set, the data is located in columns A and B, rows 2 – 6, with row 1 being the variable labels. We note that the data in column A represents the independent variable, drug time, and the data in column B represent the dependent variable, reaction time. We specify the column **B** data in the **Quantitative** box for the Dependent variable, *Y*, and the column **A** data in the **Quantitative** box for the independent, or Explanatory, variable, *X*. We also check the **Variable labels** box in this menu.

Click on the **Outputs** tab (shown in Figure 11.5) to specify the type of output desired. To build the least squares prediction equation, we check the **Standardized coefficients** box. Click **OK** to build the model.

Figure 11.4



Y / Dependent variables:

Quantitative: B1:B6

○ Range:
● Sheet
○ Workbook

X / Explanatory variables:

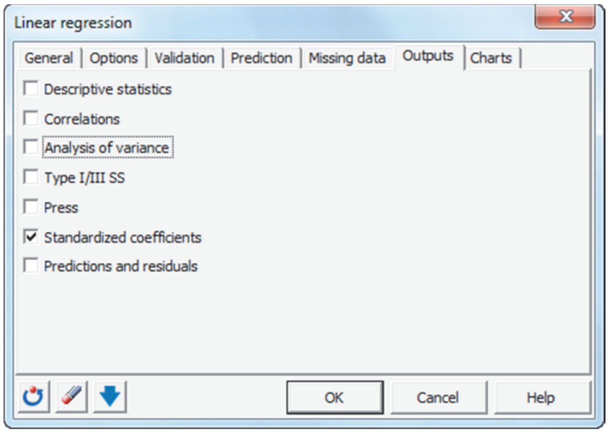
☑ Quantitative: A1:A6

☐ Qualitative:

☑ Variable labels

☐ Observation labels:

Figure 11.5



☐ Descriptive statistics

☐ Correlations

☐ Analysis of variance

☐ Type I/III SS

☐ Press

☑ Standardized coefficients

☐ Predictions and residuals

The XLSTAT output is shown in Figure 11.6.

Figure 11.6

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-0.1000	0.6351	-0.1575	0.8849	-2.1211	1.9211
DRUG_X	0.7000	0.1915	3.6556	0.0354	0.0906	1.3094

Equation of the model:

TIME_Y = -0.10000+0.70000*DRUG_X

We see that the printout above contains the least squares prediction equation. We note that it is identical to the prediction equation provided in the text.

11.3 Testing the Regression Model

The first step in a regression analysis is to estimate the relationship between the two variables by finding the least squares prediction equation. The next step is to determine if there is a statistically significant relationship between the two variables. This involves conducting a test of hypothesis to determine if the independent variable, *x*, is a useful linear predictor of the dependent variable, *y*. We illustrate how to conduct this test in the following example.

Exercise 11.2: Use Example 11.3 found in the *Statistics* text.

Problem: Suppose an experiment involving five subjects is conducted to determine the relationship between the percentage of a certain drug in the bloodstream and the length of time it takes to react to a stimulus. The results are shown in Table 11.2 and are saved in the STIMULUS data file.

Table 11.2

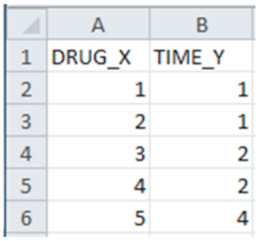
Subject	Percentage, x, of Drug	Reaction Time, y
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Consider the least squares regression equation found in Exercise 11.1. Conduct a test (at $\alpha = .05$) to determine whether the reaction time (y) is linearly related to the amount of drug (x).

Solution:

We solve Exercise 11.2 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **STIMULUS** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 11.7.

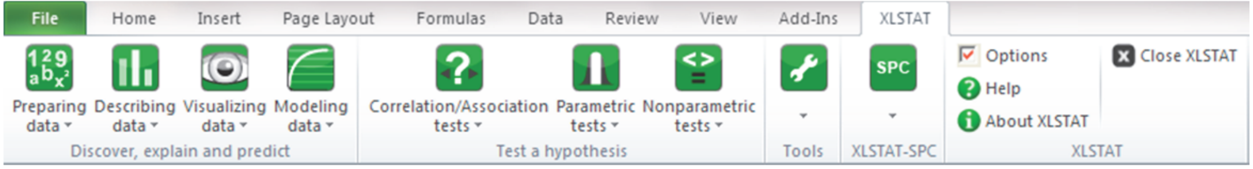
Figure 11.7



	A	B
1	DRUG_X	TIME_Y
2	1	1
3	2	1
4	3	2
5	4	2
6	5	4

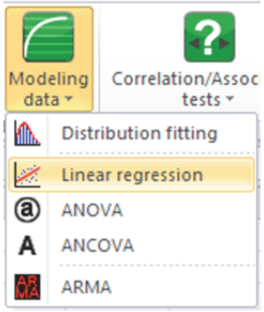
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 11.8.

Figure 11.8



To test the least squares regression model, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 11.9.

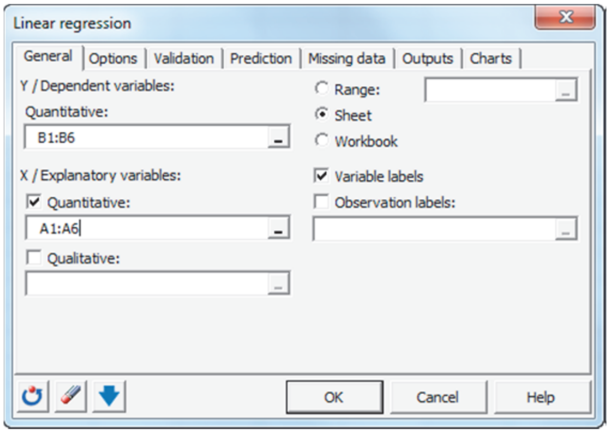
Figure 11.9



This opens the **Linear regression** menu shown in Figures 11.10-11.11. We need to first specify the location of the data that is to be analyzed. In our data set, the data is located in columns A and B, rows 2 – 6, with row 1 being the variable labels. We note that the data in column A represents the independent variable, drug time, and the data in column B represent the dependent variable, reaction time. We specify the column **B** data in the **Quantitative** box for the Dependent variable, *Y*, and the column **A** data in the **Quantitative** box for the independent, or Explanatory, variable, *X*. We also check the **Variable labels** box in this menu.

Click on the **Outputs** tab (shown in Figure 11.11) to specify the type of output desired. To test the least squares prediction equation, we check the **Standardized coefficients** box. Click **OK** to test the model.

Figure 11.10



Y / Dependent variables:

Quantitative:

B1:B6

○ Range:

● Sheet

○ Workbook

X / Explanatory variables:

☑ Quantitative:

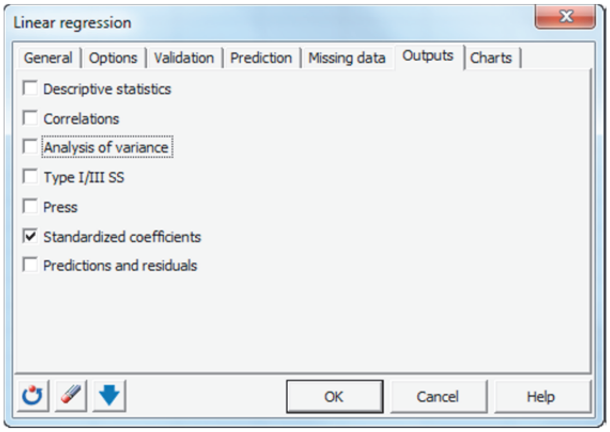
A1:A6

□ Qualitative:

☑ Variable labels

□ Observation labels:

Figure 11.11



□ Descriptive statistics

□ Correlations

□ Analysis of variance

□ Type I/III SS

□ Press

☑ Standardized coefficients

□ Predictions and residuals

The XLSTAT output is shown in Figure 11.12.

Figure 11.12

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-0.1000	0.6351	-0.1575	0.8849	-2.1211	1.9211
DRUG_X	0.7000	0.1915	3.6556	0.0354	0.0906	1.3094

Equation of the model:

TIME_Y = -0.10000+0.70000*DRUG_X

We see that the printout above contains the desired test. We note that the test statistic ($t = 3.6556$) and the p-value ($p = .0354$) are identical to the values provided in the text.

11.4 The Coefficients of Correlation and Determination

The two most common descriptive measures to describe the usefulness of the regression are coefficient of correlation and the coefficient of determination. Both descriptive measures are very easy to generate using **XLSTAT**. We illustrate with the following example.

Exercise 11.3: Use Example 11.5 found in the *Statistics* text.

Problem: Suppose an experiment involving five subjects is conducted to determine the relationship between the percentage of a certain drug in the bloodstream and the length of time it takes to react to a stimulus. The results are shown in Table 11.3 and are saved in the STIMULUS data file.

Table 11.3

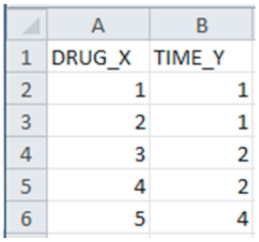
Subject	Percentage, x, of Drug	Reaction Time, y
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Calculate the coefficients of correlation and determination for the drug reaction data.

Solution:

We solve Exercise 11.3 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **STIMULUS** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 11.13.

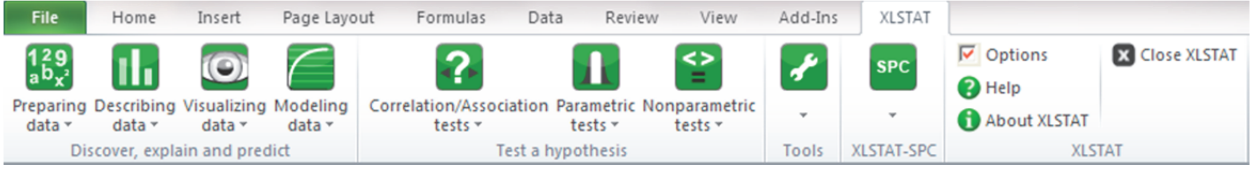
Figure 11.13



	A	B
1	DRUG_X	TIME_Y
2	1	1
3	2	1
4	3	2
5	4	2
6	5	4

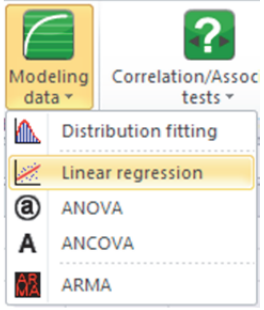
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 11.14.

Figure 11.14



To calculate the coefficients of correlation and determination, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 11.15.

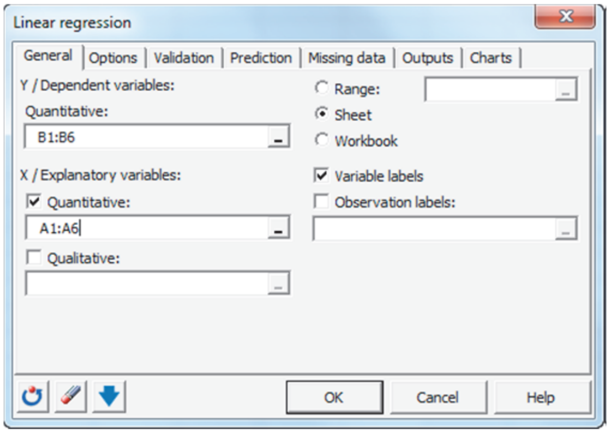
Figure 11.15



This opens the **Linear regression** menu shown in Figures 11.16-11.17. We need to first specify the location of the data that is to be analyzed. In our data set, the data is located in columns A and B, rows 2 – 6, with row 1 being the variable labels. We note that the data in column A represents the independent variable, drug time, and the data in column B represent the dependent variable, reaction time. We specify the column **B** data in the **Quantitative** box for the Dependent variable, *Y*, and the column **A** data in the **Quantitative** box for the independent, or Explanatory, variable, *X*. We also check the **Variable labels** box in this menu.

Click on the **Outputs** tab (shown in Figure 11.17) to specify the type of output desired. To calculate the coefficients of correlation and determination, we check the **Correlations** box. Click **OK** to calculate the coefficients of correlation and determination.

Figure 11.16



Y / Dependent variables:

Quantitative: B1:B6

X / Explanatory variables:

☑ Quantitative: A1:A6

☐ Qualitative:

○ Range:

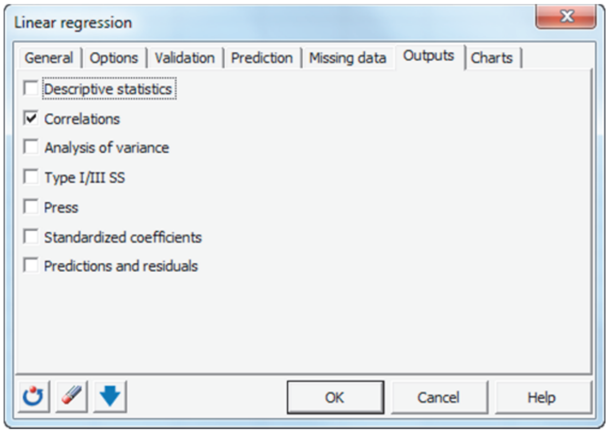
● Sheet

○ Workbook

☑ Variable labels

☐ Observation labels:

Figure 11.17



☐ Descriptive statistics

☑ Correlations

☐ Analysis of variance

☐ Type I/III SS

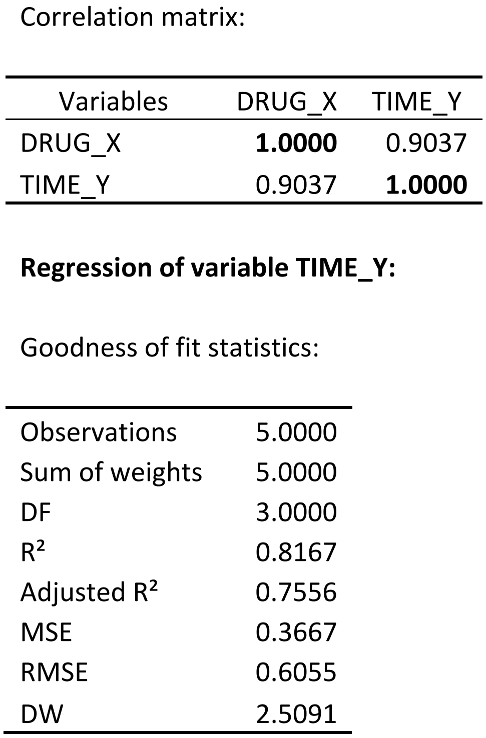
☐ Press

☐ Standardized coefficients

☐ Predictions and residuals

The XLSTAT output is shown in Figure 11.18.

Figure 11.18



```
Correlation matrix:

Variables      DRUG_X   TIME_Y
DRUG_X         1.0000   0.9037
TIME_Y         0.9037   1.0000

Regression of variable TIME_Y:

Goodness of fit statistics:

Observations      5.0000
Sum of weights    5.0000
DF                3.0000
R²                0.8167
Adjusted R²       0.7556
MSE               0.3667
RMSE              0.6055
DW                2.5091
```

We see that the printout above contains both the coefficient of correlation (r = .9037) and the coefficient of determination (r^2 = .8167). We compare these values to those found in the text and find they are identical.

11.5 Using the Model for Estimation and Prediction

The final step in the simple linear regression analysis is to use the model to estimate and predict values of the dependent variable, y, for specified settings of the independent variable, x. We illustrate this procedure using the following example.

Exercise 11.4: Use Example 11.6 and 11.7 found in the *Statistics* text.

Problem: Suppose an experiment involving five subjects is conducted to determine the relationship between the percentage of a certain drug in the bloodstream and the length of time it takes to react to a stimulus. The results are shown in Table 11.4 and are saved in the STIMULUS data file.

Table 11.4

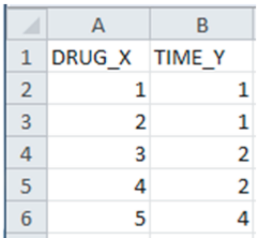
Subject	Percentage, x, of Drug	Reaction Time, y
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

a. Find a 95% confidence interval for the mean for all reaction times when the concentration for the drug in the bloodstream is 4%.
b. Find a 95% prediction interval for a single reaction time when the concentration for the drug in the bloodstream is 4%.

Solution:

We solve Exercise 11.4 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **STIMULUS** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 11.19.

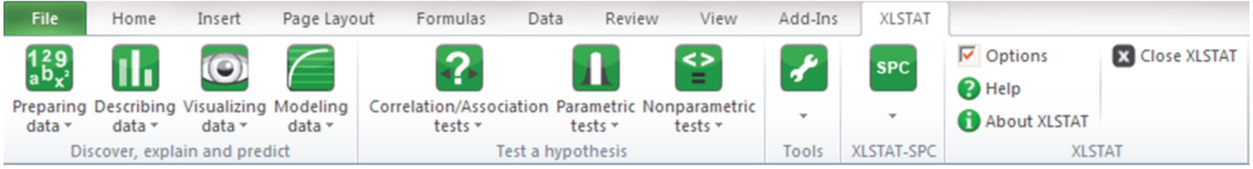
Figure 11.19



	A	B
1	DRUG_X	TIME_Y
2	1	1
3	2	1
4	3	2
5	4	2
6	5	4

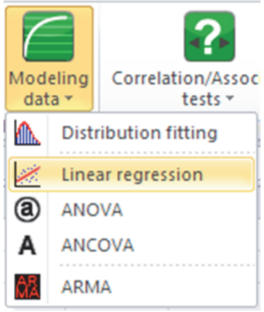
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 11.20.

Figure 11.20



To create the desired confidence intervals, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 11.21.

Figure 11.21

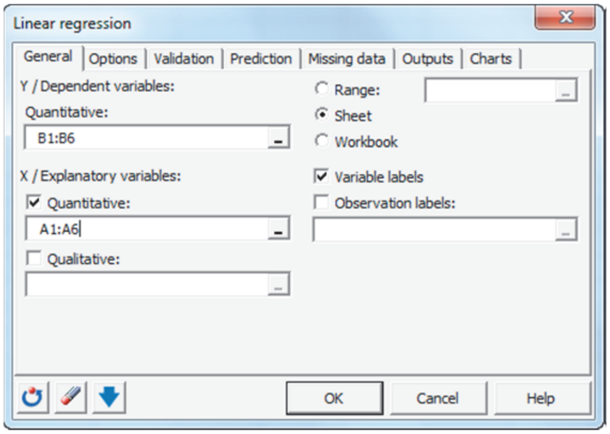


This opens the **Linear regression** menu shown in Figures 11.22-11.24. We need to first specify the location of the data that is to be analyzed. In our data set, the data is located in columns A and B, rows 2 – 6, with row 1 being the variable labels. We note that the data in column A represents the independent variable, drug time, and the data in column B represent the dependent variable, reaction time. We specify the column **B** data in the **Quantitative** box for the Dependent variable, *Y*, and the column **A** data in the **Quantitative** box for the independent, or Explanatory, variable, *X*. We also check the **Variable labels** box in this menu.

Click on the **Outputs** tab (shown in Figure 11.23) to specify the type of output desired. To create the confidence and prediction intervals, we check the **Predictions and residuals** (along with any other analyses we desire) box.

Click on the **Prediction** tab (shown in Figure 11.24) to identify the values of the independent variable, *x*, that we want to predict for. **XLSTAT** will automatically create the prediction and confidence intervals for the values of *x* contained in the data set. To use a specific value of *x*, we make sure that value is located in a cell within the worksheet and identify that cell in the **Quantitative Explanatory variables** box. In this example, we check the **Prediction** box and identify the independent variable value, 4, located in cell **A5** in the **Quantitative Explanatory variables**

Figure 11.22



Y / Dependent variables:

Quantitative: B1:B6

X / Explanatory variables:

☑ Quantitative: A1:A6

□ Qualitative:

○ Range:

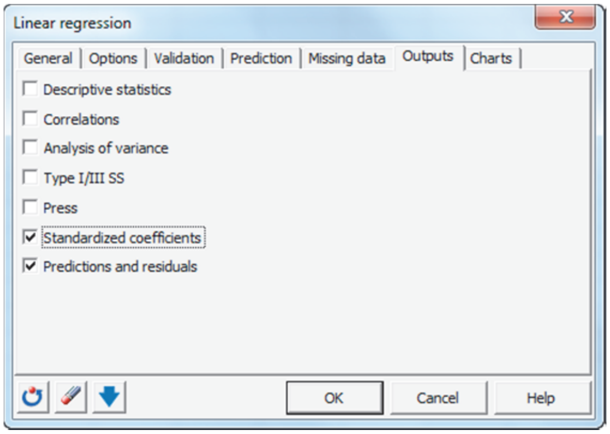
● Sheet

○ Workbook

☑ Variable labels

□ Observation labels:

Figure 11.23



□ Descriptive statistics

□ Correlations

□ Analysis of variance

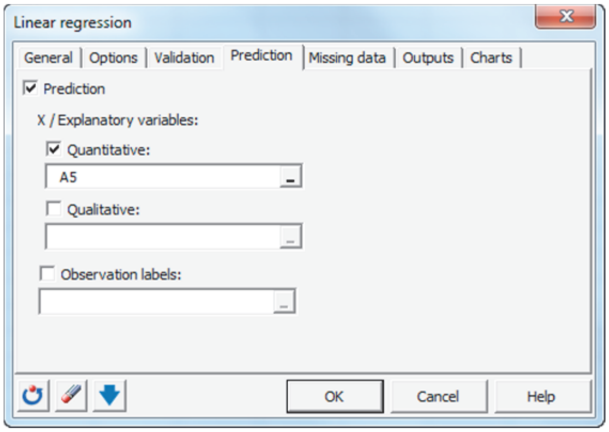
□ Type I/III SS

□ Press

☑ Standardized coefficients

☑ Predictions and residuals

Figure 11.24



☑ Prediction

X / Explanatory variables:

☑ Quantitative: A5

□ Qualitative:

□ Observation labels:

The XLSTAT output is shown in Figure 11.25.

Figure 11.25

Predictions and residuals:

				Lower bound	Upper bound	Lower bound	Upper bound
				95%	95%	95%	95%
Observation	DRUG_X	TIME_Y	Pred(TIME_Y)	(Mean)	(Mean)	(Observation)	(Observation)
Obs1	1.0000	1.0000	0.6000	-0.8927	2.0927	-1.8376	3.0376
Obs2	2.0000	1.0000	1.3000	0.2445	2.3555	-0.8972	3.4972
Obs3	3.0000	2.0000	2.0000	1.1382	2.8618	-0.1110	4.1110
Obs4	4.0000	2.0000	2.7000	1.6445	3.7555	0.5028	4.8972
Obs5	5.0000	4.0000	3.4000	1.9073	4.8927	0.9624	5.8376

Predictions for the new observations:

Observation	Pred(TIME_Y)	Lower bound 95% (Mean)	Upper bound 95% (Mean)	Lower bound 95% (Observation)	Upper bound 95% (Observation)
PredObs1	2.7000	1.6445	3.7555	0.5028	4.8972

The top part of the printout contains the confidence and prediction intervals for each of the five sampled data values. The bottom part contains the confidence and prediction interval for the one value of the independent variable, $x = 4$, that we specified. We compare these values to the endpoints found in the text and see that they are identical.

11.6 Technology Lab

The Technology Lab consists of problems for the student to practice the techniques presented in each lesson. Each problem is taken from the homework exercises within the *Statistics* text and includes an **Excel** data set (when applicable) that should be used to create the desired output. The completed output has been included with each problem so that the student can verify that he/she is generating the correct output.

1. **Ideal height of your mate.** Anthropologists theorize that humans tend to choose mates who are similar to themselves. This includes choosing mates who are similar in height. To test this theory, a study was conducted on 147 Cornell University students. (*Chance*, Summer 2008) Each student was asked to select the height of their ideal spouse or life partner. The researchers fit the simple linear regression model, $E(y) = \beta_0 + \beta_1 x$, where y = ideal partner's height (in inches) and x = student's height (in inches). The data for the study are saved in the IDHEIGHT data file.

a. Fit the model to the data and find the least squares prediction equation.
b. Use the results to determine if a positive linear relationship exists between the students' heights and their ideal partners' heights.
c. Calculate the coefficients of correlation and determination for this data.
d. Use the model to create a 95% confidence interval for $E(y)$ and a 95% prediction interval for y for a student who is 68 inches tall.

XLSTAT Output

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	86.0177	4.7863	17.9717	< 0.0001	76.5578	95.4776
Height	-0.2605	0.0704	-3.7021	0.0003	-0.3995	-0.1214

Equation of the model:

IdealHt = 86.01770-0.26046*Height

Correlation matrix:

Variables	Height	IdealHt
Height	**1.0000**	-0.2939
IdealHt	-0.2939	**1.0000**

Goodness of fit statistics:

Observations	147.0000
Sum of weights	147.0000
DF	145.0000
R^2	0.0864
Adjusted R^2	0.0801
MSE	13.2741
RMSE	3.6434
DW	0.7921

Predictions for the new observations:

Observation	Pred(IdealHt)	Std. dev. on pred. (Mean)	Lower bound 95% (Mean)	Upper bound 95% (Mean)	Lower bound 95% (Observation)	Upper bound 95% (Observation)
PredObs1	68.3068	0.3006	67.7127	68.9009	61.0813	75.5322

2. **Lobster fishing study.** An article in the *Bulletin of Marine Science* (April, 2010) studied teams of fisherman fishing for the red spiny lobster in Baja California Sur, Mexico. Two variables measured for each of 8 teams from the Punta Abreojos (PA) fishing cooperative were y = total catch of lobsters (in Kilograms) during the season and x = average percentage of traps allocated per day to exploring areas of unknown catch (called *search frequency*). These data, saved in the TRAPSPAC data file, are shown below.

Total Catch	Search Frequency
2785	35
6535	21
6695	26
4891	29
4937	23
5727	17
7019	21
5735	20

a. Fit the model to the data and find the least squares prediction equation.
b. Use the results to determine if a negative linear relationship exists between the total catch and the search frequency values.
c. Calculate the coefficients of correlation and determination for this data.
d. Use the model to create a 95% confidence interval for *E(y)* and a 95% prediction interval for *y* for a search frequency of 25%.

XLSTAT Output

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	9658.2436	1617.7621	5.9701	0.0010	5698.6308	13617.8564
Search Frequency	-171.5726	65.7578	-2.6092	0.0402	-332.5206	-10.6247

Equation of the model:

Total Catch = 9658.24359-171.57265*Search Frequency

Correlation matrix:

Variables	Search Frequency	Total Catch
Search Frequency	**1.0000**	-0.7291
Total Catch	-0.7291	**1.0000**

Goodness of fit statistics:

Observations	8.0000
Sum of weights	8.0000
DF	6.0000
R²	0.5315
Adjusted R²	0.4535
MSE	1011836.5442
RMSE	1005.9009
DW	1.9061

Predictions for the new observations:

Observation	Pred(Total Catch)	Lower bound 95% (Mean)	Upper bound 95% (Mean)	Lower bound 95% (Observation)	Upper bound 95% (Observation)
PredObs1	5368.9274	4483.7139	6254.1408	2752.5956	7985.2591