Chapter 12
Multiple Regression and Model Building

12.1 Introduction

Chapter 12 in *Statistics* introduces the topic of **multiple** regression analysis to the reader. While Chapter 11 served as the introduction to the general concepts of simple linear regression, Chapter 12 expands these concepts to modeling with several variables. In addition, Chapter 12 examines some of the problems associated with regression analysis and gives methods of detecting and solving these problems.

We utilize Chapter 12 examples to build on the linear regression base developed in the preceding chapter. Through the use of the linear regression data analysis tool, **XLSTAT** allows the user to build more sophisticated models than the linear models of Chapter 11. We examine both the model building methods and the residual analysis options offered within **XLSTAT**. We will use the chapter examples that are given in the text to illustrate these methods.

Please note that the multiple regression material is not covered 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
12.1	166	Example 12.3	GFCLOCKS
12.2	170	Example 12.4	GFCLOCKS
12.3	173	Example 12.6	GFCLOCKS
12.4	176	Example 12.12	CARNATIONS
12.5	178	Example 12.13	EXECSAL
12.6	181		GFCLOCKS

12.2 Multiple Regression Models

We have seen in Chapter 11 how to use XLSTAT to build a simple linear regression model using one independent variable, x. The next step in our regression process in to add more independent variables into the regression model. XLSTAT allows for this using the same menus as seen in the simple linear regression chapter. We use an example from the text below.

Exercise 12.1: We use Example 12.3 found in the *Statistics* text.

Problem: A collector of antique grandfather clocks knows that the price (y) received for the clocks increases linearly with the age (x_1) of the clocks. Moreover, the collector hypothesizes that the auction price (y) of the clocks will increase linearly as the number of bidders (x_2) increases. Thus, the following model is hypothesized:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

A sample of 32 auction prices of grandfather clocks, along with their age and the number of bidders is shown in Table 12.1 and saved in the GFCLOCKS data file. Use XLSTAT to fit the model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, and answer the following question:

a. Test the hypothesis that the mean auction price of a clock increases as the number of bidders increases when age is held constant, that is, $\beta_2 > 0$. Use $\alpha = .05$.

Table 12.1

Age(x_1)	Number of Bidders (x_2)	Auction Price (y)	Age(x_1)	Number of Bidders (x_2)	Auction Price (y)
127	13	1,235	170	14	2,131
125	12	1,080	182	8	1,550
127	7	845	162	12	1,884
150	9	1,522	184	10	2,041
156	6	1,047	143	6	845
182	12	1,979	159	9	1,483
156	12	1,822	108	14	1,055
132	10	1,253	175	8	1,545
137	9	1,297	108	6	729
123	9	946	179	9	1,792
137	15	1,713	121	15	1,175
127	12	1,024	187	8	1,593
137	8	1,147	121	7	785
153	6	1,092	125	7	744
127	13	1,152	194	5	1,356
126	10	1,336	168	7	1,262

Solution:

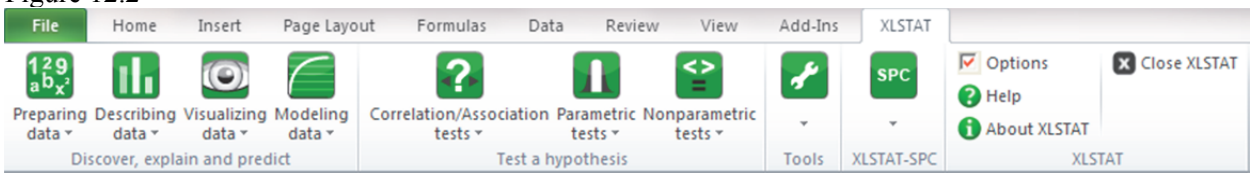
We solve Exercise 12.1 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **GFCLOCKS** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 12.1.

Figure 12.1

	A	B	C
1	AGE	NUMBIDS	PRICE
2	127	13	1235
3	115	12	1080
4	127	7	845
5	150	9	1522
6	156	6	1047
7	182	11	1979
8	156	12	1822
9	132	10	1253
10	137	9	1297
11	113	9	946
12	137	15	1713

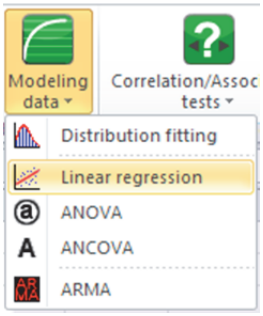
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 12.2.

Figure 12.2



To find the least squares regression model, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 12.3.

Figure 12.3



This opens the **Linear regression** menu shown in Figures 12.4-12.5. 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 thru C, rows 2 – 33, with row 1 being the variable labels. We note that the data in columns A and B represents the independent variables, age and number of bidders, and the data in column C represent the dependent variable, price. We specify the column **C** data in the **Quantitative** box for the Dependent variable, *Y*, and the column **A and B** 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 12.5) to specify the type of output desired. To generate the typical regression output, we check the **Descriptive Statistics, Correlations, Analysis of Variance,** and **Standardized coefficients** boxes. Click **OK** to build the model.

Figure 12.4

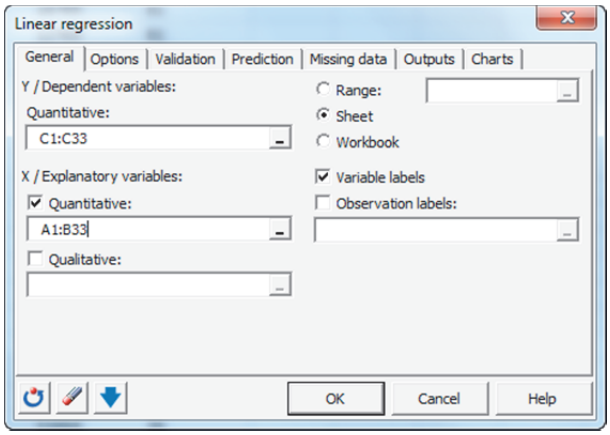
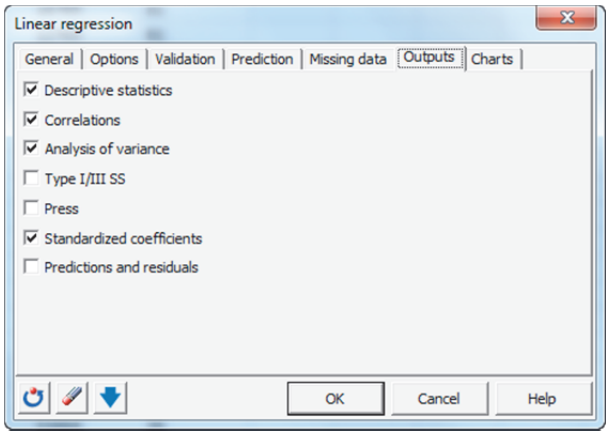


Figure 12.5



The XLSTAT output is shown in Figure 12.6.

Figure 12.6

Correlation matrix:

Variables	AGE	NUMBIDS	PRICE
AGE	**1.0000**	-0.2537	0.7296
NUMBIDS	-0.2537	**1.0000**	0.3952
PRICE	0.7296	0.3952	**1.0000**

Goodness of fit statistics:

Observations	32.0000
Sum of weights	32.0000
DF	29.0000
R^2	0.8923
Adjusted R^2	0.8849
MSE	17818.1565
RMSE	133.4847
DW	1.8720

Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	4283062.9601	2141531.4801	120.1882	< 0.0001
Error	29	516726.5399	17818.1565		
Corrected Total	31	4799789.5000			

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-1338.9513	173.8095	-7.7036	< 0.0001	-1694.4316	-983.4711
AGE	12.7406	0.9047	14.0820	< 0.0001	10.8902	14.5910
NUMBIDS	85.9530	8.7285	9.8474	< 0.0001	68.1011	103.8048

Figure 12.4 shows the typical regression output provided by XLSTAT. In addition to the test for β_2 desired in this exercise, XLSTAT provides many other useful regression results. We point out the most important features of the printout generated by XLSTAT. Compare these values to those shown in the *Statistics* text.

XLSTAT Printout Values	Description of Values
R Square = .8923	Coefficient of determination
s = 133.5	Standard Deviation
Constant Coefficient = -1338.95 13	Estimate of β_0
AGE Coefficient = 12.7406	Estimate of β_1
NUMBIDS Coefficient = 85.9530	Estimate of β_2
t-ratio for NUMBIDS = 9.8474	Test Statistic for testing β_2
P-value for NUMBIDS = 0.0001	P-value for testing β_2
t-ratio for AGE = 14.0820	Test Statistic for testing β_1
P-value for AGE = 0.0001	P-value for testing β_1
F-ratio = 120.1882	Global F-Test Statistic
P-value for F-test = 0.0001	P-value for Global F-test

We find the t-value for the test of β_2 is equal to t = 9.85 and the p-value for the desired test is p = .0001. The estimates of the β coefficients can be found in the Value column in the printout. Our estimates of β_0, β_1, and β_2 are -1338.9513, 12.7406 and 85.9530, respectively. We refer you to the text for more detailed information regarding the interpretations and conclusion that should be made for these values.

The next step of a regression analysis is to test all the hypothesized variables simultaneously. We refer to this process as checking the usefulness of the model. This process is illustrated in the following example.

Exercise 12.2: We use Example 12.4 found in the *Statistics* text.

Problem: A collector of antique grandfather clocks knows that the price received for the clocks increases linearly with the age of the clocks. Moreover, the collector hypothesizes that the auction price of the clocks will increase linearly as the number of bidders increase. Thus, the following model is hypothesized:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

A sample of 32 auction prices of grandfather clocks, along with their age and the number of bidders is shown in Table 12.1. The model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ is fit to the data. Use XLSTAT to :

a. Find and interpret the adjusted coefficient of determination, R_a^2.

b. Conduct the global F-test of model usefulness at the $\alpha = .05$ level of significance.

Solution:

We need to generate the multiple regression model hypothesized above using XLSTAT. The printout generated must include the adjusted coefficient of determination and the global-F test and p-value information. Fortunately, the standard XLSTAT regression output yields both of the desired values.

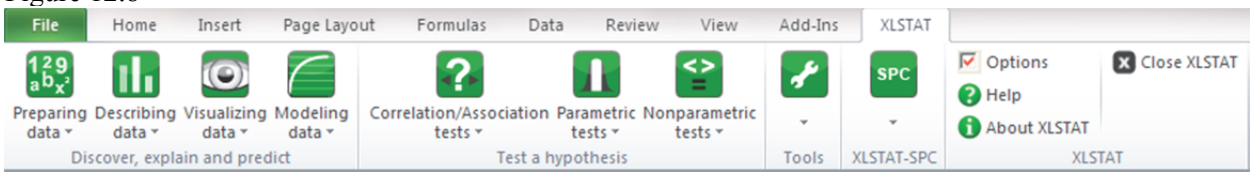
We solve Exercise 12.2 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **GFCLOCKS** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 12.7.

Figure 12.7

	A	B	C
1	AGE	NUMBIDS	PRICE
2	127	13	1235
3	115	12	1080
4	127	7	845
5	150	9	1522
6	156	6	1047
7	182	11	1979
8	156	12	1822
9	132	10	1253
10	137	9	1297
11	113	9	946
12	137	15	1713

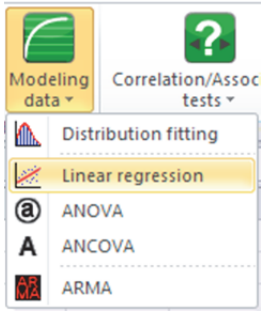
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 12.8.

Figure 12.8



To find the least squares regression model, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 12.9.

Figure 12.9



This opens the **Linear regression** menu shown in Figures 12.10-12.11. 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 thru C, rows 2 – 33, with row 1 being the variable labels. We note that the data in columns A and B represents the independent variables, age and number of bidders, and the data in column C represent the dependent variable, price. We specify the column **C** data in the **Quantitative** box for the Dependent variable, *Y*, and the column **A and B** 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 12.11) to specify the type of output desired. To generate the typical regression output, we check the **Descriptive Statistics, Correlations, Analysis of Variance,** and **Standardized coefficients** boxes. Click **OK** to build the model.

Figure 12.10

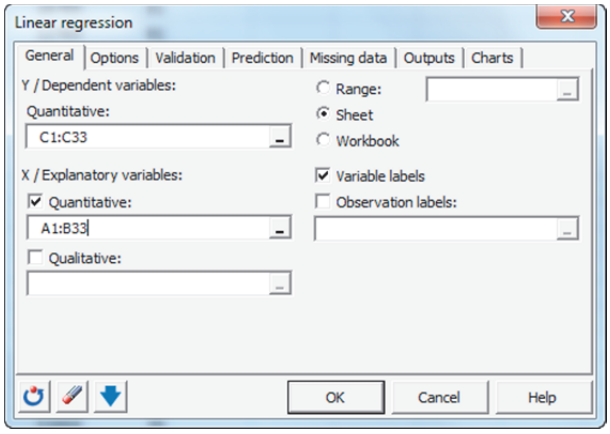
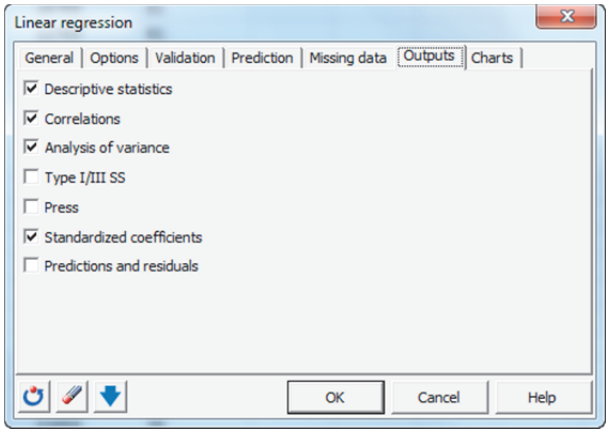


Figure 12.11



The XLSTAT output is shown in Figure 12.12.

Figure 12.12

Correlation matrix:

Variables	AGE	NUMBIDS	PRICE
AGE	**1.0000**	-0.2537	0.7296
NUMBIDS	-0.2537	**1.0000**	0.3952
PRICE	0.7296	0.3952	**1.0000**

Goodness of fit statistics:

Observations	32.0000
Sum of weights	32.0000
DF	29.0000
R^2	0.8923
Adjusted R^2	0.8849
MSE	17818.1565
RMSE	133.4847
DW	1.8720

Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	4283062.9601	2141531.4801	120.1882	< 0.0001
Error	29	516726.5399	17818.1565		
Corrected Total	31	4799789.5000			

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-1338.9513	173.8095	-7.7036	< 0.0001	-1694.4316	-983.4711
AGE	12.7406	0.9047	14.0820	< 0.0001	10.8902	14.5910
NUMBIDS	85.9530	8.7285	9.8474	< 0.0001	68.1011	103.8048

The adjusted coefficient of determination R_a^2 is listed as the Adjusted R^2 in the Goodness of Fit Statistics table above. The R_a^2 value of $R_a^2 = .8849$ can be compared to the value shown in the text. The global F statistic and p-value are shown in the Analysis of Variance table in the printout above. The global F statistic is F = 120.1882 and the p-value is 0.0001, which are identical to the values shown in the text.

The next step in the model building process of a regression analysis is to add various types of regression terms to the model. Whether the terms added are interactions, quadratics, or qualitative terms, the process within XLSTAT is the same. We illustrate this process by adding an interaction component to the preceding example to illustrate. Please note that the process of adding quadratic and qualitative terms to the regression model is identical to the process demonstrated in the next example.

Exercise 12.3: We use Example 12.6 found in the *Statistics* text.

Problem: Suppose the collector of grandfather clocks, having observed many auctions, believes that the *rate of increase* of the auction price with age will be driven upward by a large number of bidders. Thus, instead of a relationship in which the rate of the price is the same for any number of bidders, the collector believes the slope of the price-age relationship increases as the number of bidders increases. Consequently, the interaction model is proposed:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

a. Test the overall utility of the model using the global F-test at $\alpha = .05$.
b. Test the hypothesis (at $\alpha = .05$) that the price-age slope increases as the number of bidders increases - that is, that age and number of bidders, x_2, interact positively.

Solution:

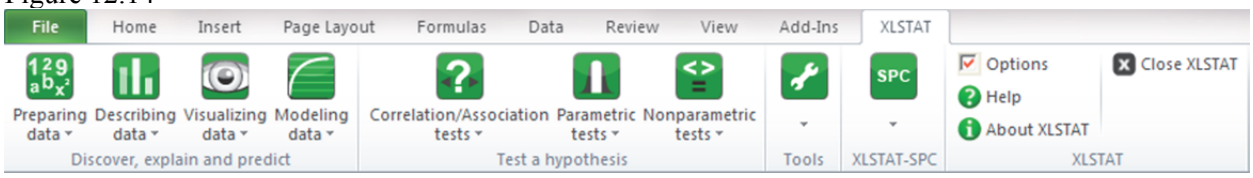
We solve Exercise 12.2 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **GFCLOCKS** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 12.13.

Figure 12.13

	A	B	C	D
1	AGE	NUMBIDS	AGE-BID	PRICE
2	127	13	1651	1235
3	115	12	1380	1080
4	127	7	889	845
5	150	9	1350	1522
6	156	6	936	1047
7	182	11	2002	1979
8	156	12	1872	1822
9	132	10	1320	1253
10	137	9	1233	1297
11	113	9	1017	946
12	137	15	2055	1713

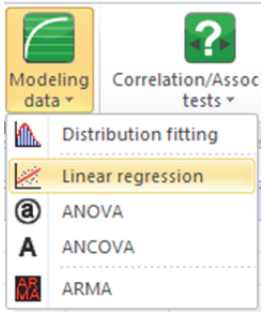
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 12.14.

Figure 12.14



To find the least squares regression model, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 12.15.

Figure 12.15



This opens the **Linear regression** menu shown in Figures 12.16-12.17. 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 thru D, rows 2 – 33, with row 1 being the variable labels. We note that the data in columns A - C represents the independent variables, age, number of bidders, and the interaction between them, and the data in column D represent the dependent variable, price. We specify the column **D** data in the **Quantitative** box for the Dependent variable, *Y*, and the column **A thru C** 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 12.17) to specify the type of output desired. To generate the typical regression output, we check the **Descriptive Statistics, Correlations, Analysis of Variance,** and **Standardized coefficients** boxes. Click **OK** to build the model.

Figure 12.16

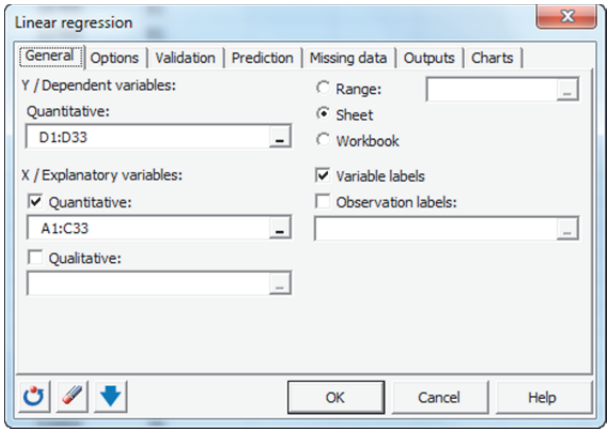
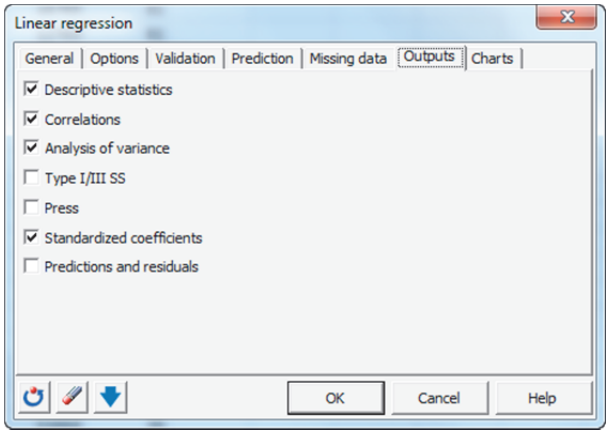


Figure 12.17



The XLSTAT output is shown in Figure 12.18.

Figure 12.18

Correlation matrix:

Variables	AGE	NUMBIDS	AGE-BID	PRICE
AGE	**1.0000**	-0.2537	0.3635	0.7296
NUMBIDS	-0.2537	**1.0000**	0.7916	0.3952
AGE-BID	0.3635	0.7916	**1.0000**	0.8585
PRICE	0.7296	0.3952	0.8585	**1.0000**

Goodness of fit statistics:

Observations	32.0000
Sum of weights	32.0000
DF	28.0000
R^2	0.9539
Adjusted R^2	0.9489
MSE	7905.7905
RMSE	88.9145
DW	2.4190

Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	3	4578427.3668	1526142.4556	193.0411	< 0.0001
Error	28	221362.1332	7905.7905		
Corrected Total	31	4799789.5000			

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	320.4580	295.1413	1.0858	0.2868	-284.1115	925.0275
AGE	0.8781	2.0322	0.4321	0.6690	-3.2845	5.0408
NUMBIDS	-93.2648	29.8916	-3.1201	0.0042	-154.4950	-32.0346
AGE-BID	1.2978	0.2123	6.1123	< 0.0001	0.8629	1.7328

The global F-test is shown on the printout as F-ratio = 193.0411 and the t-test for the interaction is shown as t-ratio = 6.1123. Compare these values to the ones shown in the text.

12.3 Comparing Two Regression Models

We have seen how XLSTAT can be used to fit regression models with just quantitative variables and regression models with just qualitative variables. For more complicated models, XLSTAT allows the user to input both quantitative and qualitative variables into a single multiple regression model. By specifying the appropriate columns in the Excel data set to the appropriate locations within the XLSTAT linear regression menu, any number of quantitative and qualitative variables can be added to the regression model.

The final step in the model building topic is to develop a method that allows the user to compare two regression models to determine which is the better predictor of the dependent variable. Section 12.9 in the text details the F-test for testing a portion of the regression model. By fitting two separate models within XLSTAT, it is possible to calculate the partial-F test statistic that the book details. We demonstrate with the following example.

Exercise 12.4: We use Example 12.12 found in the *Statistics* text.

Problem: A botanist conduct an experiment to study the growth of carnations as a function of the temperature x_1 (^{o}F) in a greenhouse and the amount of fertilizer x_2 [kilograms (kg) per plot] to the soil. Twenty-seven plots of equal size were treated with fertilizer in amounts varying between 50 and 60 kg per plot and were mechanically kept at constant temperatures between 80 and 100^{o}F. Small carnation plants [approximately 15 centimeters (cm) in height] were planted in each plot, and their height y (cm) was measured after a six-week growing period. The resulting data are shown in Table 12.2 and are contained in the CARNATIONS data file.

a. Fit a complete second-order model to the data.
b. Do the data provide sufficient evidence to indicate that the second-order terms in the model β_3, β_4, and β_5 contribute information for the prediction of y?

Table 12.2

x_1	x_2	y	x_1	x_2	y	x_1	x_2	y
80	50	50.8	90	50	63.4	100	50	46.6
80	50	50.7	90	50	61.6	100	50	49.1
80	50	49.4	90	50	63.4	100	50	46.4
80	55	93.7	90	55	93.8	100	55	69.8
80	55	90.9	90	55	92.1	100	55	72.5
80	55	90.9	90	55	97.4	100	55	73.2
80	60	74.5	90	60	70.9	100	60	38.7
80	60	73	90	60	68.8	100	60	42.5
80	60	71.2	90	60	71.3	100	60	41.4

Solution:

In order to compare the determine if the second-order terms contribute information for predicting y, both a full model (containing the second-order terms) and a reduced model (that does not contain the second-order terms) must be fit in XLSTAT. We utilize the procedures covered in the last section to fit both models (using Data File **CARNATIONS**). The corresponding regression output for both models is shown in Figures 12.19 and 12.20.

Figure 12.19
Analysis of variance:
Full Model

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	5	8402.2645	1680.4529	596.3239	< 0.0001
Error	21	59.1784	2.8180		
Corrected Total	26	8461.4430			

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-5127.8991	110.2960	-46.4922	< 0.0001	-5357.2722	-4898.5260
TEMP	31.0964	1.3444	23.1301	< 0.0001	28.3005	33.8922
FERT	139.7472	3.1401	44.5047	< 0.0001	133.2171	146.2773
TEMPSQ	-0.1334	0.0069	-19.4636	< 0.0001	-0.1476	-0.1191
FERTSQ	-1.1442	0.0274	-41.7401	< 0.0001	-1.2012	-1.0872
TEM_FERT	-0.1455	0.0097	-15.0124	< 0.0001	-0.1657	-0.1253

Figure 12.20
Analysis of variance:
Reduced Model

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	1789.9344	894.9672	3.2195	0.0577
Error	24	6671.5085	277.9795		
Corrected Total	26	8461.4430			

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	106.0852	55.9450	1.8962	0.0700	-9.3796	221.5500
TEMP	-0.9161	0.3930	-2.3312	0.0285	-1.7272	-0.1050
FERT	0.7878	0.7860	1.0023	0.3262	-0.8344	2.4099

Compare these two printouts versus the SAS printouts found in the text. We refer you to the text for more detailed information regarding the interpretations and conclusion that should be made for these values.

12.4 Stepwise Regression

One of the tools that regression analysis provides for selecting variables to include in a regression model is the stepwise regression technique. As discussed in the text, the stepwise regression tool is a useful variable-screening tool when a large number of potential independent variables exist. The stepwise technique is illustrated with the following example from the text.

Exercise 12.5: We use Example 12.13 found in the *Statistics* text.

Problem: An international management consulting company develops multiple-regression models for executive salaries of its client firms. The consulting company has found that models which use the natural logarithm of salary as the dependent variable have better predictive power than those using salary as the dependent variable. A preliminary step in the construction of these models is the determination for the most important independent variable. For one firm, 10 potential independent variables (7 quantitative and 3 qualitative) were measured in a sample of 100 executives (saved in the EXECSAL data file). Since it would be very difficult to construct a complete second-order model with all of the 10 independent variables, use stepwise regression to decide which of the 10 variables should be included in the building for the final model for the logarithm of executive salaries.

Solution:

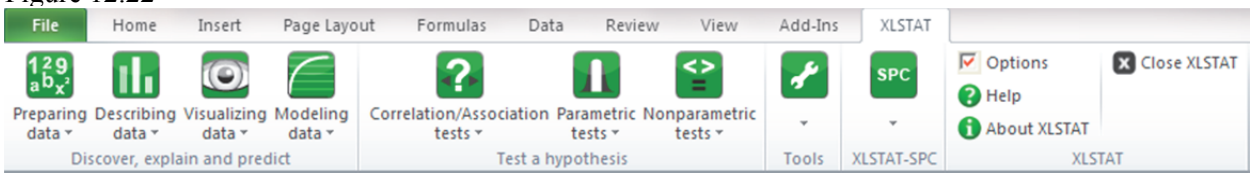
We solve Exercise 12.5 utilizing the **Linear regression** menu presented in **XLSTAT**. **Open** the data file **EXECSAL** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 12.21.

Figure 12.21

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
2	1	11.4436	12	15	1	240	170	1	44	5	0	21
3	2	11.7753	25	14	1	510	160	1	53	9	0	28
4	3	11.3874	20	14	0	370	170	1	56	5	0	26
5	4	11.2172	3	19	1	170	170	1	26	9	0	24
6	5	11.6553	19	12	1	520	150	1	43	7	0	27
7	6	11.1619	14	13	0	420	160	1	53	9	0	27
8	7	11.6457	18	18	1	290	170	1	43	7	0	22
9	8	11.1927	2	17	1	200	180	1	31	10	0	26
10	9	11.5954	14	13	1	560	180	1	43	7	0	23
11	10	11.136	4	16	1	230	160	1	36	10	0	25
12	11	11.5327	8	18	1	540	150	1	39	8	1	21

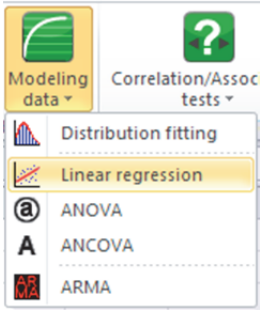
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 12.22.

Figure 12.22



To conduct a stepwise regression of the data, we click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 12.23.

Figure 12.23



This opens the **Linear regression** menu shown in Figures 12.24-12.25. 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 B thru L, rows 2 – 101, with row 1 being the variable labels. We note that the data in columns C - L represents the independent variables, x_1 through x_{10}, and the data in column B represent the dependent variable, y. We specify the column **B** data in the **Quantitative** box for the Dependent variable, y, and the column **C thru L** data in the **Quantitative** box for the independent, or Explanatory, variables, x_1 through x_{10}. We also check the **Variable labels** box in this menu.

Click on the **Options** tab (shown in Figure 12.25) to specify the type of stepwise regression model selection technique. To conduct a stepwise regression, we check the **Model Selection** box, and specify the **Stepwise** technique from the pull-down menu. Click **OK** to create the desired output.

Figure 12.24

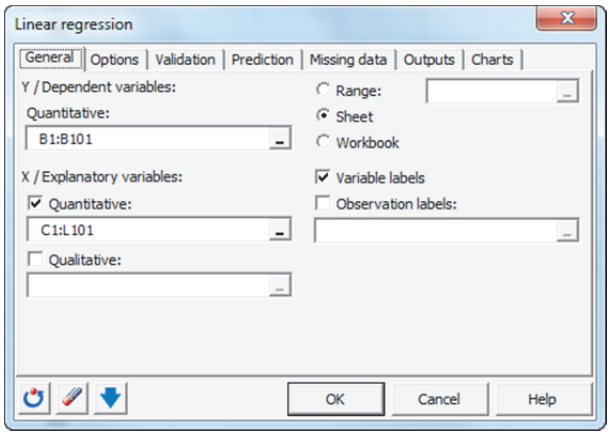
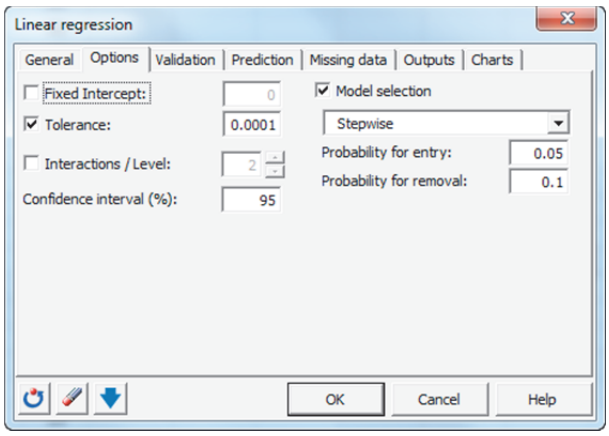


Figure 12.25



The XLSTAT output is shown in Figure 12.26.

Figure 12.26

Summary of the variables selection:

No. of variables	Variables	Variable IN/OUT	Status	MSE	R^2
1	X1	X1	IN	0.0260	0.6190
2	X1 / X3	X3	IN	0.0173	0.7492
3	X1 / X3 / X4	X4	IN	0.0112	0.8391
4	X1 / X2 / X3 / X4	X2	IN	0.0065	0.9075
5	X1 / X2 / X3 / X4 / X5	X5	IN	0.0056	0.9206

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	9.9619	0.1011	98.5777	< 0.0001	9.7613	10.1626
X1	0.0273	0.0010	26.5005	< 0.0001	0.0252	0.0293
X2	0.0291	0.0033	8.7188	< 0.0001	0.0225	0.0357
X3	0.2247	0.0164	13.7424	< 0.0001	0.1922	0.2572
X4	0.0005	0.0000	11.0643	< 0.0001	0.0004	0.0006
X5	0.0020	0.0005	3.9469	0.0002	0.0010	0.0029
X6	0.0000	0.0000				
X7	0.0000	0.0000				
X8	0.0000	0.0000				
X9	0.0000	0.0000				
X10	0.0000	0.0000				

We compare this printout to the one shown in the text. We note that the first five variables were selected as variables to include in the multiple regression analysis.

12.5 Residual Analysis

So far we have covered model building and model testing within the XLSTAT program. The last topic to address is the topic of residual analysis. As specified in the text, the topic of residual analysis requires the construction of several different graphical displays that are readily available from the multiple regression menu within XLSTAT. We illustrate how to generate these plots using the following example from the text.

Exercise 12.6: We use the data from the grandfather clocks data that is used throughout this chapter to generate the plots necessary to complete a residual analysis.

Solution:

The data for the grandfather clock example used throughout this chapter is shown in Table 12.3 and saved in the data file **GFCLOCKS**. The interaction model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

is again fit to these (modified) data. Use XLSTAT to generate all corresponding residual analysis printouts.

Table 12.3

Age(x_1)	Number of Bidders (x_2)	Auction Price (y)	Age(x_1)	Number of Bidders (x_2)	Auction Price (y)
127	13	1,235	170	14	2,131
125	12	1,080	182	8	1,550
127	7	845	162	12	1,884
150	9	1,522	184	10	2,041
156	6	1,047	143	6	v845
182	12	1,979	159	9	1,483
156	12	1,822	108	14	1,055
132	10	1,253	175	8	1,545
137	9	1,297	108	6	729
123	9	946	179	9	1,792
137	15	1,713	121	15	1,175
127	12	1,024	187	8	1,593
137	8	1,147	121	7	785
153	6	1,092	125	7	744
127	13	1,152	194	5	1,356
126	10	1,336	168	7	1,262

We begin by building the interaction model as we did earlier in this chapter. The results from building the interaction model are shown again in Figure 12.27.

Figure 12.27

Analysis of variance:

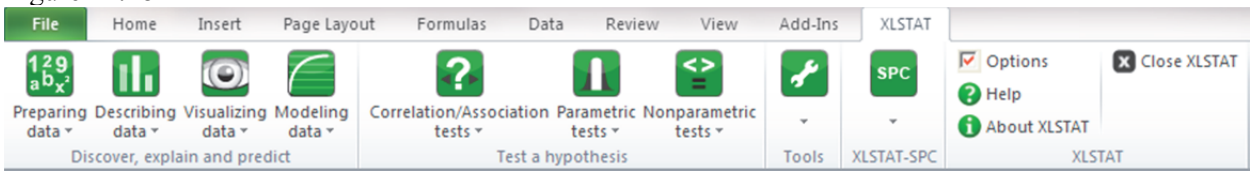
Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	3	4578427.3668	1526142.4556	193.0411	< 0.0001
Error	28	221362.1332	7905.7905		
Corrected Total	31	4799789.5000			

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	320.4580	295.1413	1.0858	0.2868	-284.1115	925.0275
AGE	0.8781	2.0322	0.4321	0.6690	-3.2845	5.0408
NUMBIDS	-93.2648	29.8916	-3.1201	0.0042	-154.4950	-32.0346
AGE-BID	1.2978	0.2123	6.1123	< 0.0001	0.8629	1.7328

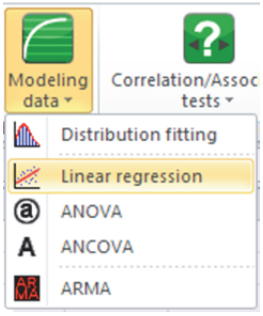
To generate the residual analysis printouts, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 12.28.

Figure 12.28



We click on the **Modeling data** menu and select the **Linear regression** option shown in Figure 12.29.

Figure 12.29



This opens the **Linear regression** menu shown in Figures 12.30-12.31. 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 thru D, rows 2 – 33, with row 1 being the variable labels. We note that the data in columns B - D represents the independent variables, x_1, x_2, and x_1x_2, and the data in column A represent the dependent variable, y. We specify the column **A** data in the **Quantitative** box for the Dependent variable, y, and the column **B thru D** data in the **Quantitative** box for the independent, or Explanatory, variables, x_1, x_2, and x_1x_2. We also check the **Variable labels** box in this menu.

Click on the **Chars** tab (shown in Figure 12.31) to specify the type of output desired. To generate the residual plots, we check the **Predictions and residuals** box (we need to make sure that the **Predictions and residuals** box is also checked in the Output tab). Click **OK**.

Figure 12.30

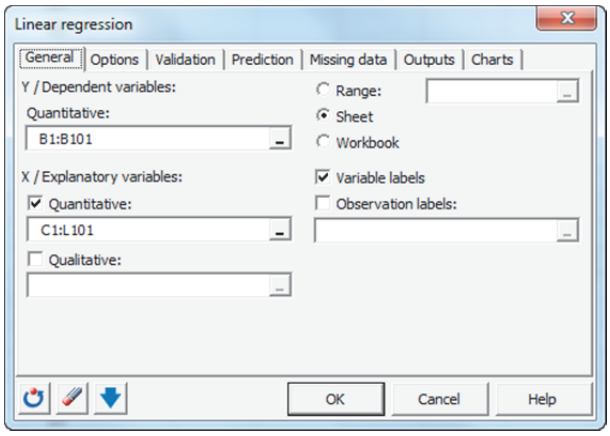
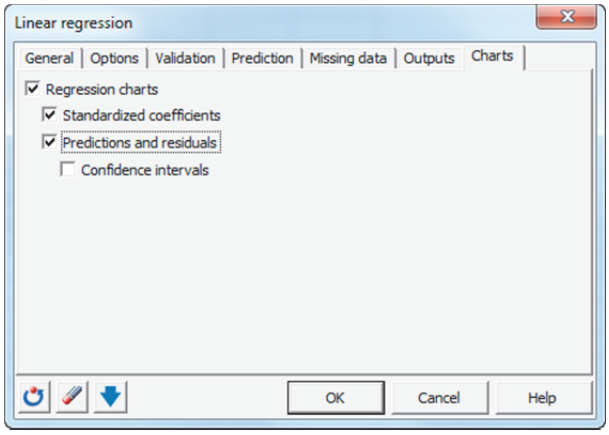
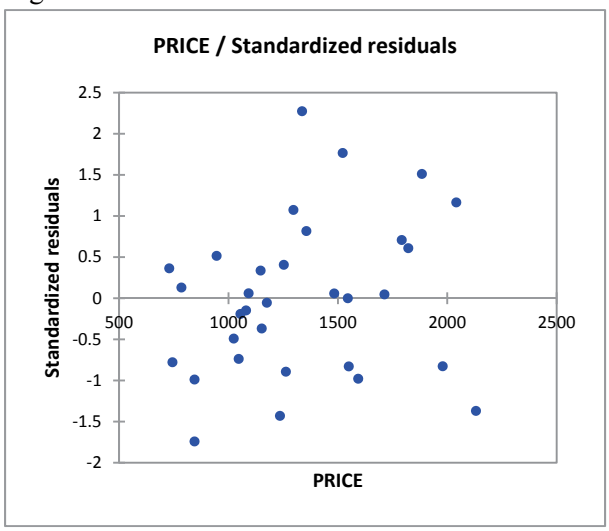


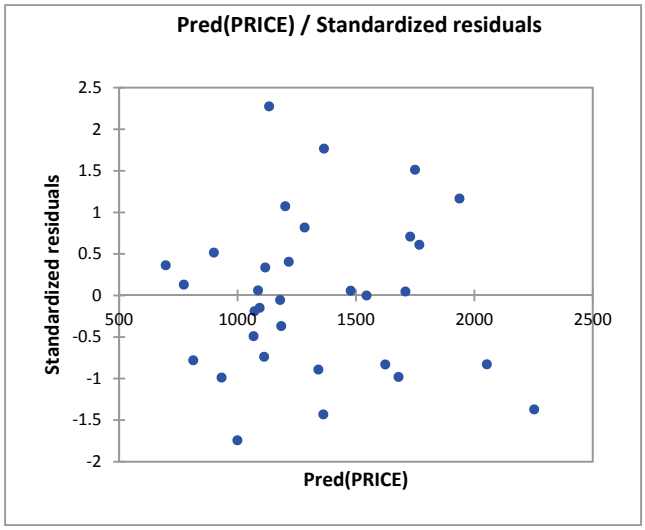
Figure 12.31

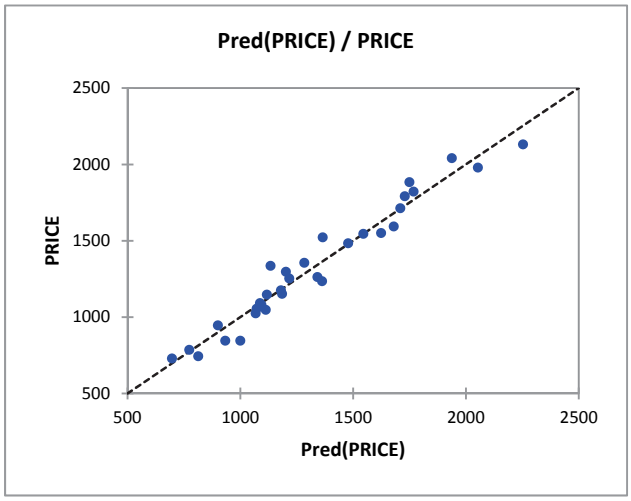


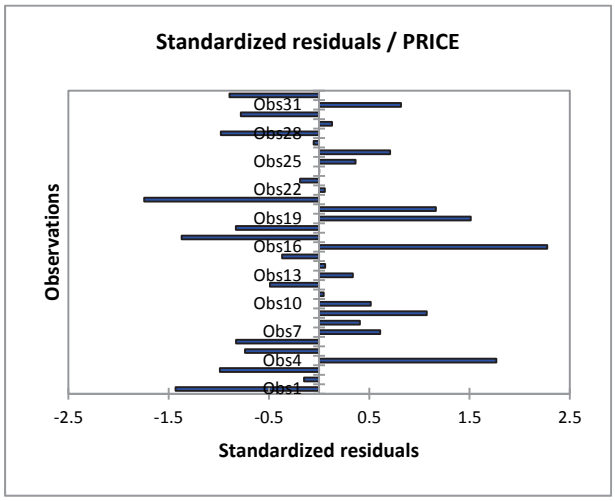
The XLSTAT residual plots are shown in Figure 12.32.We use these plots to conduct the residual analysis discussed in the text.

Figure 12.32









12.6 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 **Failure times of silicon wafer microchips**. Researchers at National Semiconductor experimented with tin-lead solder bumps used to manufacture silicon wafer integrated circuit chips (International Wafer Level Packaging Conference, No. 3-4, 3005). The failure time of the microchips (in hours) was determined at different solder temperatures (degrees Centigrade). The data for one experiment are given in the table and saved in the WAFER data set. Use the wafer data and the procedures you learned in this chapter to build a quadratic regression model that predicts failure time (y) based on solder temperature (x). Be sure to conduct a residual analysis also.

XLSTAT Output

Goodness of fit statistics:

Observations	22.0000
Sum of weights	22.0000
DF	19.0000
R^2	0.9415
Adjusted R^2	0.9354
MSE	473531.9339
RMSE	688.1366
DW	1.3566

Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	144830279.6194	72415139.8097	152.9256	< 0.0001
Error	19	8997106.7442	473531.9339		
Corrected Total	21	153827386.3636			

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	154242.9143	21868.4738	7.0532	< 0.0001	108471.6726	200014.1561
TEMP	-1908.8504	303.6636	-6.2861	< 0.0001	-2544.4255	-1273.2753
TempxTemp	5.9289	1.0476	5.6593	< 0.0001	3.7362	8.1217