

Chapter 2

Looking at Data–Relationships

In many studies, both independent and dependent variables typically arise either as part of a controlled experiment or as an observational study. Before any specific model is imposed that can be tested statistically, it is important to judge graphically whether any relationship is justified. If we let x denote the explanatory (independent) variable and y the response (dependent) variable, then we might plot in Cartesian coordinates all pairs (x_i, y_i) of observed values. This is called a **Scatterplot**.

2.1 Scatterplots

The steps involved in creating a scatterplot are similar to those for producing a **Histogram** using the **ChartWizard**. The instructions are based on **Excel 97/98/2000/2001**. **Excel 5/95** users should make corresponding changes.

Table 2.1: Fuel Consumption and Speed

Speed	10	20	30	40	50	60	70	80
Fuel	21.00	13.00	10.00	8.00	7.00	5.90	6.30	6.95

Speed	90	100	110	120	130	140	150
Fuel	7.57	8.27	9.03	9.87	10.79	11.77	12.83

Example 2.1. (Exercise 2.10, page 122 in the text.) How does the fuel consumption of a car change as its speed increases? Table 2.1 gives data for a British Ford Escort. Speed is measured in kilometers per hour, and fuel consumption is measured in liters of gasoline used per 100 kilometers travelled. Make a scatterplot with speed on the horizontal axis.

Creating a Scatterplot

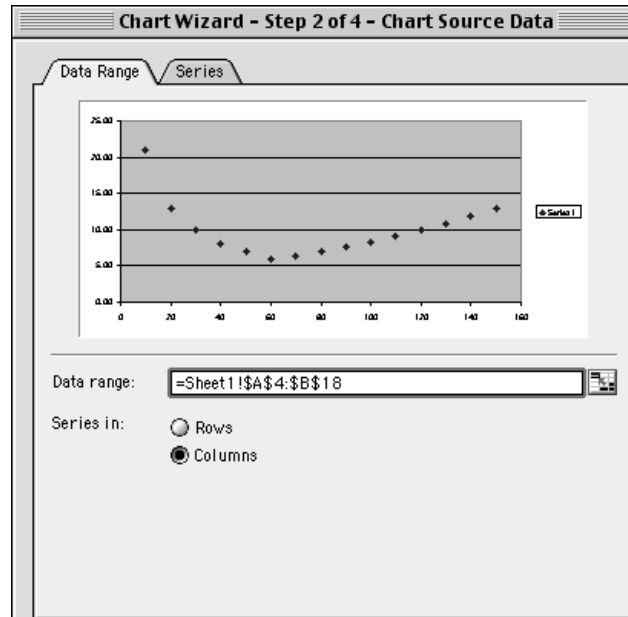


Figure 2.1: ChartWizard—Step 2

Enter the data from Table 2.1 into consecutive columns of a worksheet (for instance in cells A4:B18 and as shown in Figure 2.3 later in this section).

Step 1. Select cells A4:B22 and click on the **Chart Wizard**. From the choice of charts select **XY (Scatter)** for Chart type on the left and select the top Chart sub-type **Scatter** on the right. Click Next.

Step 2. The next dialog box previews the chart and allows any changes to be made to the data range (see Figure 2.1). Click Next.

Step 3. The **Chart Option** dialog box appears (Figure 2.2).

- Click the **Titles** tab and enter “Fuel vs. Speed” for Chart title, “Speed (km/hr)” for Value (X) Axis, and “Fuel used (liters)” for Value (Y) Axis.
- Click the **Legend** tab. Clear the Show Legend check box. Click Next.
- Click the **Gridlines** tab and make sure that the Major gridlines box is checked for both axes. Click Next.

Step 4. In the last step select the radio button to enter the chart as an object on the current sheet. Click Finish. The scatterplot appears embedded on your workbook (Figure 2.3).

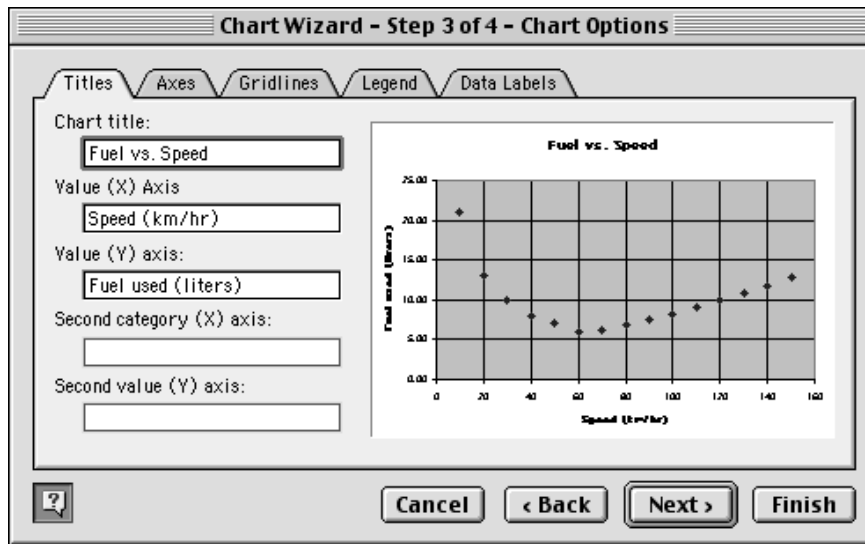


Figure 2.2: Preview of Scatterplot—Step 3

Enhancing a Scatterplot

The scatterplot may be enhanced using editing tools, some of which were described in Chapter 1. Activate the scatterplot by clicking once within its border to access new commands that become available under the Menu Bar. For instance, compare the pull-down options under **Insert**, **Format**, **Tools** as well as the new **Chart**.

Changing Scale

Excel uses a range from 0 to 100% as the default, and sometimes the scatterplot will show unwanted blank space. The scatterplot can be edited to change the maximum or minimum X axis value by double clicking the X axis to produce the **Format Axis** dialog box. Equivalently you can select the X axis by clicking once and then choose **Format** from the Menu Bar. Make any changes you wish by selecting the appropriate tabs at the top of the dialog box. If desired, the Y axis may similarly be selected for editing. Refer to the discussion for enhancing a histogram in Section 1.1, applies to any chart whether it is a histogram or a scatterplot.

Changing Titles

You can change titles on the scatterplot after you have completed it. Click on the X axis title “Speed (km/hr)” to select it and begin typing. Similarly the Y axis title and the chart title can be changed.

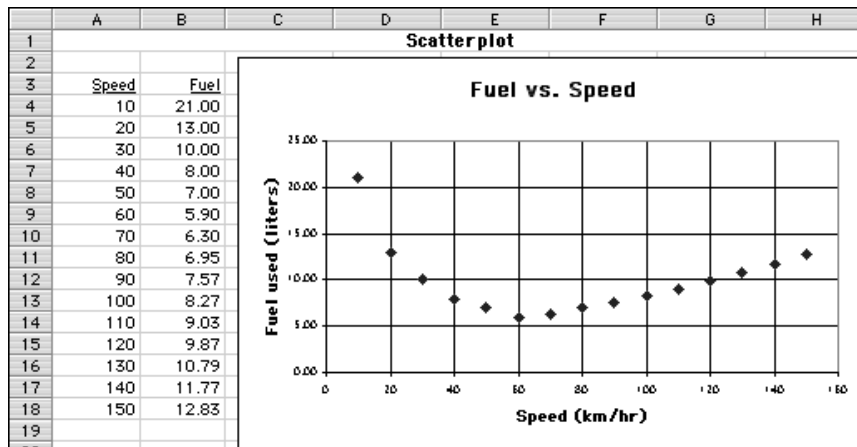


Figure 2.3: Speed and Fuel Data and Scatterplot Embedded on Sheet

Labeling a Data Point

By default Excel uses diamonds to plot the points. Suppose, for presentation purposes, you wish to use a different shape (and color) to represent a point and also to label a point. In particular suppose you wish to label the point representing minimum speed as “Min”. The following steps describe how to achieve this.

1. Activate the chart and click on the minimum observation.
2. Hold down the **Control** key (**Windows**) or **Command** key (**Macintosh**) and with your mouse pointer **select** the minimum point. Release the mouse button and select the point again. The pointer becomes a four-pointed plus sign (Figure 2.4). You can now access new commands under **Format** on the Menu Bar which let you edit the selected point.

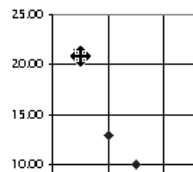


Figure 2.4: Selecting a Point

3. For **Excel 5/95** choose **Insert – Data Labels** from the Menu Bar to open the **Format Data Point** dialog box. For **Excel 97/98/2000/2001** choose **Format – Selected Data Point...** from the Menu Bar to open a corresponding **Format Data Point** dialog box. Under the **Data Labels**

tab select the radio button for Show Value. Click OK. Excel attaches the y value 21.00 to this point on the scatterplot and encloses it within a bordered selection box ready for editing. Type “min” (which appears in the **Formula Bar**) and press enter. The selection box now contains the word “min”. You can select and move it and then **deselect** by clicking elsewhere.

Changing the Marker and Color of a Data Point

Sometimes you may want to make a point stand out by changing its symbol and color on the scatterplot. We show how this is done using the “min” point above.

1. Activate the chart and click on the “min” observation as before.
2. Hold down the **Control** key (**Windows**) or **Command** key (**Macintosh**) and select the point so that the pointer again becomes a four-pointed plus sign.

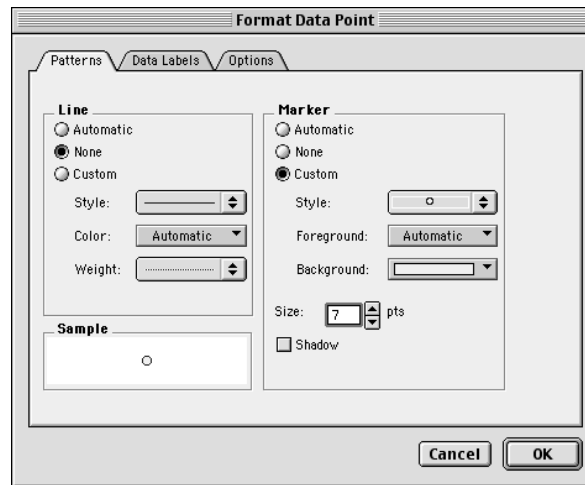


Figure 2.5: Changing the Default Marker

3. For **Excel 5/95**, choose **Insert – Data Labels** from the Menu Bar to open the **Format Data Point** dialog box. For **Excel 97/98/2000/2001**, choose **Format – Selected Data Point...** from the Menu Bar to open a corresponding **Format Data Point** dialog box. Under the **Patterns** tab, leave the **Line** selection as **None**. Under **Marker**, select a marker type from the pull-down list for **Style**, and also select a Foreground and Background color and size (Figure 2.5). In **Excel 5/95** the size of the marker cannot be changed and there is no **Options** tab. Your selection is previewed in the small **Sample** box in the lower portion of the dialog box. Click OK. Figure 2.4 shows the result of the above editing.

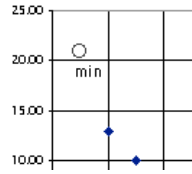


Figure 2.6: Editing a Point

Normal Quantile Plots

There are several ways to assess whether a data set is normal. An analytic approach beyond the level of this book was developed by S. Shapiro and M. B. Wilk (An analysis of variance test for normality, *Biometrika* **52**, pp. 591–611, 1965).

	A	B	C	D	E	F
1	Newcomb's Measurements					
2	of the Speed of Light					
3						
4	28	26	33	24	34	-44
5	27	16	40	-2	29	22
6	24	21	25	30	23	29
7	31	19	24	20	36	32
8	36	28	25	21	28	29
9	37	25	28	26	30	32
10	36	26	30	22	36	23
11	27	27	28	27	31	27
12	26	33	26	32	32	24
13	39	28	24	25	32	25
14	29	27	28	29	16	23

Figure 2.7: Newcomb Data

A simple graphical approach is to construct a histogram and compare the observed counts with the 68-95-99.7% rule. A more sensitive version of this idea is to order the observations and examine their distribution visually, using a scatterplot involving the corresponding expected quantiles of a normal curve. Normal data will tend to fall on a straight line. (This is the basis for the Shapiro-Wilk test.)

Although Excel does not provide a normal quantile plot, one can easily be constructed. The expected value of the i th order statistic (the i th largest in increasing magnitude) of a sample of size n from a $N(0, 1)$ distribution can be approximated by the percentile

$$z_{(i)} = \text{NORMSINV} \left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}} \right)$$

which is the value of a standard normal such that the area to the left is $\frac{i - \frac{3}{8}}{n + \frac{1}{4}}$.

Then plot $z_{(i)}$ on the vertical axis against $x_{(i)}$ on the horizontal axis where $x_{(i)}$ is the i th largest from the data set $\{x_1, x_2, \dots, x_n\}$ using the **Chart Wizard**.

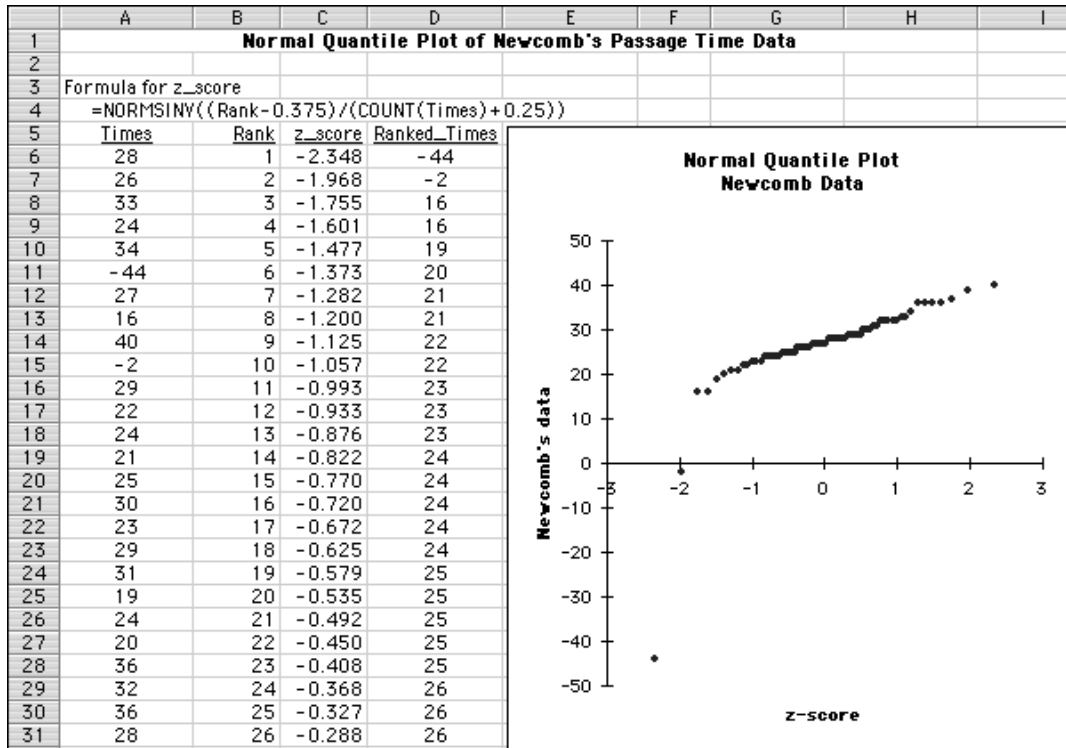


Figure 2.8: Normal Quantile Plot

Example 2.2. (Example 1.29, page 79 in the text.) Figure 2.7 contains 66 measurements of the speed of light made by Simon Newcomb between July and September 1882. The measurements give the deviation from 24,800 nanoseconds baseline. Construct a normal quantile plot of the data.

Solution

1. Referring to columns A:D in Figure 2.8, reenter the data in cells A6:A71 of a workbook and the label “Times” in A5. From the Menu Bar, choose **Data – Sort** to sort the data in increasing order and enter the sorted data in D6:D71.
2. Enter the label “Rank” in B5 followed by the integers $\{1, 2, \dots, 66\}$ in B6:B71.
3. Enter the label “z_score” in C5. Name the ranges “Times” and “Rank.” Then select cell C6 and enter

$$= \text{NORMSINV}((\text{Rank} - 0.375)/(\text{COUNT}(\text{Times}) + 0.25))$$

Click the fill handle at the lower right corner of C6 and fill to C71.

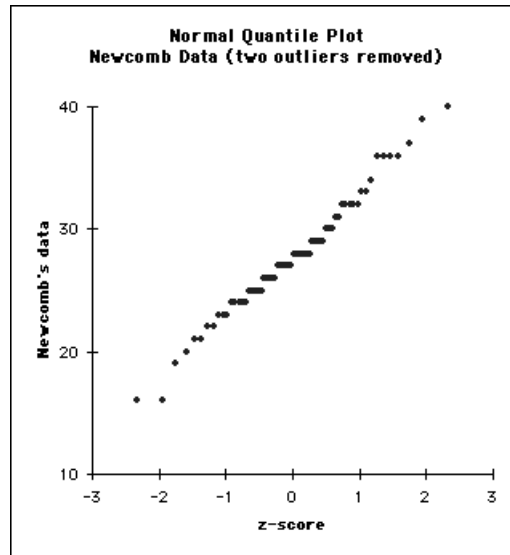


Figure 2.9: Normal Quantile Plot—Outliers Removed

4. Select a cell to locate the quantile plot, then click the **Chart Wizard** button. Choose scatterplot as the chart, enter the data range C5:D71, and complete the remaining steps. Enhance the chart to reproduce the scatterplot shown in Figure 2.8 which also shows a portion of the data on the worksheet.

Two outliers are apparent $\{-44, -2\}$. Remove them from the plot and redo the calculations. The result is Figure 2.9, whose straight line appearance is an indication that a normal distribution fits the data quite well after the two outliers have been removed.

2.2 Correlation

The correlation between two variables x and y measures the strength of the linear association between them. For n pairs (x_i, y_i) , $1 \leq i \leq n$, of data points the sample correlation coefficient is defined to be

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where \bar{x} and \bar{y} are the sample means of the $\{x_i\}$ and $\{y_i\}$, respectively, and s_x and s_y are the corresponding sample standard deviations.

Using the CORREL Function

The most direct way to find the correlation for Example 2.1 is by the **Formula Palette** for **Excel 97/98/2000/2001** or the **Function Wizard** for **Excel 5/95** with the function **CORREL**, which computes the correlation coefficient.

Example 2.3. (Exercise 2.21, page 131 in the text.) Find the correlation r between the heights of the women and men in Table 2.2

Table 2.2: Heights of Women and Men

Women (x)	66	64	66	65	70	65
Men (y)	72	68	70	68	71	65

Solution

1. Enter the above data into two adjacent columns in a worksheet, say, cells A1:B6. Select an empty cell where you want the correlation to appear and invoke either the **Function Wizard** or the **Formula Palette**. In each case, select **Statistical** for Function Category and **CORREL** for Function Name.
2. Enter A1:A6 for **Array1** and B1:B6 for **Array2**. You may enter by hand or click and drag on the workbook over the range A1:A6, press Tab on the keyboard, then click and drag over the range B1:B6, and finally click OK (or Finish, for **Excel 5/95**). The answer 0.565 appears in the cell you selected.

Using the ToolPak

Correlation between two variables can also be calculated using the **Correlation** tool in the **Analysis ToolPak**. This tool is most effective, however, for determining pairwise correlations for multivariate data sets for which repeated use of the **CORREL** function would be inefficient.

This tool prints out a matrix of correlations. Such a matrix is helpful in multiple regression in deciding which variables to include in a model.

Example 2.4. Darlene Gordon of the Purdue University School of Education provided the data partly shown in Figure 2.10. The data were collected on 78 seventh-grade students in a rural Midwestern school. The research concerned the relationship between the students' "self-concept" and their academic performance. The variables are OBS, a subject identification number; GPA, grade point index; IQ, score on an IQ test; AGE, age in years; SEX, female (F) or male (M); SC, overall score on the Piers-Harris self-concept scale; and C1–C6, "cluster scores" for specific aspects of self-concept: C1 = behavior, C2 =

	A	B	C	D	E	F	G	H	I	J	K	L
1	Self-concept and Academic Performance											
2	OBS	GPA	IQ	AGE	SEX	SC	C1	C2	C3	C4	C5	C6
3	1	7.9	111	13	2	67	15	17	13	13	11	9
4	2	8.3	107	12	2	43	12	12	7	7	6	6
5	3	4.6	100	13	2	52	11	10	5	8	9	7
6	4	7.5	107	12	2	66	14	15	11	11	9	9
7	5	8.9	114	12	1	58	14	15	10	12	11	6
8	6	7.6	115	12	2	51	14	11	7	8	6	9
9	7	7.7	111	13	2	71	15	17	12	14	11	10
10	8	2.4	97	13	2	51	10	12	5	11	5	6
11	9	6.0	100	13	1	49	12	9	6	9	6	7
12	10	8.8	112	13	2	51	15	16	4	9	5	8
13	11	7.5	104	12	1	35	12	5	3	2	4	7
14	12	5.5	89	13	1	54	16	8	3	11	7	7
15	13	7.2	104	13	2	54	16	14	6	7	2	7
16	14	7.6	102	13	1	64	14	12	8	10	12	9
17	15	4.7	91	14	1	56	14	13	8	10	7	8
18	16	8.2	114	13	1	69	15	15	9	12	11	9
19	17	7.8	114	13	1	55	14	11	6	11	11	9
20	18	7.6	103	12	1	65	16	16	5	12	11	9

Figure 2.10: Self-concept and Academic Performance Data Set

school status, C3 = physical, C4 = anxiety, C5 = popularity, and C6 = happiness.

Find the correlations between the response variable GPA and each of the explanatory variables IQ, AGE, SEX, SC, and C1–C6. Of all the explanatory variable, IQ does the best job of explaining GPA in a simple linear regression. How do you know this without doing all the regressions?

Solution. Figure 2.10 shows the first 30 sets of observations in this data set, to which the following instructions apply. For the full set merely select the appropriate Input range.

1. From the Menu Bar choose **Tools – Data Analysis** and in the dialog box highlight **Correlation** and click OK.
2. In the next dialog box, **Correlation**, enter A2:L32 for **Input range** (most conveniently done by clicking and dragging over this range on the workbook and pressing the Tab key). Check the box **Labels in first row** and point to cell N1 for **Output range**. Click OK.

Excel Output

The output appears in N1:Z13, as shown in Figure 2.11. In view of symmetry, only half the correlation matrix is required. From Column P we read off the correlations between GPA and the explanatory variables. The largest correlation involving GPA is with IQ and is 0.709.

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1		OBS	GPA	IQ	AGE	SEX	SC	C1	C2	C3	C4	C5	C6
2	OBS	1											
3	GPA	0.318	1										
4	IQ	0.520	0.709	1									
5	AGE	0.110	-0.164	-0.299	1								
6	SEX	-0.087	-0.051	0.211	-0.103	1							
7	SC	0.255	0.654	0.415	0.094	-0.098	1						
8	C1	-0.116	0.562	0.099	-0.080	-0.248	0.685	1					
9	C2	0.109	0.679	0.450	0.050	0.020	0.874	0.633	1				
10	C3	0.292	0.601	0.578	0.068	0.027	0.790	0.316	0.733	1			
11	C4	0.192	0.484	0.318	0.150	-0.021	0.857	0.474	0.787	0.662	1		
12	C5	0.230	0.590	0.399	0.071	-0.320	0.828	0.480	0.667	0.717	0.760	1	
13	C6	0.325	0.644	0.396	0.209	-0.124	0.812	0.517	0.685	0.676	0.572	0.681	1

Figure 2.11: Correlation ToolPak Output

2.3 Least-Squares Regression

We have seen how to plot two variables against each other in a scatterplot and have calculated the correlation coefficient to measure the strength of the linear association between them. It is useful to have an analytic relationship between the explanatory variable x and the response variable y of the form

$$y = f(x)$$

for predicting y from x . Such a relationship is called a simple (meaning one explanatory variable) **regression curve**. The simplest curve is a straight line

$$y = a + bx$$

called the regression line of y on x . The regression line represents, under certain assumptions, the mean response at each specified value x .

The method used to determine the coefficients a and b goes back at least to the great mathematician Gauss and is called the **Principle of Least-Squares**. Gauss himself recognized that the criterion was arbitrary and he used it because the coefficients a and b were then solvable in closed form. (Additional reasons connected with the errors being normal are presented in more advanced treatments.)

For a given x_i , we call

$$\hat{y}_i = a + bx_i$$

the **predicted** value and

$$e_i = y_i - \hat{y}_i$$

the **residual**. The **error sum of squares** is defined to be

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

By differentiating with respect to a and b , we can solve for the values that minimize $\sum_{i=1}^n e_i^2$. These are the values used in the regression line. They are given by the formulas

$$\begin{aligned} \text{slope} \quad b &= r \frac{s_y}{s_x} \\ \text{intercept} \quad a &= \bar{y} - b\bar{x} \end{aligned}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, r is the correlation coefficient, and

$$\begin{aligned} (n-1)s_x^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \\ (n-1)s_y^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \end{aligned}$$

Fitting a Line to Data

Excel provides three built-in methods for regression analysis: **Trendline**, the **Regression** tool in the **Analysis ToolPak**, and regression functions such as **FORECAST** and **TREND**. For merely graphing a regression line and providing its equation and the coefficient of determination r^2 , the **Trendline** command suffices. We will consider the **Regression** tool in Chapters 10 and 11, as well as regression functions.

Linear Trendline

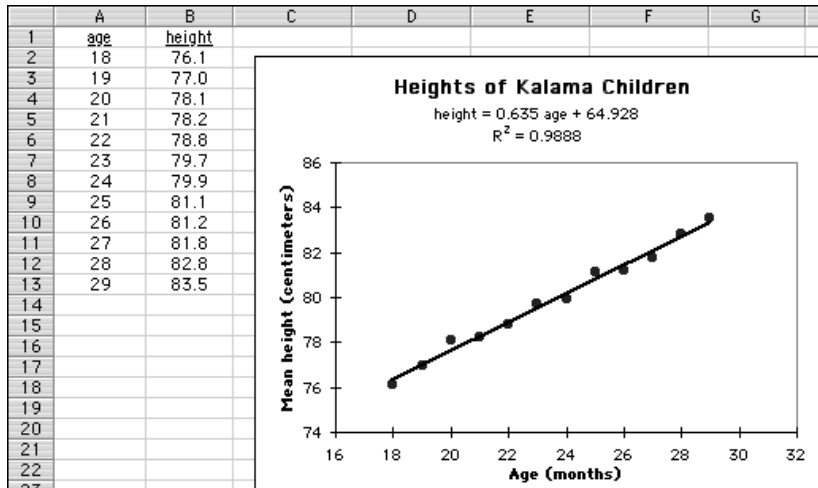


Figure 2.12: Kalama Data and Scatterplot

We use the **Linear Trendline** to insert a curve on a scatterplot. The trendline can be added to any scatterplot even after the **Regression** tool is used.

Example 2.5. (Example 2.10, page 135 in the text.) Columns A, B of Figure 2.12 give the data for a group of children in Kalama, an Egyptian village that was the site of a study of nutrition in developing countries. The explanatory variable x is age, the response variable y is height. Fit the least-squares regression line to the data.

Solution. We first construct a scatterplot of the data (also shown in Figure 2.12) to verify that a linear model is appropriate.

1. Enter the data in cells A2:B13 of a workbook. Use the **ChartWizard** to create a scatterplot and edit it (primarily to change the horizontal scale), as discussed previously, so that it appears as shown in Figure 2.12. The scatterplot shows an approximate linear relationship, so it is appropriate to fit the data pairs with a straight line.

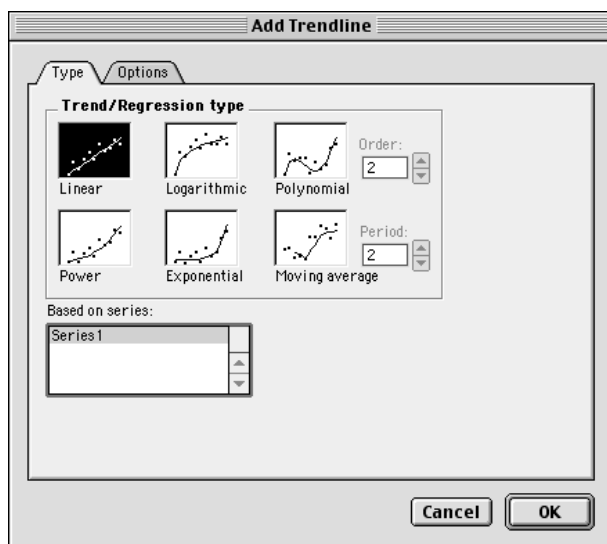


Figure 2.13: Trendline Type

2. Activate the chart for editing and select the data by **clicking on one of the points**. The points appear highlighted, the **Name** box in the **Formula Bar** shows S1, and in the text entry area we can read

$$= \text{SERIES}(\text{,Sheet1!\$A\$2 : \$A\$13, Sheet1!\$B\$2 : \$B\$13, 1})$$

meaning that the series has been selected. (Refer to the online help for more information on this function and the Introduction for the meaning of Sheet1! notation.)

3. For **Excel 5/95**, choose **Insert – Trendline** from the Menu Bar; for **Excel 97/98/2000/2001**, choose **Chart – Add Trendline...** from the Menu Bar. Then proceed as follows. Click the **Type** tab and select **Linear** (Figure 2.13). **Excel 97/98/2000/2001** have an additional text area (**Based on series**) in the blank space at the bottom of this figure. Click the **Options** tab and select the radio button **Automatic:Linear (Series1)**. Check the boxes **Display Equation on Chart** and **Display r-squared Value on Chart**. Make sure that the **Set Intercept** box is clear. Click **OK**. The regression line is superimposed on the scatterplot, its equation $y = 0.635x + 64.928$ is displayed, and the coefficient of determination $R^2 = 0.9888$ (in Excel's notation) is inserted on the scatterplot.
4. The output may be edited as previously with other charts for presentation purposes. For instance, activate the chart, and click on the rectangular box surrounding the equation; the border turns a darker grey. Use the **Decimal** tool to increase or decrease the number of decimal points. Edit the text by replacing x with “age” and y with “height.” Move $R^2 = 0.9888$ from its location on the graph to a more convenient place. The final result with the regression line appears as Figure 2.14.

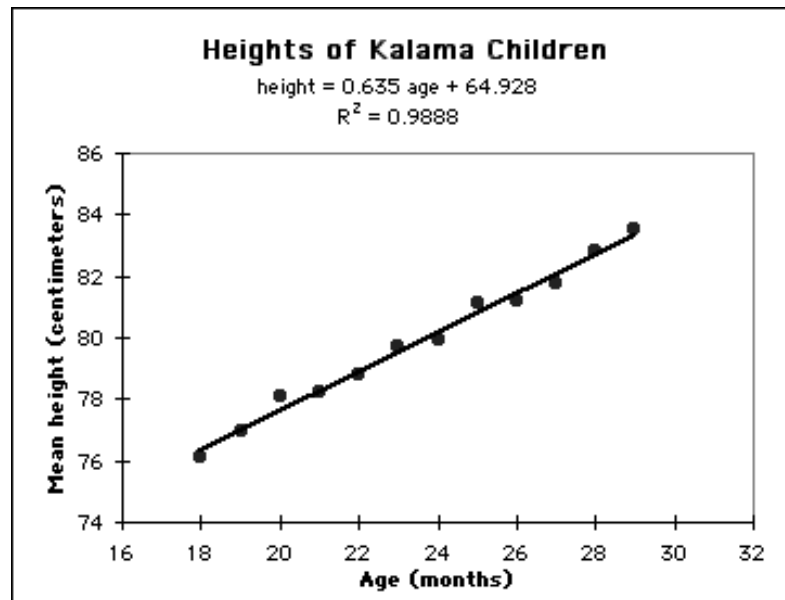


Figure 2.14: Regression Line and Scatterplot

Residuals

No discussion of regression is complete without an analysis of residuals, which provide evidence of how well the regression model fits. We defer discussion of this topic to Chapter 10 where the **Regression** tool will be introduced. However, in Figure 2.15 we show a scatterplot of residuals against age for the Kalama example. There does not appear to be any discernable pattern in the plot, indicating that a straight-line fit is appropriate.



Figure 2.15: Residual Plot