

Chapter 2

Looking at Data—Relationships

This chapter studies relationships between variables. Following an approach similar to Chapter 1, relationships are displayed with graphs; the strength of a linear relationship is described by a number; and then straight-lines are used as models for relationships between two quantitative variables. We also learn how to turn complex relationships into linear ones. Discussion of relations between categorical variables is postponed until Chapter 9.

All graphs and statistical computations in this chapter are performed in the second platform **Fit Y by X** of the **Analyze** menu.

2.1 Displaying Relationships with Graphs

A *scatterplot* displays the relationship between two quantitative variables. *Side-by-side boxplots* and *side-by-side means diamonds* display the relationship between a categorical explanatory variable and a quantitative response variable. In JMP, you specify the role (response or categorical) and modeling type of each variable and JMP automatically performs the appropriate methodology.

2.1.1 Two Quantitative Variables: Scatterplots

Scatterplots are created whenever the **Fit Y by X** platform is called and both variables are quantitative.

IPS Figure 2.1 State SAT scores

Figures 2.1 and 2.2 of the textbook show scatterplots to investigate the relationship between a state's mean SAT mathematics score and the percent of its high school seniors who take the exam. The JMP data table **fg02_001.jmp** on the IPS CD-ROM contains education and related data for the states.

To create a scatterplot,

1. Select **File** ⇒ **Open** and the file **fg02_001.jmp** from the IPS CD-ROM.

fg02_001.jmp									
8/0 Cols	State	Region	Pop	SAT Verbal	SAT Math	Percent taking	Percent no HS	Teacher's pay	
51/0 Rows	1	AL	ESC	4273	565	558	8	33.1	31.3
	2	AK	PAC	607	521	513	47	13.4	49.6
	3	AZ	MTN	4428	525	521	28	21.3	32.5
	4	AR	WSC	2510	566	550	6	33.7	29.3
	5	CA	PAC	31878	495	511	45	23.8	43.1
	6	CO	MTN	3823	536	538	30	15.6	35.4
	7	CT	NE	3274	507	504	79	20.8	50.3
	8	DE	SA	725	508	495	66	22.5	40.5
	9	DC	SA	543	489	473	50	26.9	43.7
	10	FL	SA	14400	498	496	48	25.6	33.3

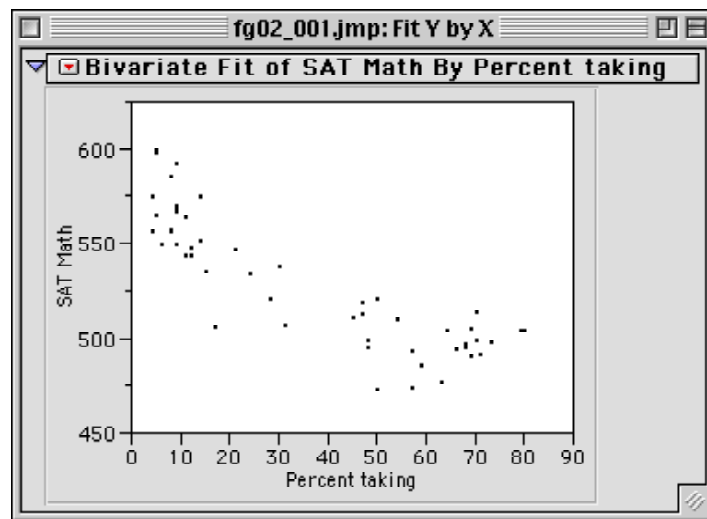
2. Select **Analyze** ⇒ **Fit Y by X**.

Since **SAT Math** is the *response variable* and **Percent taking** is the *explanatory variable*,

3. Select the column **SAT Math** and click **Y, Response**.

4. Select the column **Percent taking** and click **X, Factor**.

5. Press **OK**.



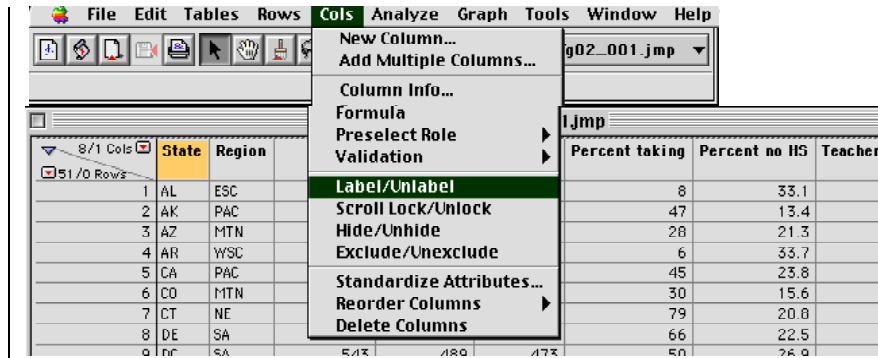
Identifying Individuals on the Scatterplot

If you hover the cursor over a data point, the row number of the state that the point represents is displayed. The **Label/Unlabel** command in the **Cols** menu tells JMP to use a column's values to identify points in plots.

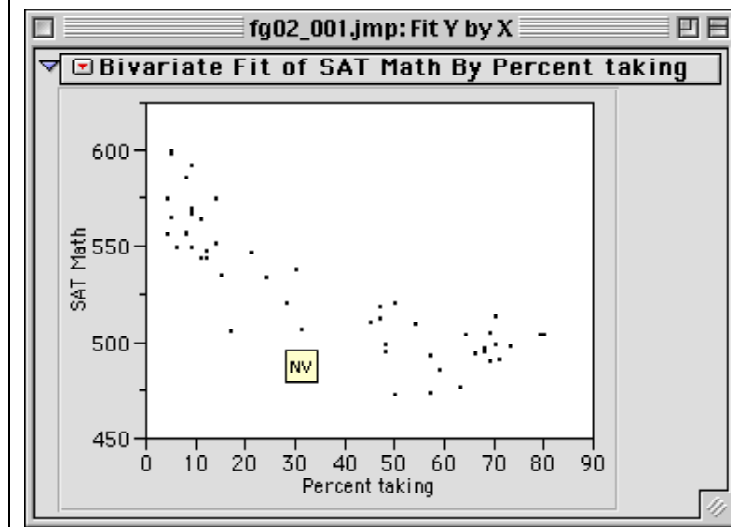
IPS Figure 2.1 State SAT scores (cont'd.)

1. Select the column **State** in the data table.

2. Select **Cols** ⇒ **Label/Unlabel**.




Return to the **fg02_001.jmp: Fit Y by X** window directly or by using the **Window** menu. Move the cursor over the data points. JMP now displays the value of the variable **State**.



Remarks

The scatterplot can be enhanced in several ways (see Section 0.6 in Chapter 0 for details):

- Increase (or decrease) the size of the plot by selecting a corner of the plot and dragging.
- Scroll either axis by moving the hand tool over the numbers.
- Modify tick marks and the increment between numbers by double-clicking on a scale data value.
- Modify or enhance an axis name by double-clicking on the axis name.
- Create editable notes to be displayed and stored with the plot using , the annotate tool.

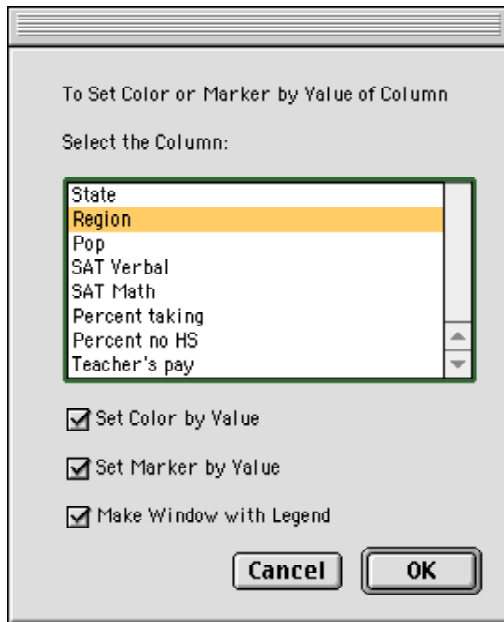
Adding Categorical Variables to Scatterplots

IPS Figure 2.1 State SAT scores (cont'd.)

The Census Bureau groups the states into regions of the country. To investigate regional patterns, we might wish to assign the points on the scatterplot associated with the individual states different colors and symbols depending on which region they are in. To do this, we change the *state* of the rows/individuals. (See Section 0.3.2 in Chapter 0 for more details on row states.) The *row state characteristic* that we use

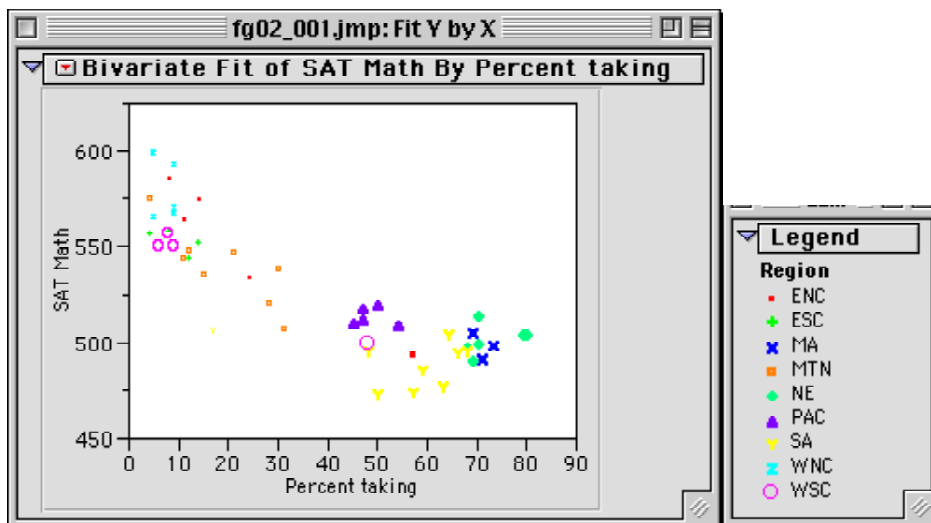
here is called **Color or Marker by Column**. We wish to color and mark the points differently depending on the values of a variable or column.

1. Select **Rows** ⇒ **Color or Marker by Column**.



2. Select **Region**, and check **Set Marker by Value** and **Make Window with Legend**.
3. Press **OK**.

Return to the **fg02_001: Fit Y by X** window and compare two regions.



2.1.2 A Categorical Explanatory Variable and a Quantitative Response Variable: Side-by-Side Boxplots and Means Diamonds

To display a relationship between a categorical explanatory variable and a quantitative response variable, we make a side-by-side comparison of the distributions of the response for each category. Some tools for such comparisons are

- Side-by-side boxplots. See IPS Figure 1.16 on the comparison of calories (the quantitative response) for beef, meat, and poultry hot dogs (the categories).
- Side-by-side point plots.
- Side-by-side means diamonds.

All three plots are available in JMP. We will use the hot dog data to illustrate.

IPS Table 1.9 Differences among types of hot dogs

People who are concerned about health may prefer low-calorie hot dogs and ask “Are there any systematic differences among the three types of hot dogs in calories?” In other words, “Are the type and calorie content of a hot dog related?”

The JMP data table **Hot Dogs** on the IPS CD-ROM contains data on 54 brands of hot dogs.

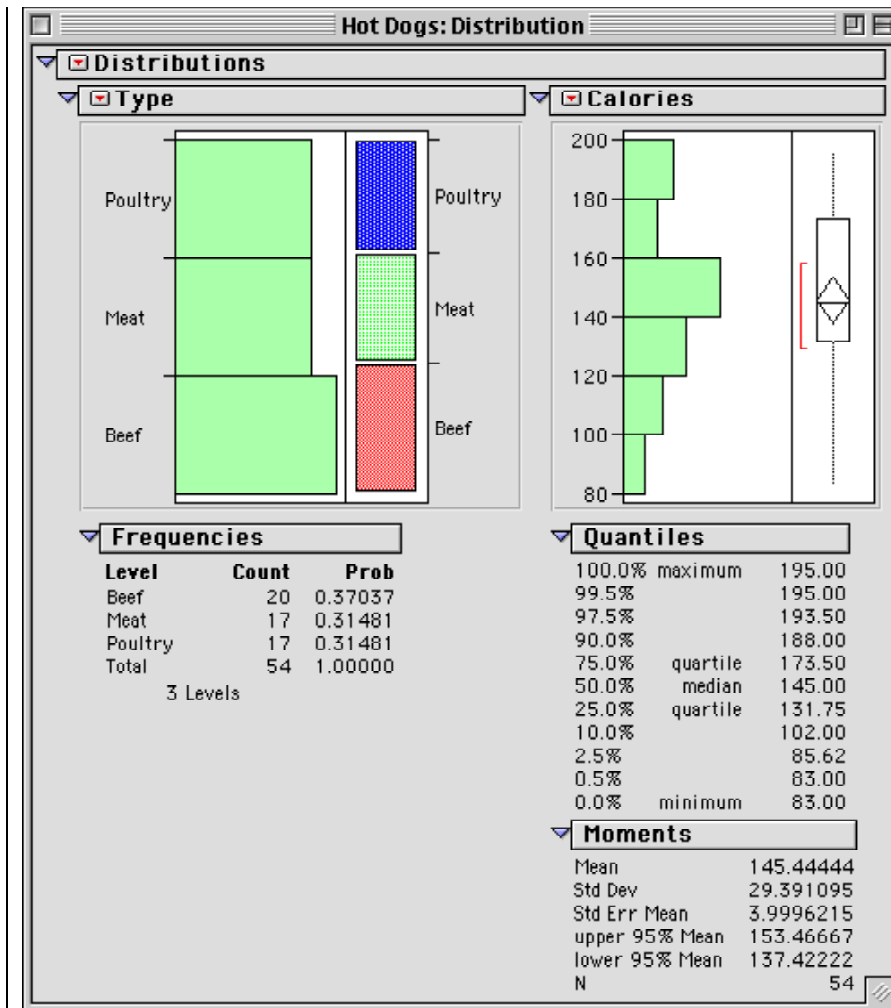
1. Select **File** ⇒ **Open** in the menu bar to open the data table.
2. Select the file **Hot Dogs.jmp** from the IPS CD-ROM.



	Product Name	Type	Taste	\$/oz	\$/lb Protein	Calories	Sodium	Protein /Fat
1	Happy Hill Supers	Beef	Bland	0.11	14.23	186	495	1
2	Georgies Skinless Beef	Beef	Bland	0.17	21.70	181	477	2
3	Special Market's Premium Be	Beef	Bland	0.11	14.49	176	425	1
4	Spike's Beef	Beef	Medium	0.15	20.49	149	322	1
5	Hungry Hugh's Jumbo Beef	Beef	Medium	0.10	14.47	184	482	1
6	Great Dinner Beef	Beef	Medium	0.11	15.45	190	587	1
7	PJB Kosher Beef	Beef	Medium	0.21	25.25	158	370	2

There are 54 brands of hot dogs and 8 variables describing each brand. First, examine the distribution of each of the variables **Type** and **Calories**.

3. Select **Analyze** ⇒ **Distribution**.
4. Select the columns **Type** and **Calories** and press **Y, Columns** and **OK**.



Remark

Notice that the calories for all three types of hot dogs are listed in one column of the JMP data table and not three, even though Table 1.9 of IPS uses three columns, one for each type of hot dog. This is because each row of a JMP data table represents an individual; in this case, it is a brand of hot dog. Since each column in a JMP data table represents a variable, the three columns of Table 1.9 in IPS are stacked into two variables—**Type** and **Calories**. This is very important to remember since all statistical computations and graphs assume that the individuals are the rows of a data table and the variables are columns.

To make a side-by-side comparison of the distributions of the calories for each type of hot dog, use the **Fit Y by X** platform.

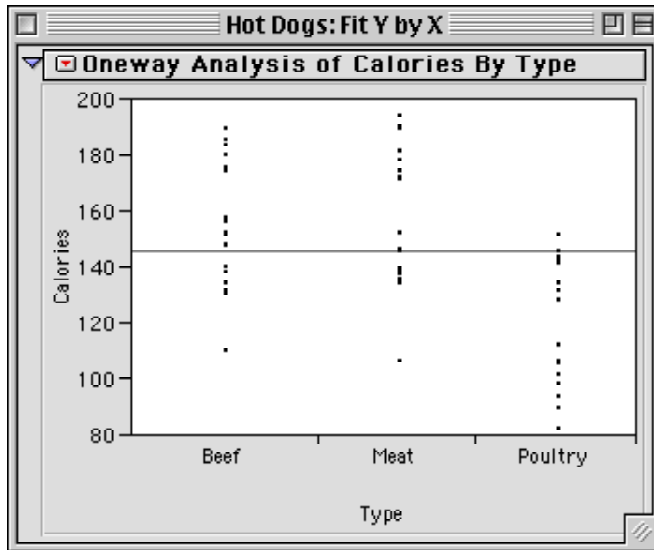
5. Select **Analyze** ⇒ **Fit Y by X**.

Since **Calories** is the *response variable* and **Type** is the *explanatory variable*,

6. Select the column **Calories** and click **Y, Response**.

7. Select the column **Type** and click **X, Factor**.

8. Press **OK**.

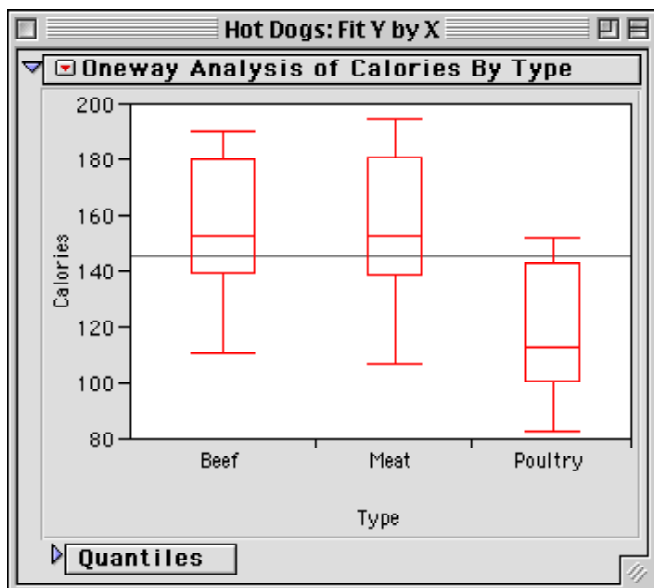


JMP presents *side-by-side point plots*. The horizontal line at 145.4 calories is the overall mean calorie content. To get *side-by-side boxplots*,

9. Press the red triangle in the **Oneway Analysis of Calories By Type** report.

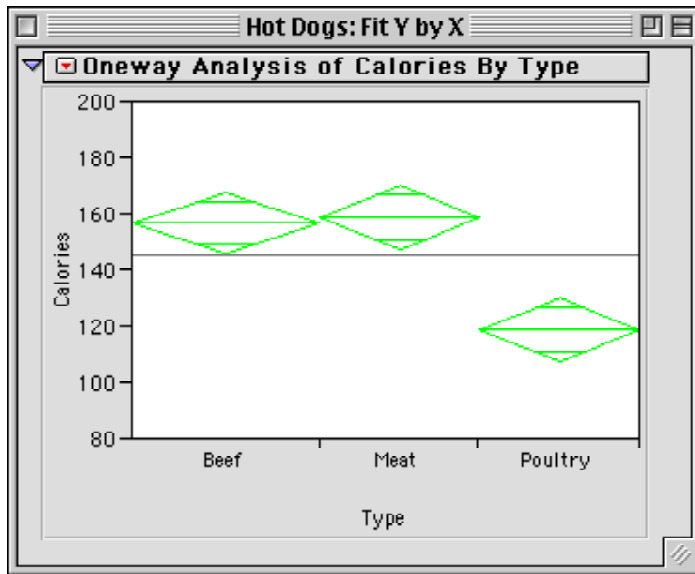
10. Select **Display Options** ⇒ **BoxPlots** from the menu that opens.

11. Deselect **Display Options** ⇒ **Points** to remove the points.




Side-by-side means diamonds are another tool, offered by JMP, for comparing the distributions of a response variable among the categories of another variable.

12. Deselect **Display Options** ⇒ **BoxPlot**
13. Select **Display Options** ⇒ **Mean Diamonds**.



In both displays, it can be seen that Poultry hot dogs have lower calories on average than either Beef or Meat hot dogs. Notice also that there appears to be no difference between Beef and Meat hot dogs.

Remark

The remainder of this chapter concentrates on relationships among quantitative variables. We will use *side-by-side means diamonds* again in conjunction with the analyses discussed in Chapters 7, 12, and 13 of the textbook. To learn more about *means diamonds*, select the  tool and click on a diamond.

2.2 Describing Relationships with Numbers: Correlation

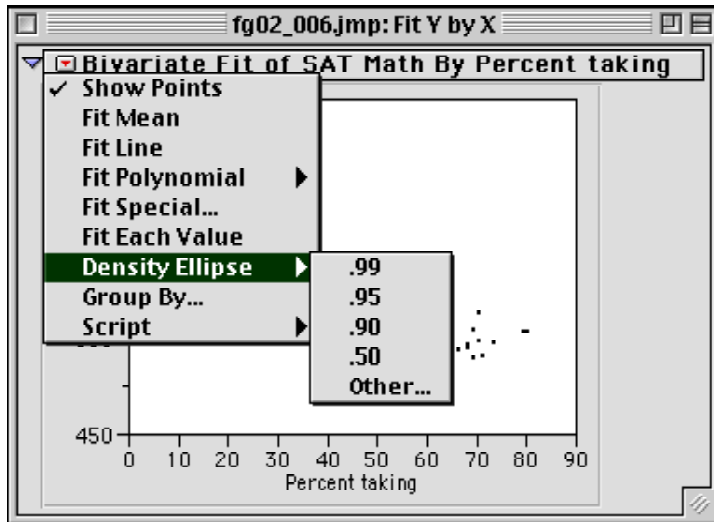
To find the *correlation* between two quantitative variables, we use the **Density Ellipse** command in the **Fit Y by X** platform in JMP INTRO.

IPS Figure 2.1 State SAT scores revisited

The scatterplot for a state's mean SAT mathematics score and the percent of its high school seniors who take the exam shows a somewhat strong negative linear relationship between the variables. Let's calculate the *correlation*. We first display the scatterplot.

1. Open the JMP data table **fg02_001.jmp** on the IPS CD-ROM if it is closed.
2. Select **Analyze** ⇒ **Fit Y by X**.
3. Select the column **SAT Math** and click **Y, Response**.

- Select the column **Percent taking** and press **X, Factor** and **OK**.
- Click on the red triangle in the **Bivariate Fit** report title and select **Density Ellipse** ⇒ **.95**. (It doesn't matter which number you select.)



- Open the **Correlation** report by clicking the disclosure button next to **Correlation**. Notice that $r = -0.86072$.

Correlation					
Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
Percent taking	35.4902	26.2864	-0.86072	0.0000	51
SAT Math	529.2745	34.83451			

Remark

In JMP IN and in the professional version of JMP, the **Multivariate** analysis platform can also be used to calculate correlations.

2.3 Models for Relationships: Least-Squares Regression

To fit the least-squares regression line, use the **Fit Line** command in the red triangle menu for scatterplots.

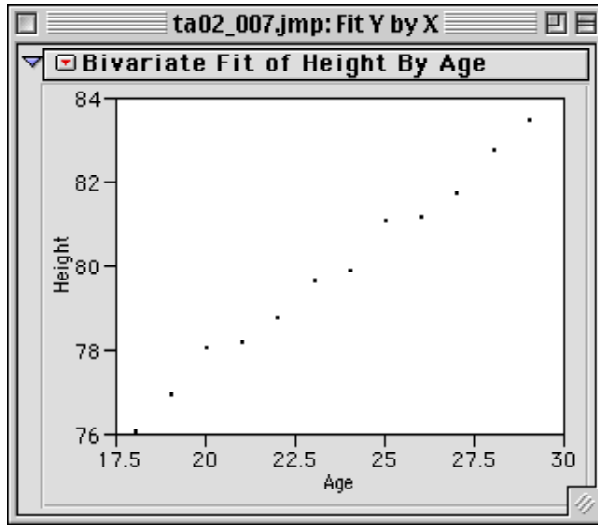
IPS Example 2.10 How do children grow?

Because the pattern of growth of children varies from child to child, we can best understand the general pattern by following the average height of a number of children. Table 2.7 of IPS presents the mean heights of a group of children in Kalama, an Egyptian village. We first create a scatterplot to examine the relationship between age and average height. Age is the explanatory variable so we wish to plot it on the x -axis. The text file **ta02_007.txt** on the IPS CD-ROM contains the pairs of values.

1. Import the file **ta02_007.txt** into a JMP data table (see Section 0.2.1 in Chapter for details) and name the columns of data values **Age** and **Height** respectively.

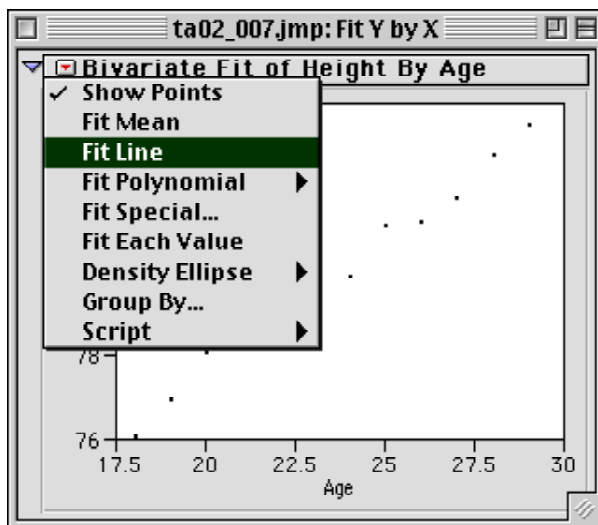
Display a scatterplot of **Height** by **Age**.

2. Select **Analyze** ⇒ **Fit Y by X**.
3. Select **Height** and **Y, Response**.
4. Select **Age**, and press **X, Factor** and **OK**.

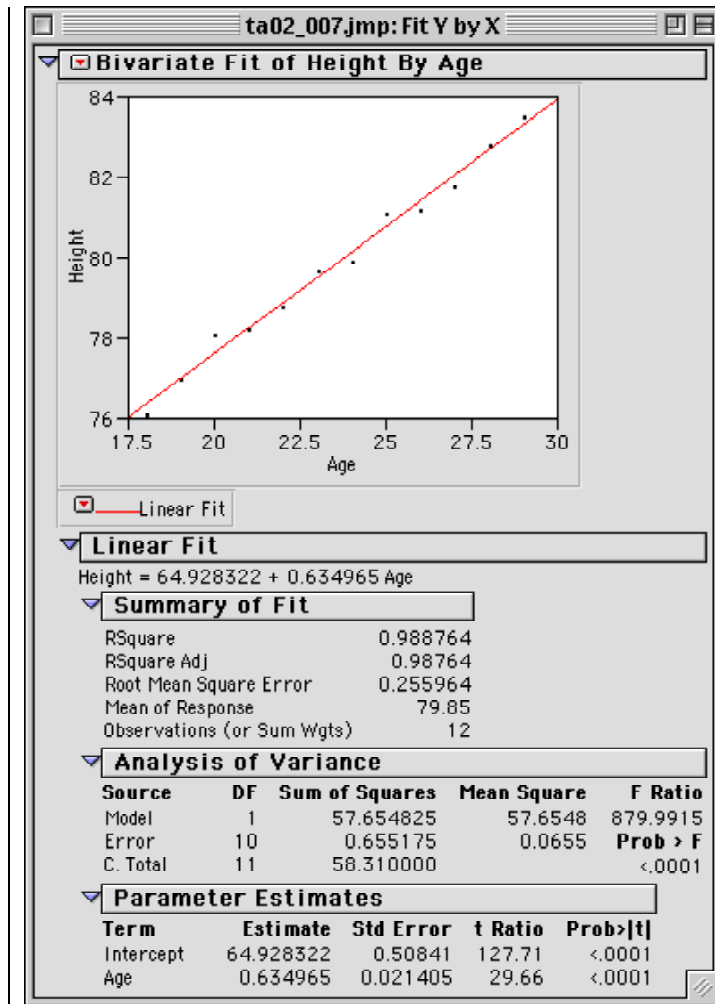


The plot shows a strong linear relationship with no outliers. A straight line will serve as a good model for the relationship between the **age** and **height** of Kalama children.

Fit the least-squares regression line to the data.



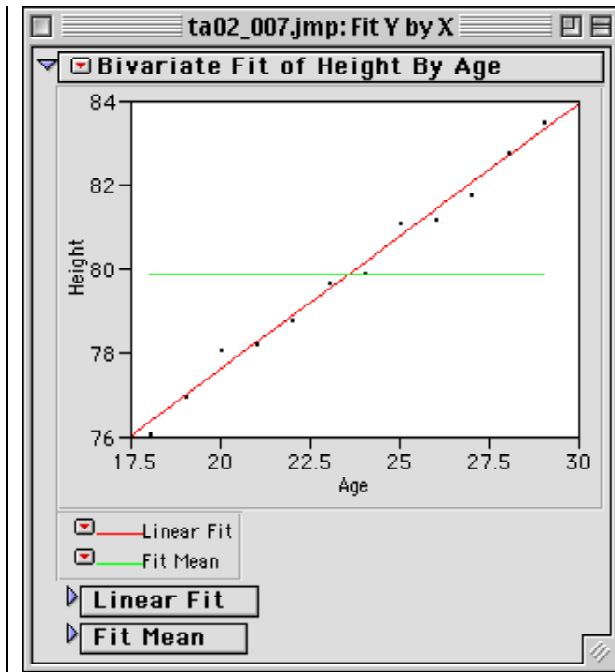
5. Press the red triangle and select **Fit Line** from the pop-up menu.



RSquare, r^2

The least-squares equation can be found directly under the **Linear Fit** title bar **Height = 64.9 + 0.635 Age**; *RSquare*, (r^2), can be found directly under the **Summary of Fit** title bar **RSquare = $r^2 = 0.988764$** . Recall that r^2 is the proportion of the variability in **Height** that is explained by the least squares regression of **Height** on **Age**. To see the variability in **Height** better,

Select **Fit Mean** from the pop-up menu next to **Bivariate Fit of Height by Age**.




Compare the variability of the points about the horizontal green line with that about the tilted red line. The vertical distances of the points from the red (regression) line are considerably less than from the horizontal green line.

Prediction

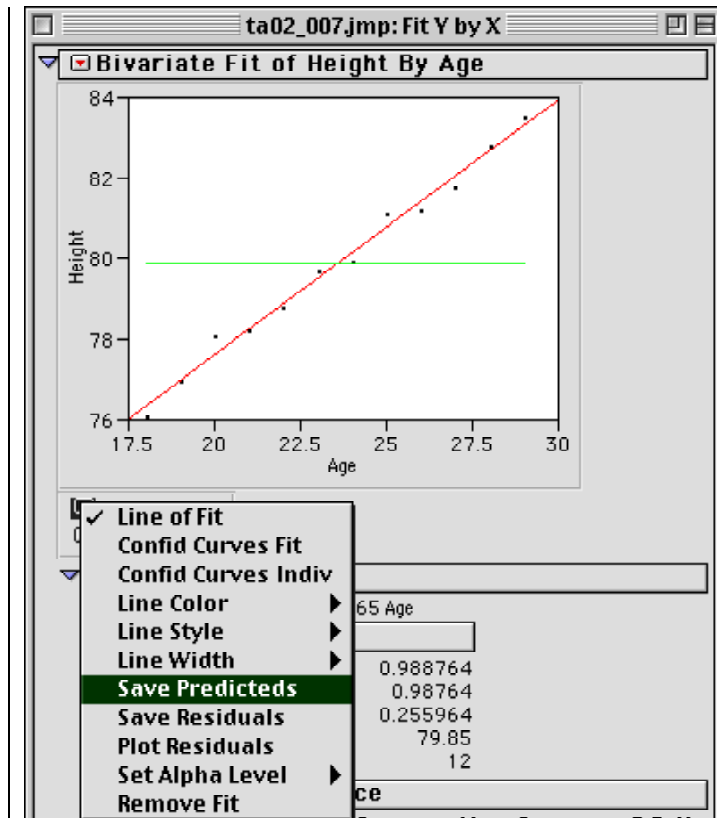
We can use JMP to *predict* the response for a specific value of the explanatory variable x . For example, we might want to predict the mean height of Kalama children at 20 months of age.

IPS Example 2.10 How do children grow? (cont'd.)

1. Select the **crosshair** tool  from the **Tools** palette.
2. Place the cursor, now resembling a crosshair, on the line directly above 20 and press. The *predicted value* of the mean **Height**, 77.6 inches, for children at 20 months of age is displayed.

JMP can also store the *predicted values* for each of the individuals in the data table.

3. Press the red triangle that is directly below the scatterplot and next to the **Linear Fit** title bar, and select **Save Predicteds** from the menu that opens.



4. Select the **ta02_007** data table window and notice that a new column **Predicted Height** was created to hold the predicted value for each observation.

	Age	Height	Predicted Height
1	18	76.1	76.3576923
2	19	77	76.9926573
3	20	78.1	77.6276224
4	21	78.2	78.2625874
5	22	78.8	78.8975524

2.4 Assessing the Fit: Residuals, Outliers, and Influential Observations

Besides fitting models that describe the overall pattern of a relationship, JMP helps to assess the appropriateness of a fitted model and to identify striking deviations from that model. The professional statistician does these tasks first before examining r^2 , the equation of the line, or predicting the response for a specific value of the explanatory variable.

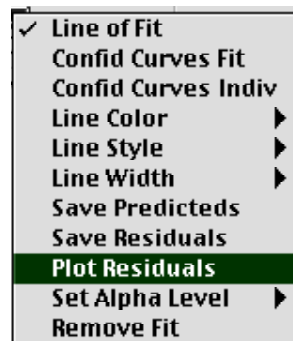
Residuals

Residuals are the vertical deviations of the observed data points from the corresponding predicted values on the least-squares regression line. As such, they represent deviations of the regression model from the

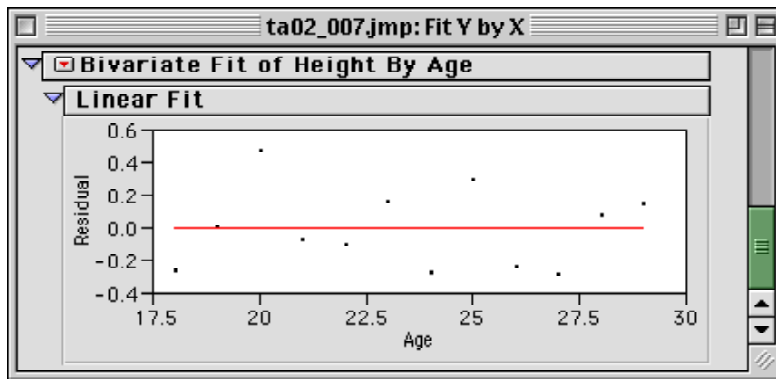
data points, and a plot of the residuals can help you assess the appropriateness of a regression line as a model for the data. Plotting is a task at which JMP excels. With one command, you can obtain a residual plot; with another command, you can tell JMP to calculate all the residuals and store them in the original data table for later use.

IPS Example 2.10 How do children grow? (ont'd.)

1. Bring the report window **ta02_007: Fit Y by X** forward. (If you no longer have the window available, repeat the first 5 steps in Section 2.3.)
2. To plot the residuals against the explanatory variable for the linear fit, select **Plot Residuals** from the pop-up menu located directly below the scatterplot next to **Linear Fit** (not the one next to **Fit Mean**, if you are using the report from Section 2.3).



Since the plot is a random band of points centered at zero, the least-squares model, **Height = 64.9 + 0.635 Age**, provides an appropriate description of the relationship between height and age.



3. To save the residuals to the JMP data table, select **Save Residuals** from the red triangle menu located directly below the scatterplot next to **Linear Fit**.

ta02_007.jmp					
4/1 Cols	Age	Height	Predicted Height	Residuals Height	
12/0 Rows	1	18	76.1	76.3576923	-0.2576923
	2	19	77	76.9926573	0.00734266
	3	20	78.1	77.6276224	0.47237762
	4	21	78.2	78.2625874	-0.0625874
	5	22	78.8	78.8975524	-0.0975524

Outliers and Influential Observations

In addition to judging the appropriateness of a regression line as a model for the data, we need to look for striking individual points—*outliers* and *influential observations*. *Outliers* are points that are outlying in the y , or vertical, direction while points that are outlying in the x , or horizontal, direction are potentially *influential observations*. Both can be identified using residual plots. To judge the influence of a point outlying in the x direction, we must find the regression line with and without the suspect point.

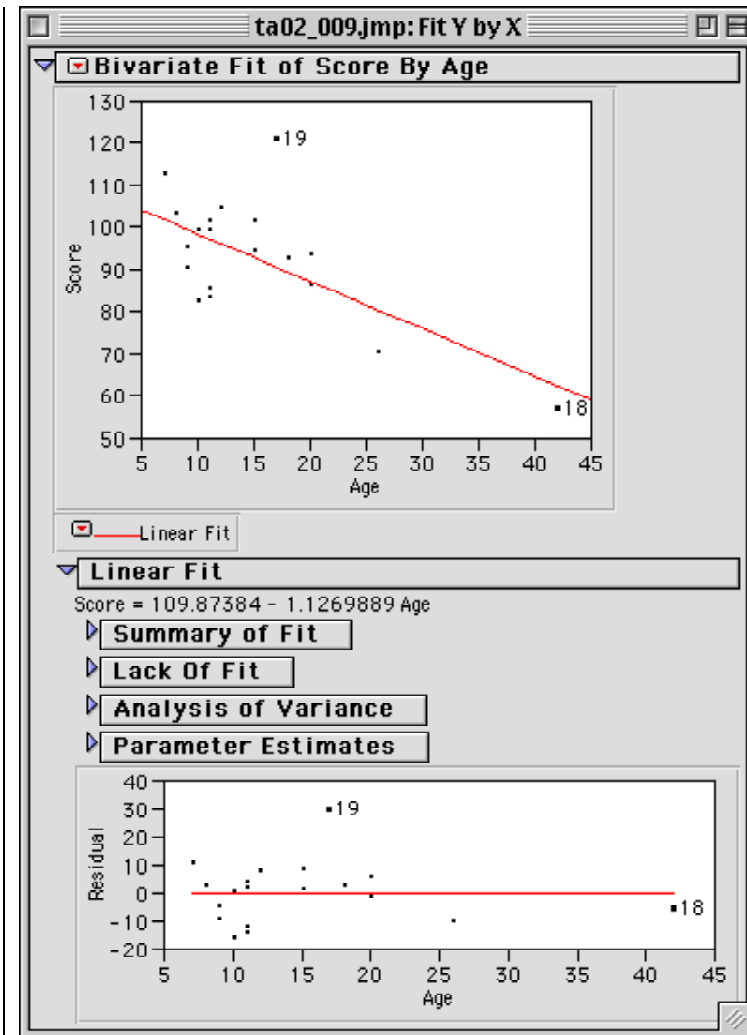
IPS Example 2.18 Age at which a child begins to talk

Example 2.18 of the textbook presents data from a cognitive study of children designed to investigate the relationship of the age at which a child begins to talk and a later score on a test of mental ability. Let's determine if there are any outliers or influential observations in the data. The data file **ta02_009.txt** on the IPS CD-ROM contains the data from this study.

1. Import the text file **ta02_009.txt** on the IPS CD-ROM into a JMP data table (see Section 0.2.1 in Chapter 0 for details) and name the columns **Child**, **Age**, and **Score**.

Fit the least-squares regression line and obtain a residual plot.

2. Select **Analyze** ⇒ **Fit Y by X**.
3. Select **Score** and press **Y, Response**.
4. Select **Age**, and press **X, Factor** and **OK**.
5. Press the red triangle on the **Bivariate Fit of Score by Age** title bar and select **Fit Line** from the menu that opens.
6. Press the red triangle next to **Linear Fit**, which is directly below the scatterplot, and select **Plot Residuals**.
7. Identify the children associated with the outlying points by holding the cursor over them.

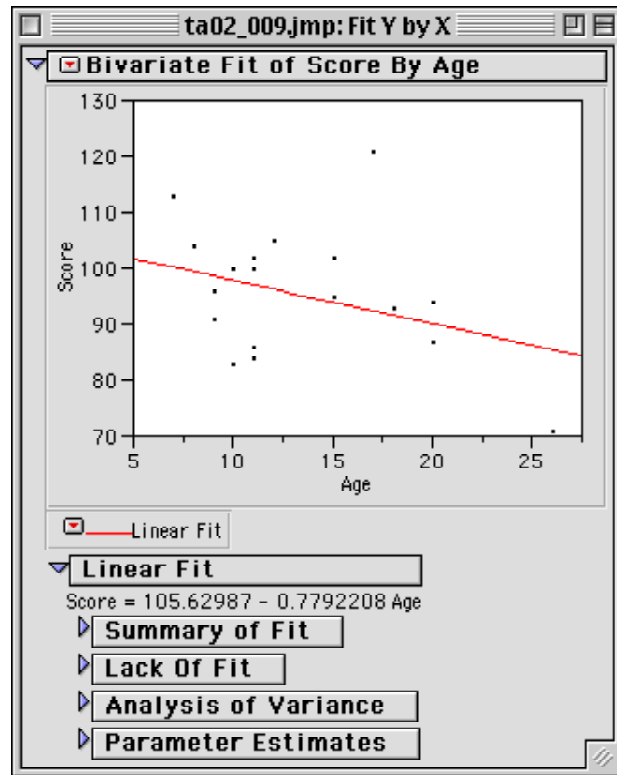


Child 19 is an *outlier*. Its vertical deviation from the model is much larger than for other children. The model does not fit this child well. Child 18, while having a small residual, is outlying in the x direction and, as such, is potentially *influential*. Click on it and notice that the corresponding point in the scatterplot of **Score by Age** is highlighted. To investigate the influence of child 18 on the fitted line, you need to exclude this child and refit the least-squares regression line. To do this, simply change the row state of child (row) 18 in the data table to **Exclude**. (Row states are discussed in more detail in Section 0.4 in Chapter 0.)

8. Select **Window** \Rightarrow **ta02_009** to bring the data table to the front.
9. If row 18 (child 18) is not highlighted, select **Row 18**.
10. Select **Rows** \Rightarrow **Exclude/Include** and notice that the exclusion symbol \emptyset appears in the row number area next to row 18 at the left of the data grid.
11. Select **Window** \Rightarrow **ta02_009: Fit Y by X** to bring the report to the front.
12. To have JMP automatically duplicate the analysis without child 18, press the red triangle on the **Bivariate Fit of Score by Age** title bar at the top.

13. Select **Script** ⇒ **Redo Analysis**.

Compare this scatterplot and the equation of this line with those in the previous report that included child 18. The slope of the least squares regression line has been substantially changed. Child 18 is an influential observation.



2.6 Transforming Relationships

Nonlinear relationships between variables can sometimes be changed into linear ones by transforming one or both of the variables. To transform variables in JMP, you create a new variable using the formula editor.

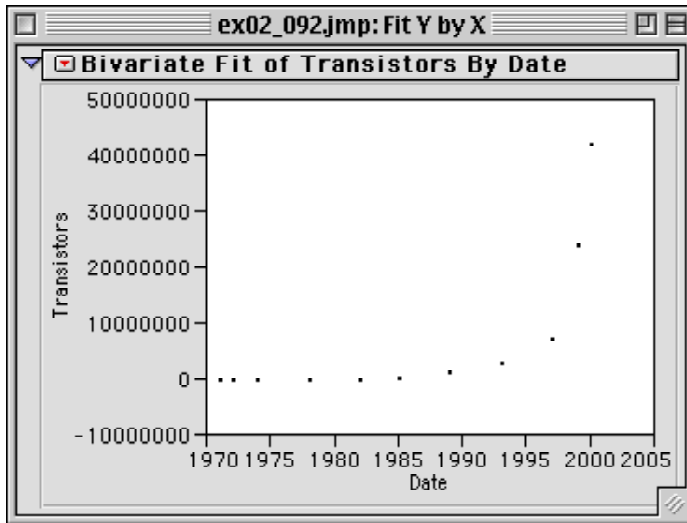
IPS Exercise 2.92 Moore's law

Gordon Moore, one of the founders of Intel Corporation, predicted that the number of transistors on an integrated circuit chip would double every 18 months. This has become known as “Moore’s law” for microprocessing power. The text file **ex02_092.txt** contains data on the number of transistors on microprocessors made by Intel since 1971. Show that a log transformation of the number of transistors indicates linear growth and hence that an exponential growth model is correct.

1. Open the JMP data table **MooresLaw.jmp** that you created in Section 0.2.2 in Chapter 0. If this JMP data table cannot be found, import the text file **ex02_092.txt** from the IPS CD-ROM into a JMP data table and name the columns **Processor**, **Date**, and **Transistors**.

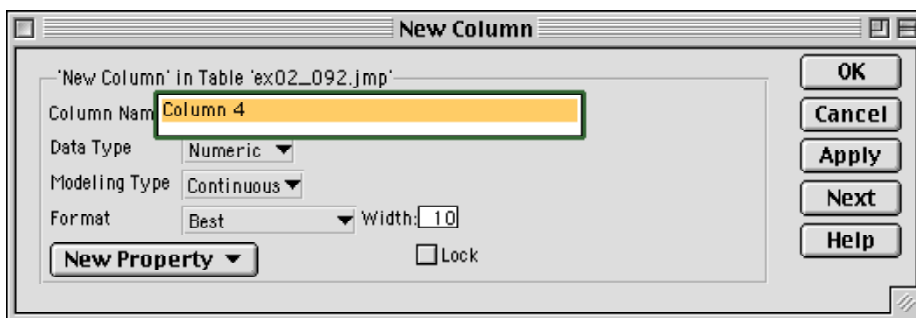
Obtain a scatterplot of the number of transistors versus time.

2. Select **Analyze** ⇒ **Fit Y by X**.
3. Select **Transistors** and press **Y, Response**.
4. Select **Date**, and press **X, Factor** and **OK**.



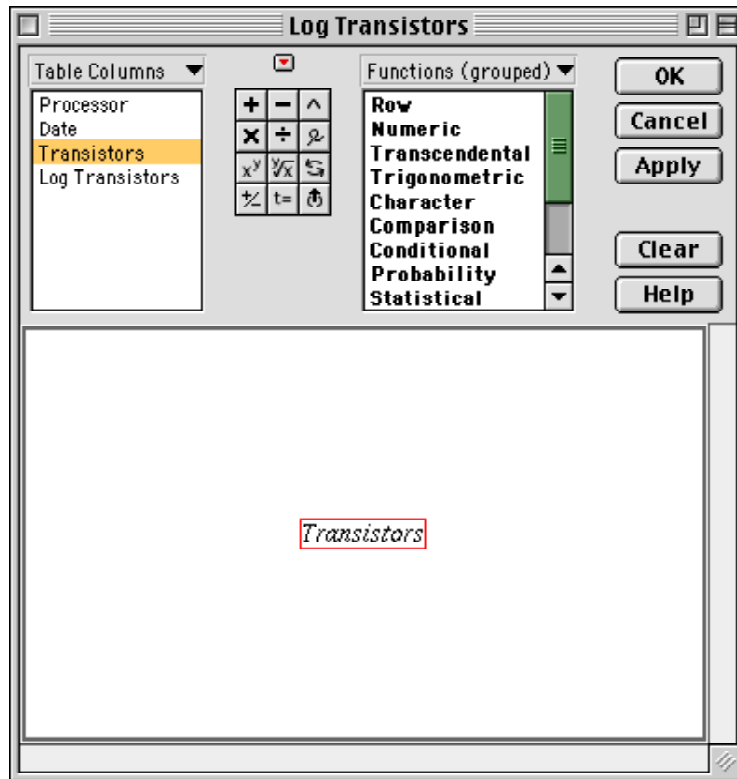
The pattern is characteristic of exponential growth—slow growth for an extended period followed by explosive growth. If it is exponential growth, a logarithmic transformation of the number of transistors will straighten the pattern. We will create a new variable to hold the logarithms of the number of transistors and plot it against time.

5. Go to the JMP data table **ex02_092** and select **Cols** ⇒ **New Column** to create a column of logarithms of the values in the column **Transistors**.



6. Enter **Log Transistors** in the **Column Name** field of the **New Column** panel.
7. Select **Fixed Dec** from the **Format** menu and enter **3**.
8. Press **New Property** and select **Formula**.
9. From the list of columns, select **Transistors**.

10. From the list of **Functions (grouped)**, select **Transcendental** ⇒ **Log**.



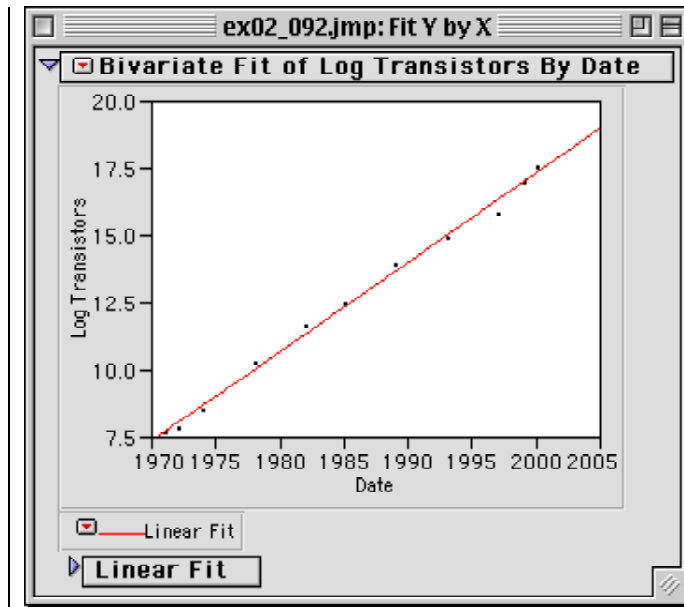
11. Press **OK** and **OK** again.

	Processor	Date	Transistors	Log Transistors
1	4004	1971	2250	7.719
2	8008	1972	2500	7.824
3	8080	1974	5000	8.517
4	8086	1978	29000	10.275

Now plot the logarithm of the number of transistors against time and fit a least-squares regression line to the data.

12. Select **Analyze** ⇒ **Fit Y by X**.
- Select **Log Transistors** and press **Y, Response**.
 - Select **Date**, and press **X, Factor** and **OK**.
13. Press the red triangle next to the **Bivariate Fit of Log Transistors by Date** title bar and select **Fit Line** from the menu that opens.

The relationship between the logarithms of the number of transistors and time is quite linear. Hence, exponential growth is a good model for the relationship of the number of transistors and time.



2.7 Summary

All graphs and statistical computations in this chapter are performed in the second platform **Fit Y by X** of the **Analyze** menu.

Graph/Computation	Command
Displaying relationships	Analyze ⇒ Fit Y by X
Scatterplots	
Side-by-side boxplots	
Side-by-side means diamonds	
Describing relationships	Analyze ⇒ Fit Y by X ⇒ Density Ellipse
Models for relationships	Analyze ⇒ Fit Y by X ⇒ Fit Line

To transform a variable, create a new variable using the **New Column** command, on the **Cols** menu, and the formula editor.

2.8 Exercises

Use JMP to help carry out the following exercises from the textbook. Data for the exercises can be imported from text files on the IPS CD-ROM with names corresponding to the exercise number or associated table number and a suffix of *.txt*.

1. Wine consumption and heart attacks. Exercise 2.7.
2. World record times for 10,000-meter races. Exercise 2.15.
3. Fidelity Investments sector funds. Exercise 2.17.

4. Color attraction for cereal leaf beetles. Exercise 2.18.
5. Vanguard International Growth Fund and the EAFE index. Exercise 2.27.
6. Speed and gas consumption. Exercise 2.29.
7. World record times for 10,000-meter races. Exercise 2.45. The **Group By** command in the **Fitting** menu (next to the scatterplot name) can tell JMP to fit separate regression lines for men and women.
8. Literacy. Exercise 2.55.
9. Hot dogs. Exercise 2.60. The JMP data table **HotDogs.jmp** on the IPS CD-ROM contains the data.
10. Poverty and MD's. Exercise 2.64.
11. Life span and body weight. Exercise 2.105.
12. CSDATA. Exercise 2.132.