Chapter 9
Descriptive Statistics for Bivariate Data

9.1 Introduction

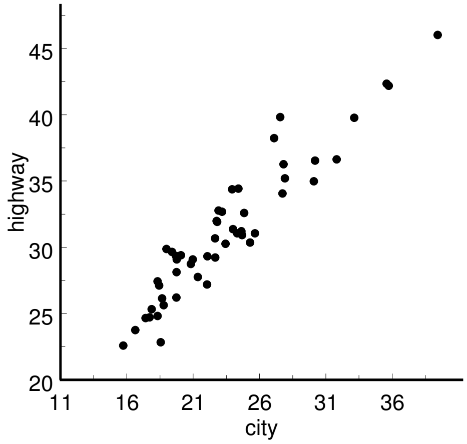
We discussed univariate data description (methods used to explore the distribution of the values of a single variable) in Chapters 2 and 3. In this chapter we will consider bivariate data description. That is, we will discuss descriptive methods used to explore the joint distribution of the pairs of values of a pair of variables. The **joint distribution** of a pair of variables is the way in which the pairs of possible values of these variables are distributed among the units in the group of interest. When we measure or observe pairs of values for a pair of variables, we want to know how the two variables behave together (the joint distribution of the two variables), as well as how each variable behaves individually (the marginal distributions of each variable).

In this chapter we will restrict our attention to bivariate data description for two quantitative variables. We will make a distinction between two types of variables. A **response variable** is a variable that measures the response of a unit to natural or experimental stimuli. A response variable provides us with the measurement or observation that quantifies a relevant characteristic of a unit. An **explanatory variable** is a variable that can be used to explain, in whole or in part, how a unit responds to natural or experimental stimuli. This terminology is clearest in the context of an experimental study. Consider an experiment where a unit is subjected to a treatment (a specific combination of conditions) and the response of the unit to the treatment is recorded. A variable that describes the treatment conditions is called an explanatory variable, since it may be used to explain the outcome of the experiment. A variable that measures the outcome of the experiment is called a response variable, since it measures the response of the unit to the treatment. For example, suppose that we are interested in the relationship between the gas mileage of our car and the speed at which our car is driven. We could perform an experiment by selecting a few speeds and then driving our car at these speeds and calculating the corresponding mileages. In this example the speed at which the car is driven is the explanatory variable and the resulting mileage is the response variable. There are also situations where both of the variables of interest are response variables. For example, in the Stat 214 example we might be interested in the relationship between the height and weight of a student; the height of a student and the weight of a student are both response variables. In this situation we might choose to use one of the response variables to explain or predict the other, *e.g.*, we could view the height of a student as an explanatory variable and use it to explain or predict the weight of a student.

9.2 Association and Correlation

The first step in exploring the relationship between two quantitative variables X and Y is to create a graphical representation of the ordered pairs of values (X, Y) which constitute the data. A **scatterplot** is a graph of the n points with coordinates (X, Y) corresponding to the n pairs of data values. When both of the variables are response variables, the labeling of the variables and the ordering of the coordinates for graphing purposes is essentially arbitrary. However, when one of the variables is a response variable and the other is an explanatory variable, we need to adopt a convention regarding labeling and ordering. We will label the response variable Y and the explanatory variable X and we will use the usual coordinate system where the horizontal axis (the X–axis) indicates the values of the explanatory variable X and the vertical axis (the Y–axis) indicates the values of the response variable Y. With this standard labeling convention, the scatterplot is also called a plot of Y versus X. Some of the scatterplots in this section employ jittering (small random displacements of the coordinates of points) to more clearly indicate points which are very close together.

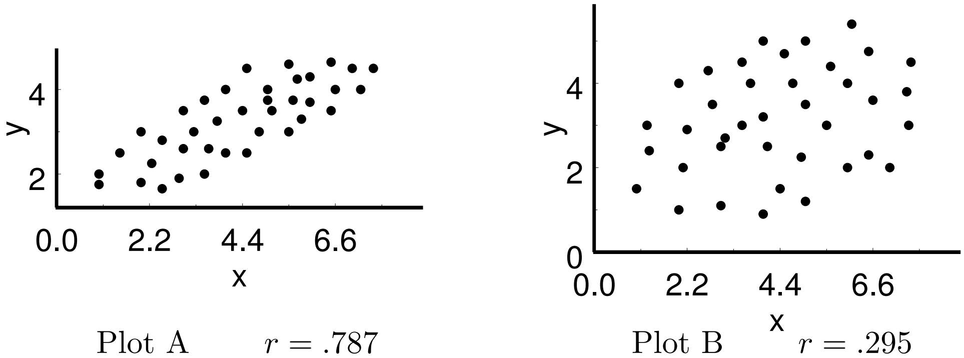
Figure 1. Subcompact car highway mileage versus city mileage.



A scatterplot of the highway EPA mileage of a subcompact car model versus its city EPA mileage, for the $n = 51$ subcompact car models of the example in Section 3.1 (excluding the 5 unusual models), is given in Figure 1. There is an obvious trend or pattern in the subcompact car mileage scatterplot of Figure 1. A subcompact car model with a higher city mileage value tends to also have a higher highway mileage value. This relationship is an example of positive association. We can also see that the trend in this example is more linear than nonlinear. That is, the trend in the subcompact car mileage scatterplot is more like points scattered about a straight line than points scattered about a curved line.

The two plots in Figure 2 illustrate positive linear association. Moving to the right in the X direction we see that the points tend to move upward in the Y direction. That is, as the value of X increases the value of Y tends to increase as well. This linear association (linear trend) is stronger in plot A than it is in plot B. The quantity r provided with these plots is a measure of linear association which will be explained later.

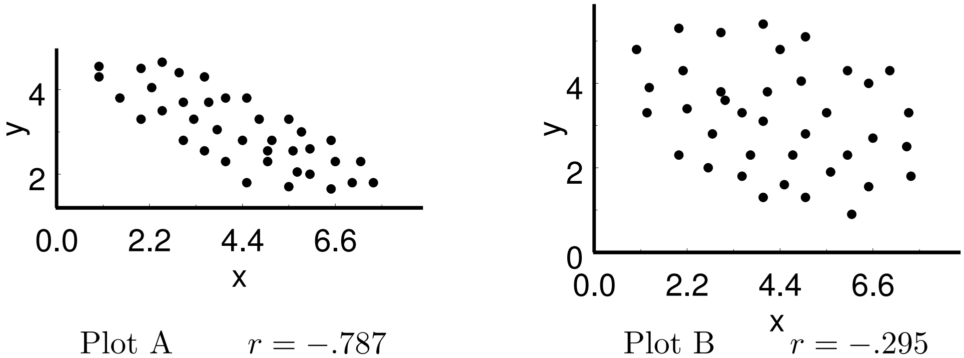
Figure 2. Examples of positive linear association



Plot A $r = .787$ Plot B $r = .295$

The two plots in Figure 3 illustrate negative linear association. Moving to the right in the X direction we see that the points tend to move downward in the Y direction. That is, as the value of X increases the value of Y tends to decrease. Again, this linear association (linear trend) is stronger in plot A than it is in plot B.

Figure 3. Examples of negative linear association.



Plot A $r = -.787$ Plot B $r = -.295$

We might describe the points in a scatterplot as forming a point cloud. A useful heuristic approach to the idea of linear association is provided by picturing an ellipse drawn around the point cloud. By envisioning ellipses drawn around the points in the plots of Figures 2 and 3, we can make the following observations. When there is positive linear association, the long direction (major axis) of the ellipse slopes upward; and when there is negative linear association, the long direction of the ellipse slopes downward. Moreover, the width of the ellipse in the direction perpendicular to the long direction (the minor axis) indicates the strength of the linear association. That is, a narrower ellipse indicates stronger linear association than does a wider ellipse. Please note that it is the width of the ellipse and not the steepness of the long direction of the ellipse that indicates strength of linear association.

It is difficult, even with a lot of experience, to determine precisely how strong the linear association between two variables is from a scatterplot. Therefore we need to define a numerical summary statistic that can be used to quantify linear association.

We first need to quantify the location of the center of the point cloud in the scatterplot. We will use the two means $\overline{X}$ and $\overline{Y}$ to quantify the center (location) of the point cloud in the X–Y plane. That is, the point with coordinates $(\overline{X}, \overline{Y})$ will serve as our quantification of the center of the point cloud (the center of the ellipse around the data).

To motivate the statistic that we will use to quantify linear association we need to describe the notions of positive and negative linear association relative to the point $(\overline{X}, \overline{Y})$. If X and Y are positively linearly associated, then when X is less than its mean $\overline{X}$ the corresponding value of Y will also tend to be less than its mean $\overline{Y}$; and, when X is greater than its mean $\overline{X}$ the corresponding value of Y will also tend to be greater than its mean $\overline{Y}$. Therefore, when X and Y are positively linearly associated the product $(X - \overline{X})(Y - \overline{Y})$ will tend to be positive. On the other hand, if X and Y are negatively linearly associated, then when X is less than its mean $\overline{X}$ the corresponding value of Y will tend to be greater than its mean $\overline{Y}$; and when X is greater than its mean $\overline{X}$ the corresponding value of Y will tend to be less than its mean $\overline{Y}$. Therefore, when X and Y are negatively linearly associated the product $(X - \overline{X})(Y - \overline{Y})$ will tend to be negative. This observation suggests that an average of these products of deviations from the mean, $(X - \overline{X})(Y - \overline{Y})$, averaging over all n such products, can be used to determine whether there is positive or negative linear association.

If an average of the sort described above is to be useful for measuring the strength of the linear association between X and Y, then we must standardize these deviations from the mean. Therefore, the statistic that we will use to quantify the linear association between X and Y is actually an "average" of the products of the standardized deviations of the observations from their means (Z–scores). This "average" of n values is computed by dividing a sum of n terms by $n - 1$, just as we divided by $n - 1$ in the definition of the standard deviation. Linear association is also known as **linear correlation** or simply **correlation**; and the statistic that we will use to quantify correlation is called the correlation coefficient. The **correlation coefficient** (Pearson correlation coefficient), denoted by the lower case letter r, is defined by the formula

$$r = \sum \left(\frac{X - \overline{X}}{S_X} \right) \left(\frac{Y - \overline{Y}}{S_Y} \right) / (n - 1) \ .$$

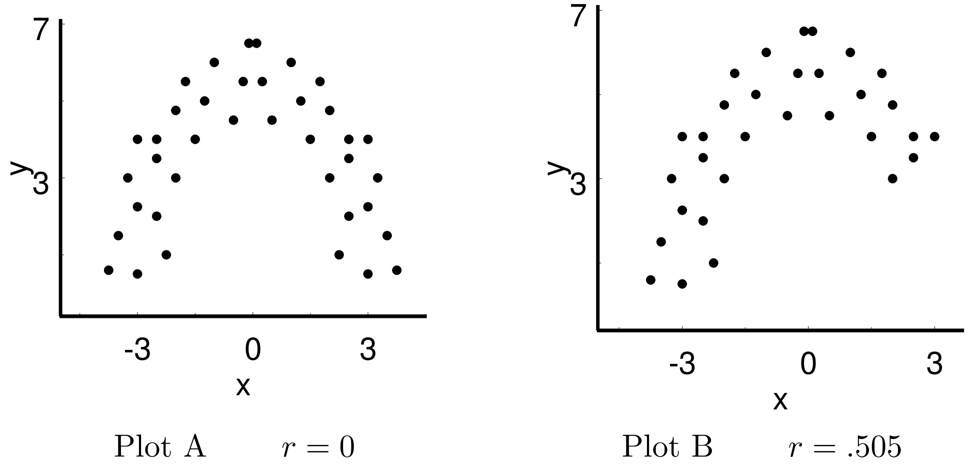
In words, the **correlation coefficient** r is the "average" of the products of the pairs of standardized deviations (Z–scores) of the observed X and Y values from their means. This formula for r is not meant to be used for computation. You should use a calculator or a computer to calculate the correlation coefficient r.

The correlation coefficient is a unitless number that is always between -1 and 1. The sign of r indicates the direction of the correlation between X and Y. A positive r indicates positive correlation and a negative r indicates negative correlation. If $r = 1$, then the variables X and Y are perfectly positively correlated in the sense that the points lie exactly on a line with positive slope. If $r = -1$, then the variables X and Y are perfectly negatively correlated in the sense that the points lie exactly on a line with negative slope. If $r = 0$, then the variables are uncorrelated, *i.e.*, there is no linear correlation between X and Y.

The magnitude of r indicates the strength of the correlation between X and Y. The closer r is to one in absolute value the stronger is the correlation between X and Y. The correlation coefficients for the plots in Figures 2 and 3 are provided below the plots. The correlation coefficient for the highway and city mileage values for the 51 subcompact car models plotted in Figure 1 is $r = .9407$ indicating that there is a strong positive correlation between the city and highway mileage values of a subcompact car.

In many situations the relationship between two variables may involve nonlinear association. The plots in Figure 4 illustrate two versions of nonlinear association. In both plots, as the value of X increases the value of Y tends to increase at first and then to decrease. In plot A of Figure 4 there is no linear association between X and Y (in this plot the ellipse would either be a circle or the long direction of the ellipse would be exactly vertical) and the correlation coefficient is zero. In plot B of Figure 4 there is a positive linear component to the nonlinear association between X and Y (the ellipse would slope upward) and the correlation coefficient is positive.

Figure 4. Examples of nonlinear association.

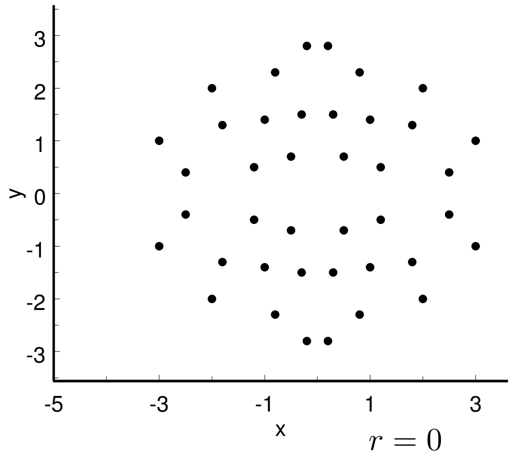


Plot A $r = 0$

Plot B $r = .505$

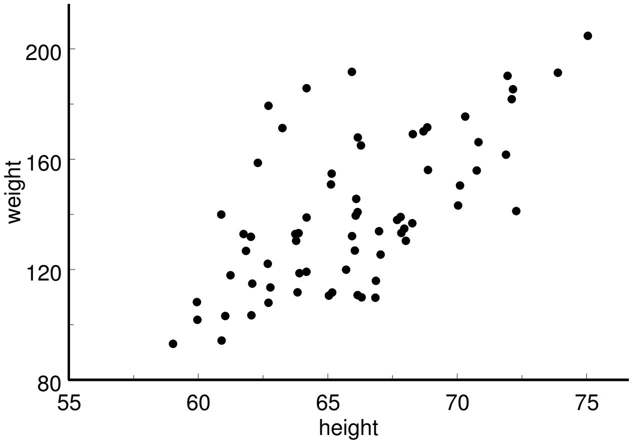
Plot A of Figure 4 illustrates a situation where there is association between X and Y but there is no correlation (no linear association). The plot of Figure 5 illustrates a situation where there is no association at all between X and Y. When there is no association, the points in the scatterplot appear to be randomly scattered about with no evidence of a trend, linear or nonlinear, and the correlation coefficient is zero. The correlation coefficient is also zero when the long direction of the ellipse around the point cloud is horizontal or vertical.

Figure 5. An example of no association.



Example. Weights and heights for the Stat 214 example. The scatterplot in Figure 6 shows the relationship between the weights (in pounds) and heights (in inches) of the $n = 67$ students in the Stat 214 example of Chapter 1.

Figure 6. Stat 214 weight and height example.



This scatterplot of weight versus height shows positive linear association between these variables. The correlation coefficient is $r = .6375$ which indicates a moderate positive correlation between the weight of a Stat 214 student and his or her height. This means that there is some tendency for a student who is heavier than average to also be taller than average; and similarly, for a student who is lighter than average to also be shorter than average.

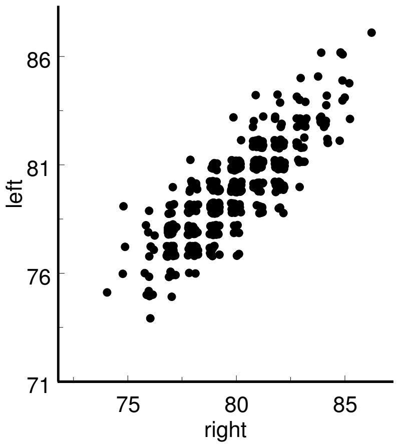
Example. Bee forewing vein length. This example is taken from Sokal and Rohlf, *Biometry*, (1969). The data are from Phillips, *Cornell Exp. Sta. Mem* **121**, (1929). These data consist of right and left forewing vein lengths (in mm times 50) for a sample of 500 worker bees. We would expect larger bees to tend to have larger wings on both sides of the body, and wing vein lengths should reflect this positive association. Thus the purpose of this example is to assess the evidence for this type of symmetry in worker bees. The data are summarized in the form of a joint frequency distribution with appended marginal frequency distributions in Table 1, and a plot of the data is provided in Figure 7. There is strong positive correlation between right and left forewing vein lengths for these bees. The correlation coefficient $r = .8372$, which quantifies the strength of linear association between these measurements, provides a measure of developmental homeostasis (physiological stability) for worker bees.

Table 1. Bee forewing vein length data.

This table provides the joint and marginal frequency distributions of the right and left forewing vein lengths (in mm times 50) for 500 worker bees.

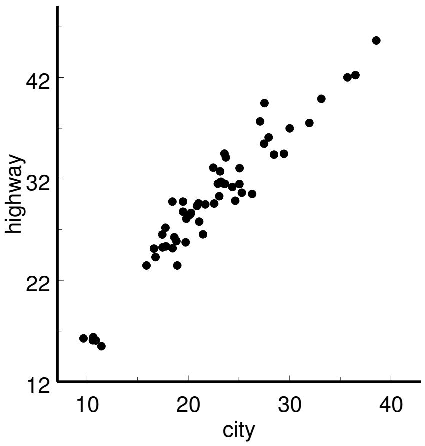
	left vein length														
right vein length	74	75	76	77	78	79	80	81	82	83	84	85	86	87	row freq.
86														1	1
85									1	1	2	2	3		9
84									3	7	2	1	1		14
83							1	4	6	16	3	1			31
82						3	13	19	12	5	2				54
81						8	18	22	19	3	1				71
80				2	6	23	46	23	1	1					102
79				10	17	34	25	8							94
78			2	14	19	12	11	1							59
77		1	4	19	14	6	1								45
76	1	5	2	4	3	1									16
75			1	1		1									3
74		1													1
column freq.	1	7	9	50	59	88	115	77	42	33	10	4	4	1	500

Figure 7. Bee forewing vein length example.



When examining a scatterplot we may find one or more unusual pairs of data values. That is, we may find that there is a point in the plot that is widely separated from the majority of the points in the plot. If the relationship between the coordinates of the unusual point agree with the overall linear pattern of the other points, then the unusual point will have the effect of strengthening the linear association between X and Y. Such an unusual point lengthens and narrows the ellipse and causes the magnitude of the correlation coefficient to increase ($|r|$ gets larger). If the relationship between the coordinates of the unusual point does not agree with the overall linear pattern of the other points, then the unusual point will have the effect of weakening the linear association between X and Y. Such an unusual point makes the ellipse wider and causes the magnitude of the correlation coefficient to decrease ($|r|$ gets smaller).

Figure 8. Subcompact car highway mileage versus city mileage (all 56 models).



The scatterplot of the highway EPA mileage of a subcompact car model versus its city EPA mileage in Figure 8 includes the values for all $n = 56$ subcompact car models (including the 5 unusual models). Here we see that the points corresponding to the five unusual car models are separated from the other points but agree with the overall linear trend. In this example including these five unusual car models increases the correlation coefficient from .9407 to .9596.

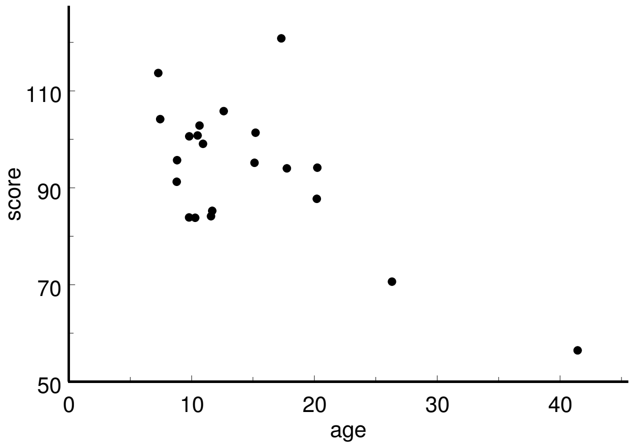
When the data include an unusual point it is a good idea to verify that the data were recorded correctly, since an error might produce an unusual point. Assuming that no data error exists, it might be a good idea to compute the correlation coefficient twice, once with all of the data and once with the unusual point not included. If there is a substantial difference between these two correlation coefficients, then appropriate comments can be added to your discussion of the problem. Another possible reason for an unusual point is the lack of enough relevant data. That is, the separation between the unusual point and the others may be due to the omission of data the inclusion of which would eliminate the separation. Therefore, a substantial difference between the two correlation coefficients (computed with and without the unusual point) might also warrant the collection of additional data for further investigation.

Example. Age at first word and Gesell test scores. This example is concerned with the relationship between the age at which a child begins to use words and the score the child attains on a test of mental ability given at a later age. The data given in Table 2 are the age at which a child spoke its first word and the score that the child attained on the Gesell adaptive test (the test was administered at a much later age). The data used in this example are from Mickey, Dunn, and Clark, *Comput. Biomed. Res.*, (1967) as reported in Rousseeuw and Leroy, *Robust Regression and Outlier Detection*, (1987). There are two response variables in this example: the age (in months) at which the child spoke its first word and the child's score on the Gesell adaptive test.

Table 2. Age at first word and Gesell score data.

child number	age at first word	Gesell score	child number	age at first word	Gesell score
1	15	95	12	9	96
2	26	71	13	10	83
3	10	83	14	11	84
4	9	91	15	11	102
5	15	102	16	10	100
6	20	87	17	12	105
7	18	93	18	42	57
8	11	100	19	17	121
9	8	104	20	11	86
10	20	94	21	10	100
11	7	113			

Figure 9. Plot of Gesell score versus age at first word.



The scatterplot given in Figure 9 shows negative linear association between these two variables. The correlation coefficient is $r = -.6403$ which indicates a moderate negative correlation between the age at which the child spoke its first word and the score that the child received on the Gesell adaptive test. This means that there is some tendency for a child who speaks its first word earlier to score higher on the Gesell test than a child who speaks its first word later.

An examination of the scatterplot for these data reveals that there are at least two unusual points in this data set.

The point (17, 121), corresponding to child number 19, is unusual in the sense that this point is separated from the other points and this pair of values does not agree with the overall negative linear trend in the data. This child spoke its first word at the age of 17 months and scored 121 on the Gesell adaptive test. These values do not fit into the overall pattern of the data for the other 20 children. Judging from the overall pattern we would expect a child that spoke its first word at age 17 months to have a Gesell score in the neighborhood of 90. Therefore, the score for this child appears to be unusually high. One possible explanation for this is that there might have been an error in recording the values for this child. This possibility was checked and it was determined that no error had been made and these are the correct values for this child. There is no justification for removing this child from the study; however, it is instructive to note that if we compute the correlation coefficient for the other 20 children omitting child number 19 we get $r = -.7561$. Thus we see that this single child (single pair of values) has a fairly large influence on the magnitude of the correlation coefficient.

The point (42, 57), corresponding to child number 18, is unusual in the sense that there is a large separation between this point and the other points in the scatterplot; but, this pair of values does agree with the overall negative linear trend in the data. These characteristics cause this pair of values to have a large influence on the linear trend in the data. This child spoke its first word at the age of 42 months and scored 57 on the Gesell adaptive test. The age at first spoken word of 42 months is very large relative to the ages at first spoken word for the other 20 children in this group. Because of the large separation between this child's age at first spoken word and the ages at first spoken word for the other children in this group, the Gesell adaptive test score of this child will exert a large influence on the overall pattern in the data. In this example the Gesell test score for this child is such that it agrees with and strengthens the overall pattern in the data. If we compute the correlation coefficient for the 20 other children omitting child number 18, we get $r = -.3349$. Therefore, without child number 18 there is a fairly small negative correlation between the age at which a child spoke its first word and the score that the child received on the Gesell adaptive test. As with child 19 we find that this single child 18 (single pair of values) has a fairly large influence on the magnitude of the correlation coefficient.

The two unusual points discussed above demonstrate the two types of unusual points we might find in a correlation problem. At this point we will examine the present example in more detail. Notice that the point (26,71), corresponding to child number 2, is also somewhat separated from the other points in the scatterplot. We see that there are two children, child number 18 and child number 2, who spoke their first words later than the majority of the children. If we omit these two children and recompute the correlation coefficient for the 18 other children we get $r = -.0340$. This shows that the evidence for a negative correlation between age at first word and Gesell test score is very highly dependent on the presence of these two children. It would be a good idea to obtain some more data so that we could determine whether these two children really are unusual or whether we simply do not have much information about children who are late in speaking their first word.

9.3 Regression

Regression analysis is used to study the dependence of a response variable on an explanatory variable. It may be helpful to think of the explanatory variable X as a measurement of an input to a system and the response variable Y as a measurement of the output of the system. If there was an exact linear functional relationship between X and Y, then the response variable Y (the output) could be expressed as a linear function of the explanatory variable X (the input). That is, if there was an exact linear relationship, then there would exist constants a and b such that, for a given value of the explanatory variable X, the corresponding value of the response variable Y could be expressed as $Y = a + bX$. If this was the case, then the points in the scatterplot would lie on the line determined by the equation $Y = a + bX$.

In practice, the linear relationship between X and Y will not be exact and the points (corresponding to the observed values of X and Y) in the scatterplot will not lie exactly on a line. Therefore, assuming that there is a linear relationship between X and Y, we want to determine a line (a linear equation relating X and Y) which adequately summarizes the linear relationship between X and Y. Another way to say this is that we want to determine the line which best fits the data. Of course we first need to decide what we mean by saying that a line fits best. Therefore, we will first define a measure of how well a line fits the data.

The measure of the quality of the fit of a line to the data that we will use is based on the vertical deviations of the observed values of the response variable Y from the corresponding values on the proposed line. The motivation for basing the measure of quality of fit on vertical deviations is that we are using the fixed values of the explanatory variable X to explain the variability in the response variable Y and variation in Y is in the vertical direction. If Y is the response variable value for a particular value of X and $\hat{Y}$ (read this as Y hat) is the value that we would have observed if the relationship was exactly linear, then the deviation $Y - \hat{Y}$ is the signed distance from the point we observed (X, Y) to the point $(X, \hat{Y})$ on the line having the same X coordinate. The deviation $Y - \hat{Y}$ is positive when the point (X, Y) is above the line and negative when the point (X, Y) is below the line. Notice that this deviation is the signed vertical distance from the line to the point (X, Y) along the vertical line through the point X on the X–axis.

If the line fits the data well, then we would expect the points in the scatterplot to be close to the line. That is, we would expect the deviations $Y - \hat{Y}$ to be small in magnitude. We would also expect a line that fits well to pass through the "middle" of the point cloud. That is, we would expect the signs of the deviations $Y - \hat{Y}$ to vary between positive and negative with no particular pattern.

The quantity that we will use to summarize the quality of fit of a line to the data is the sum of the squared deviations of the observed Y values from the $\hat{Y}$ values predicted by the line. In symbols, this quantity is $\sum(Y - \hat{Y})^2$. When comparing the fit of two lines to the data we would conclude that the line for which this sum of squared deviations is smaller provides a better fit to the data.

The **least squares regression line** is the line which yields the best fit in the sense of minimizing the sum of squared deviations of the points from the line. That is, among all possible lines, the least squares regression line is the line which yields the smallest possible sum of squared deviations. It is a mathematical fact that the least squares regression line is the line that passes through the point $(\overline{X}, \overline{Y})$ and has slope b given by the formula

$$b = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sum(X - \overline{X})^2} \ .$$

This formula for the slope of the least squares regression line is provided to show you that there is such a formula and is not meant to be used for computation. You should use a calculator or a computer to calculate the least squares regression line slope b.

If you refer back to the definition of the correlation coefficient, r, you will see that the formula for r is symmetric in X and Y. That is, interchanging the labels assigned to the two response variables has no effect on the value of r. On the other hand, the formula for the slope of the least squares regression line is clearly not symmetric in X and Y. This asymmetry reflects the fact that, in the regression context, the roles of the explanatory variable X and the response variable Y are not interchangeable.

Let $(X, \hat{Y})$ denote the coordinates of a point on the least squares regression line. The definition of the least squares regression line given above and the definition of the slope of a line imply that

$$b = \frac{\hat{Y} - \overline{Y}}{X - \overline{X}} \; .$$

Straightforward manipulation of this expression (first multiply both sides by $(X - \overline{X})$, then add $\overline{Y}$ to both sides) yields the equation

$$\hat{Y} = \overline{Y} + b(X - \overline{X})$$

for the least squares regression line. This equation is called **the mean and slope form of the equation of the least squares regression line**, since it depends on the mean $\overline{Y}$ and the slope b. Simple regrouping of terms shows that the least squares regression line equation can also be written as

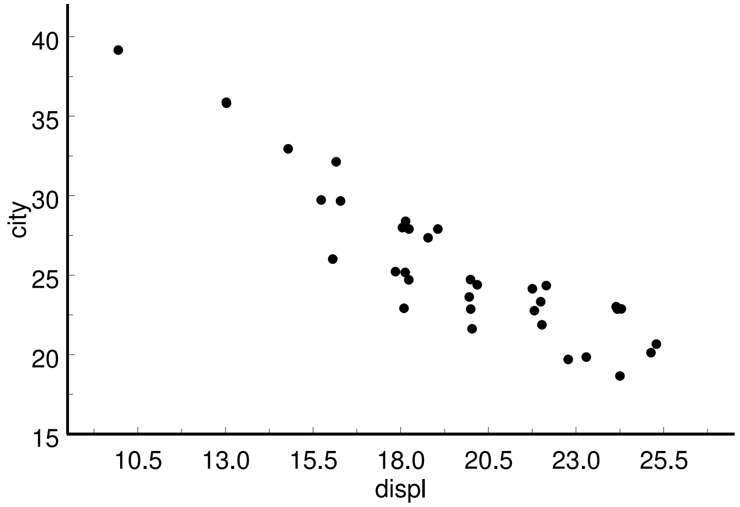
$$\hat{Y} = a + bX,$$

where $a = \overline{Y} - b\overline{X}$ is the y–intercept. You can obtain the value of the y–intercept a of the least squares regression line using a calculator or a computer. This equation is called **the intercept and slope form of the equation of the least squares regression line**, since it depends on the y–intercept a and the slope b. Of these two forms, the intercept and slope form of the equation of the least squares regression line is more convenient for most purposes.

The slope b of the least squares regression line is the (constant) rate of change of $\hat{Y}$ as a function of X, *i.e.*, if we start at a particular point $(X, \hat{Y})$ on the line and move one unit to the right in the X direction, then $\hat{Y}$ moves b units in the Y direction giving the point $(X + 1, \hat{Y} + b)$. If b is positive, the change in $\hat{Y}$ is upward ($\hat{Y}$ increases); and if b is negative, the change in $\hat{Y}$ is downward ($\hat{Y}$ decreases).

Notice that, according to the least squares regression line, $\overline{Y}$ is the response variable value that we would expect to see when $X = \overline{X}$. That is, substituting $X = \overline{X}$ into the least squares regression line equation gives $\hat{Y} = \overline{Y}$. The intercept value, a, is the response value that we would expect to see, according to the least squares regression line, when $X = 0$. That is, substituting $X = 0$ into the least squares regression line equation gives $\hat{Y} = a$.

Figure 10. Plot of EPA city mileage versus displacement, for the 35 cars with displacements no larger than 2.5 liters.



Example. Subcompact car city mileage and displacement. The displacement (size) of a car engine will clearly have an effect on the gas mileage that the car will obtain. The purpose of this example is to examine the dependence of the EPA city mileage of a subcompact car model on its engine displacement. For this example we will restrict our attention to the $n = 35$ subcompact car models with engine displacements that are no larger than 2.5 liters. For ease of discussion, we will convert the engine displacements from liters to hundreds of cubic centimeters. To convert the engine displacements of Table 4 of Section 3.1 from liters to 100 cc units we simply multiply by ten, since one liter is 1000 cc. The data for this example are also provided in Table 4 of this chapter.

From the scatterplot of city mileage versus engine displacement given in Figure 10 we see that there is strong negative linear association between the city mileage of a subcompact car and its engine displacement (the correlation coefficient is $r = -.9112$). Therefore, for subcompact car models with engine displacements no greater than 2.5 liters it makes sense to use a straight line to summarize and quantify the dependence of city mileage on engine displacement.

Let X denote the engine displacement (in 100 cc's) of a subcompact car model and let Y denote its EPA city mileage (in mpg). Some relevant summary statistics are provided in Table 3. The y–intercept is $a = 49.1590$ mpg and the slope is $b = -1.2020$ mpg/100cc; therefore, the equation of the least squares regression line is $\hat{Y} = 49.1590 - 1.2020X$.

Table 3. Subcompact car EPA city mileage and engine displacement summary statistics, for the 35 car models with displacements no larger than 2.5 liters.

$n =$	35		
$\overline{X} =$	19.4571	$r =$	-.9112
$\overline{Y} =$	25.7714	slope, $b =$	-1.2020
$S_X =$	3.6730	y–intercept, $a =$	49.1590
$S_Y =$	4.8452	$R^2 =$	.8303

The slope -1.2020 is our estimate of the constant rate of change of Y, the subcompact car's city mileage, as a function of X, the car's engine displacement. According to the least squares regression line, for each 100 cc increase (each one unit increase on the 100 cc scale of measurement) in the engine displacement of a subcompact car, we expect to see a decrease (since the slope is negative) of about 1.2020 mpg in the EPA city mileage value. Recalling that the least squares regression line passes through the point $(\overline{X}, \overline{Y})$ on substituting $\overline{X}$ into the equation we see that, according to the least squares regression line, when a subcompact car has an engine displacement of 1945.71 cc its EPA city mileage should be approximately 25.7714 mpg. The nearest engine displacement for which we have data is 1900 cc and the two subcompact car models with a 1900 cc engine displacement have actual EPA city mileages of 27 and 28 mpg; therefore, the least squares regression line prediction for a 1900 cc engine is only slightly lower than these two observed values. The y–intercept is not so easily interpreted. According to the least square regression line, when a subcompact car model has an engine displacement of 0 cc its EPA city mileage should be approximately 49.1590 mpg. Clearly this does not make sense, since an engine displacement of 0 cc is nonsensical; but there is a simple explanation. The linear relationship assumed when we fit the least squares regression line requires a constant rate of change in EPA city mileage as a function of engine displacement. However, there is no reason to expect this relationship to hold for all possible engine displacements. We might expect the rate of change to be different for very small engines than it is for the displacement range for which we have data (1000 cc to 2500 cc). Similarly, we might not expect the linear relationship we determined for the present displacement range to be valid for cars with engines having much larger displacements. Notice that if we use the least squares regression line to predict the EPA city mileage of a car with a large enough engine displacement, then we will get a (nonsensical) negative mileage value, since the least squares regression line will eventually cross the X–axis.

The least squares regression line relationship can be used to determine the value of Y that we would predict or expect to see for a given value of X. If X^* denotes a particular value of X, then, according to the least squares regression line, we would expect or predict that the corresponding value of Y would be

$$\hat{Y}(X^*) = a + bX^*$$

(read $\hat{Y}(X^*)$ as Y hat of X^*). The **predicted value** $\hat{Y}(X^*)$ is obtained by substituting X^* into the equation of the least squares regression line. Notice that $\hat{Y}(X^*)$ is the second coordinate of the point $(X^*, \hat{Y}(X^*))$ where the vertical line $X = X^*$ through X^* intersects the least squares regression line.

When using the least squares regression line to predict values we need to be aware of the danger of extrapolation. A prediction of a response value corresponding to a value X of the explanatory variable outside of the observed range of X values is called an **extrapolation**. Based on the data we have there is no way to tell whether the linear relationship summarized in the least squares regression line is appropriate for X values outside of the observed range of X values. Therefore, predictions of Y values based on the least squares regression line should be restricted to X values that are within the observed range of X values.

If we compute the predicted value $\hat{Y}$ for an observed value of X (in this context $\hat{Y}$ is called a **fitted value**), then we can use this fitted value to decompose the observed value Y as

$$Y = \hat{Y} + (Y - \hat{Y}) \ .$$

In words, this says that the observed value Y is equal to the fitted value $\hat{Y}$ plus the **residual value** $(Y - \hat{Y})$. Notice that the residual $(Y - \hat{Y})$ is the signed distance from the observed value Y (the point (X, Y)) to the fitted value $\hat{Y}$ (the point $(X, \hat{Y})$) along the vertical line through X. The residual is positive if $Y > \hat{Y}$ (the point is above the line) and negative if $Y < \hat{Y}$ (the point is below the line.) The residual would be zero if the observed value Y (the point (X, Y)) was exactly on the least squares regression line, *i.e.*, if $Y = \hat{Y}$.

The n residuals $(Y - \hat{Y})$ corresponding to the n observed data values can be used to assess the quality of fit of the least squares regression line to the data. If there was an exact linear relationship between X and Y, then the points would lie exactly on the least squares regression line and the residuals would all be zero. Therefore, when the least squares regression line fits the data well, all of the n residuals should be reasonably small in magnitude, *i.e.*, all of the points should be reasonably close to the line. A residual that is large in magnitude in one example may not be large in another example; therefore, it is important to assess the size of the residuals relative to the amount of variability in the observed Y values. If there is a systematic nonlinear pattern in the data, then we would expect to see a systematic pattern of lack of fit of the least squares regression line to the data points. Therefore, a systematic pattern in the residuals would provide evidence

Table 4. Subcompact car city mileage and displacement data, fitted, and residual values.

displacement	city mileage	fitted value	residual
X	Y	$\hat{Y}$	$Y - \hat{Y}$
10	39	37.1390	1.8610
13	36	33.5330	2.4670
13	36	33.5330	2.4670
15	33	31.1289	1.8711
16	30	29.9269	.0731
16	32	29.9269	2.0731
16	26	29.9269	-3.9269
16	30	29.9269	.0731
18	25	27.5229	-2.5229
18	25	27.5229	-2.5229
18	28	27.5229	.4771
18	28	27.5229	.4771
18	28	27.5229	.4771
18	23	27.5229	-4.5229
18	25	27.5229	-2.5229
19	27	26.3209	.6791
19	28	26.3209	1.6791
20	25	25.1189	-.1189
20	23	25.1189	-2.1189
20	22	25.1189	-3.1189
20	24	25.1189	-1.1189
20	24	25.1189	-1.1189
20	24	22.7149	1.2851
20	22	22.7149	-.7149
20	24	22.7149	1.2851
20	23	22.7149	.2851
20	23	22.7149	.2851
23	20	21.5129	-1.5129
23	20	21.5129	-1.5129
24	23	20.3109	2.6891
24	23	20.3109	2.6891
24	23	20.3109	2.6891
24	19	20.3109	-1.3109
25	20	19.1089	.8911
25	21	19.1089	1.8911

of systematic lack of fit of the least squares regression line to the data. Finally, if the variability of the observed Y values depends on the corresponding values of X, *e.g.*, the Y values corresponding to small values of X might exhibit less variability than the Y values corresponding to large values of X, then the least squares regression line may fit better for some intervals of X values than for other intervals of X values. This sort of behavior may be detectable from the relationship between the residual values and the corresponding X values.

The fitted and residual values for the subcompact car city mileage and displacement example are given in Table 4. In this example, all of the residuals except two are less than 3.2 in magnitude. The two large residual values are: -4.5229 corresponding to the Toyota Celica model with a 1800 cc engine and a city mileage of 23 mpg, and -3.9269 corresponding to the Honda Civic DOHC/VTEC model with a 1600 cc engine and a city mileage of 26 mpg. Because these two largest residuals are negative, we see that the observed city mileage values for these two car models are substantially smaller than we would expect them to be according to the least squares regression line. There is no obvious pattern in the signs of the residuals in this example. Hence, we can conclude that the least squares regression line provides a reasonable fit to the subcompact car mileage and displacement data, and that the least squares regression line is suitable as a summary and quantification of the dependence of the EPA city mileage of a subcompact car on its engine displacement.

The least squares regression line uses the values of the explanatory variable X to explain or account for the variability in the observed values of Y. Therefore, a measure of the amount of the variability in the observed Y values that is explained, or accounted for, by the least squares regression line can be used to quantify how well the least squares regression line explains the relationship between the X and Y data values.

The sum of the squares of the residuals $\sum(Y-\hat{Y})^2$ can be viewed as a measure of the variability in the data, taking the values of the explanatory variable X and the least squares regression line into account, since the fitted values depend on the X values and the least squares regression line. The sum of the squares of the deviations of the observed Y values from their mean $\sum(Y-\overline{Y})^2$ can be viewed as a measure of the variability in the observed Y values, ignoring the corresponding X values. Therefore, the difference between these two sums of squared deviations provides a measure of the amount of the variability in the observed Y values that is explained, or accounted for, by the least squares regression line. In symbols, the difference that we are referring to as a measure of the amount of the variability in the observed Y values that is accounted for by the least squares regression line is

$$\sum(Y-\overline{Y})^2-\sum(Y-\hat{Y})^2.$$

The ratio of this difference to the sum of the squared deviations from the mean $\sum(Y-\overline{Y})^2$ is known as **the coefficient of determination** and is denoted by R^2. The coefficient of determination is

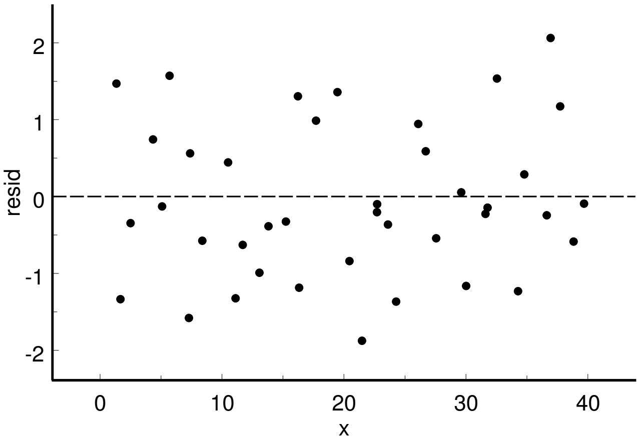
$$R^2 = \frac{\sum(Y - \overline{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \overline{Y})^2}.$$

Don't worry about having to use this formula to compute R^2. For the least square regression problem that we are considering the coefficient of determination R^2 is equal to the square of the correlation coefficient. The coefficient of determination R^2 is the proportion of the variability in the observed values of Y that is explained, or accounted for, by the least squares regression line. R^2 is a positive number between 0 and 1 and a value of R^2 close to one indicates that the least squares regression line explains the data well.

For the subcompact car mileage and displacement example, the coefficient of determination is $R^2 = .8303$. This means that, through the least squares regression line, the engine displacement of a subcompact car accounts for 83.03% of the variability in the subcompact car mileages. In other words, the least squares regression line based on engine displacement alone is actually quite successful in explaining the city mileage of a subcompact car with an engine displacement between 1 and 2.5 liters.

In addition to an examination of the residual values, we can visually assess the quality of fit of the least squares regression line to the data by plotting the line on the scatterplot of the data. A more formal graphical approach to assessing the quality of fit of the least squares regression line to the data is through a residual plot. The **residual plot** is the plot of the residuals $(Y - \hat{Y})$ versus the observed X values. An ideal residual plot should be such that all of the points are contained in a relatively narrow horizontal band centered at zero; and such that there is no obvious nonlinear pattern inside the band. An example of such an ideal residual plot is provided in Figure 11.

Figure 11. An ideal residual plot.



The residual plots in Figures 12 and 13 illustrate the two problems with the fit of the least squares regression line described earlier. The plot in Figure 12 exhibits evidence of a nonlinear trend in the data. In this example the least squares regression line is too low for small X values, it is too high for middle X values, and it is too low for large X values. We can see evidence of this behavior in the residual plot, since the residuals are positive for small X values, they are mostly negative for middle X values, and they are positive for large X values. The fan shape of the plot in Figure 13 indicates that the variability in the residuals depends on the value of X. In this example, the least squares regression line fits the data better for smaller X values than it does for larger X values. That is, the variability is smaller in the residuals corresponding to small X values than it is for the residuals corresponding to large X values.

Figure 12. A residual plot indicating poor fit.

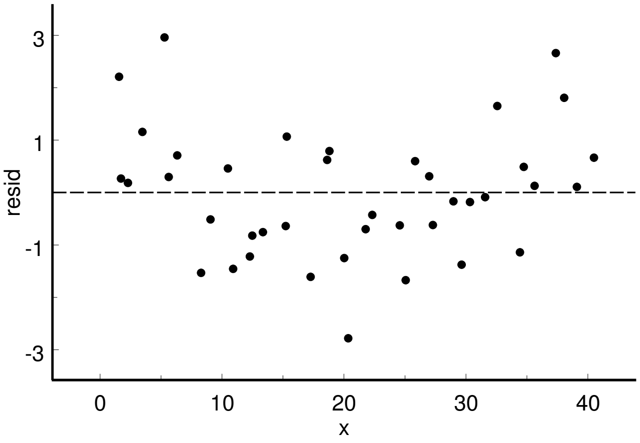
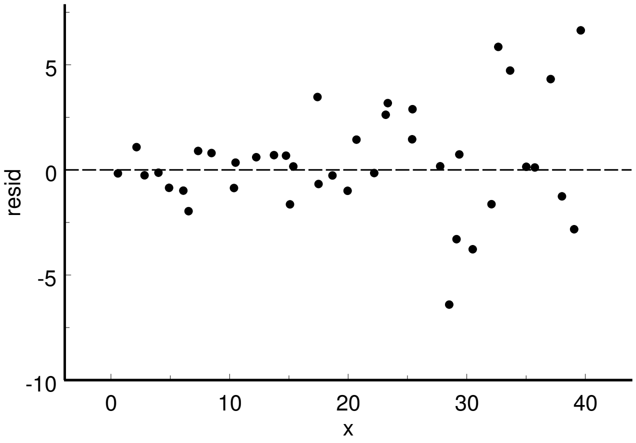


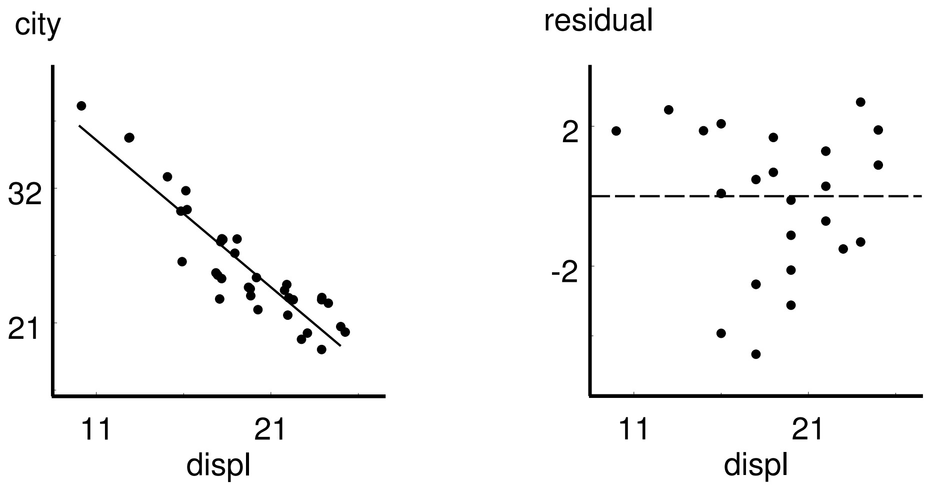
Figure 13. A residual plot indicating nonconstant variability.



When examining a residual plot you should look for obvious patterns that are supported by a reasonable number of points. Try not to let one or two slightly unusual residuals convince you that there is an overall problem. You should also be aware that an uneven distribution of X values (with more points in some X intervals than in other X intervals) may suggest problems that would disappear with the addition of a few more observations. Proper interpretation of residual plots requires practice and a sample size that is large enough to provide reliable information.

The residual plot for the subcompact car city mileage and displacement example is given in Figure 14. This figure also provides a scatterplot of the data and the fitted line. For this example the least squares regression line seems to fit the data reasonably well. However, there is a disturbing aspect of this residual plot. The first three residuals, corresponding to car models with very small engines, are positive and, as a group, these residuals are somewhat separated from the other residuals in this plot. From the scatterplot of Figure 10 we see that these three points are influential points, since the engine displacements of these three car models are small relative to the majority of the engine displacements. Before we discuss this point further, we need to discuss unusual points in the regression context.

Figure 14. Plot with fitted line and residual plot, for the 35 cars with displacements no larger than 2.5 liters.



We have already discussed the effects of unusual points on the correlation coefficient. In the regression context we will distinguish between two types of unusual points. An observation or point is said to be a **bivariate outlier** when the point does not agree with the overall linear trend in the data. A bivariate outlier may be detected visually in the scatterplot or in some examples by the fact that the corresponding residual is rather large relative to the other residuals. A scatterplot with a bivariate outlier is provided in Figure

15. For the particular situation illustrated here, the effect of the bivariate outlier would be to pull the least squares regression line upward and away from the other points in the figure. This would result in an increase in the y–intercept but little change in the slope, since this bivariate outlier is located near the middle of the observed X values.

Figure 15. A scatterplot with a bivariate outlier.

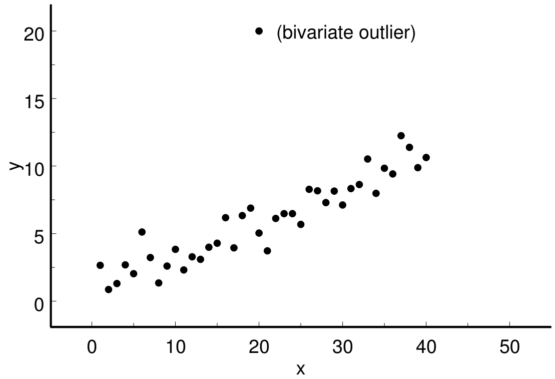
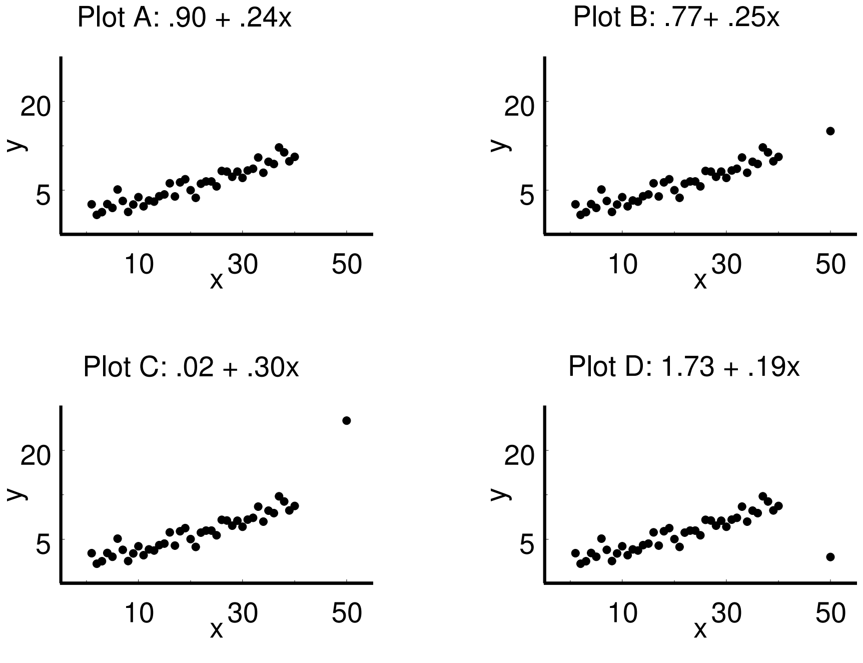


Figure 16. Scatterplots with an influential point.



An observation or point is said to be an **influential point** when the X coordinate of the point is widely separated from the X coordinates of the other points. An influential point may or may not also be a bivariate outlier. The four plots in Figure 16 illustrate the effects of an influential point on the least squares regression line slope (the equations of the regression lines are provided with the plots). The basic data for these plots is given in plot A. A single influential point is added to the data of plot A to yield the other three plots. The influential point is a bivariate outlier in plots C and D; but it is not a bivariate outlier in plot B. Notice the dramatic effect that this single point has on the slope of the regression line in plots C and D.

We will now return to our discussion of the subcompact car city mileage and displacement example and the three car models with very small engines and somewhat unusual residuals. From the listing in Table 4 we see that there are only three subcompact car models with engine displacements that are less than 1.5 liters. From the plots of Figures 10 and 14 we see that the points corresponding to these car models are influential, since their X values are somewhat separated from the other X values. These three points are not bivariate outliers, since the relationship between the city mileage and engine displacement values for these points agrees with the overall linear trend. It might be interesting to determine how much of an effect these three car models have on the least squares regression line. We are not suggesting that there is necessarily anything wrong with including the three small displacement subcompact car models in the data set, the point here is that there is not much information about small displacement cars and we would like to see just how influential these three observations are. To this end, consider the subcompact car city mileage and displacement data for the 32 car models with engine displacements between 1.5 and 2.5 liters. Summary statistics for these data are given in Table 5 and the data, fitted, and residual values are given in Table 6.

Table 5. Subcompact car EPA city mileage and engine displacement summary statistics, for the 32 car models with displacements between 1.5 and 2.5 liters.

$n =$	32		
$\overline{X} =$	20.1563	$r =$	-.8461
$\overline{Y} =$	24.7188	slope, $b =$	-1.0014
$S_X =$	2.9524	y–intercept, $a =$	44.9030
$S_Y =$	3.4941	$R^2 =$	.7160

The slope of the least squares regression line based on all 35 subcompact car models is -1.2020 mpg/100cc and the slope is -1.0014 mpg/100cc when the three small displacement car models are excluded. As we would expect from the plot of city mileage versus engine displacement, this indicates that the least squares regression line is steeper when the three small displacement car models are included than it is when they are excluded. If we include all 35 car models, then the least squares regression line indicates a decrease in the city mileage of about 1.2020 mpg for each 100cc increase in engine displacement. However, if we exclude the three small displacement car models, then the least squares regression line indicates a decrease in the city mileage of only 1.0014 mpg for each 100 cc increase in engine displacement.

Table 6. Subcompact car city mileage and displacement data, fitted, and residual values, for the 32 car models with displacements between 1.5 and 2.5 liters.

displacement	city mileage	fitted value	residual
X	Y	$\hat{Y}$	$Y - \hat{Y}$
15	33	29.8822	3.1178
16	30	28.8808	1.1192
16	32	28.8808	3.1192
16	26	28.8808	-2.8808
16	30	28.8808	1.1192
18	25	26.8780	-1.8780
18	25	26.8780	-1.8780
18	28	26.8780	1.1220
18	28	26.8780	1.1220
18	28	26.8780	1.1220
18	23	26.8780	-3.8780
18	25	26.8780	-1.8780
19	27	25.8766	1.1234
19	28	25.8766	2.1234
20	25	24.8752	.1248
20	23	24.8752	-1.8752
20	22	24.8752	-2.8752
20	24	24.8752	-.8752
20	24	24.8752	-.8752
20	24	22.8724	1.1276
20	22	22.8724	-.8724
20	24	22.8724	1.1276
20	23	22.8724	.1276
20	23	22.8724	.1276
23	20	21.8711	-1.8711
23	20	21.8711	-1.8711
24	23	20.8697	2.1303
24	23	20.8697	2.1303
24	23	20.8697	2.1303
24	19	20.8697	-1.8697
25	20	19.8683	.1317
25	21	19.8683	1.1317

A comparison of the y–intercept values suggests that the effect of the three small displacement car models on the vertical location of the least squares regression line is quite large. The y–intercept of the least squares regression line based on all 35 subcompact car models is 49.1590 mpg and the y–intercept is 44.9030 mpg when the three small displacement car models are excluded. The difference between these two y–intercepts is the vertical distance between these two lines at the nonsensical displacement value of 0 cc, *i.e.*, the vertical locations of the two the least squares regression lines, at $X = 0$, differ by 4.2560 mpg. Since the slopes of these two lines are different, it would make more sense to compare the vertical locations of these two lines at an X value that is within the observed displacement range. Predicted city mileage values for three representative values of X are given in Table 7. The representative X values are the smallest and largest values common to the two cases and a middle value $X = 20$ that is close to both X means. From these predicted values we see that excluding the three small displacement car models lowers the least squares regression line somewhat in the middle and at the lower end of the observed X range, but raises the line slightly at the upper end of the observed X range.

Table 7. Some representative predicted values.

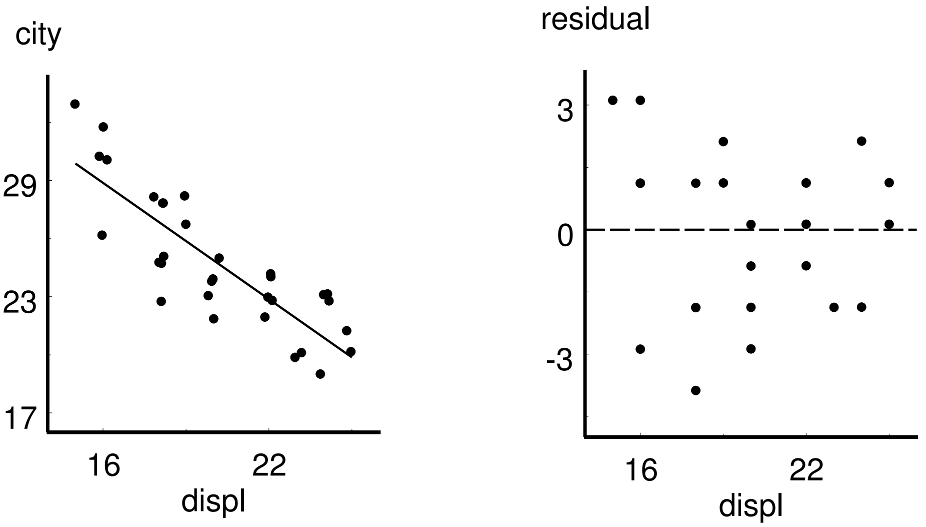
	predicted value	
displacement	including small	excluding small
15	31.1289	29.8822
20	25.1189	24.8752
25	19.1089	19.8683

The residual plot of Figure 17 does not indicate any problems with the fit of the least squares regression line when the three small displacement car models are excluded; it does show some evidence of slightly more variability for car models with smaller engine displacements. When the three small displacement car models are excluded there is only one residual that exceeds 3.2 in magnitude. The Toyota Celica model with a 1800 cc engine and a city mileage of 23 mpg has a residual of -3.8780.

The linear relationship between city mileage and displacement is somewhat weaker when the three small displacement car models are excluded. This is evident from the fact that the correlation coefficient is smaller in magnitude when these three car models are excluded. We also find that the coefficient of determination R^2 is smaller when these three car models are excluded. Since $R^2 = .7160$, we see that even without the three small displacement car models the least squares regression line still accounts for 71.6% of the variation in the subcompact car city mileage values.

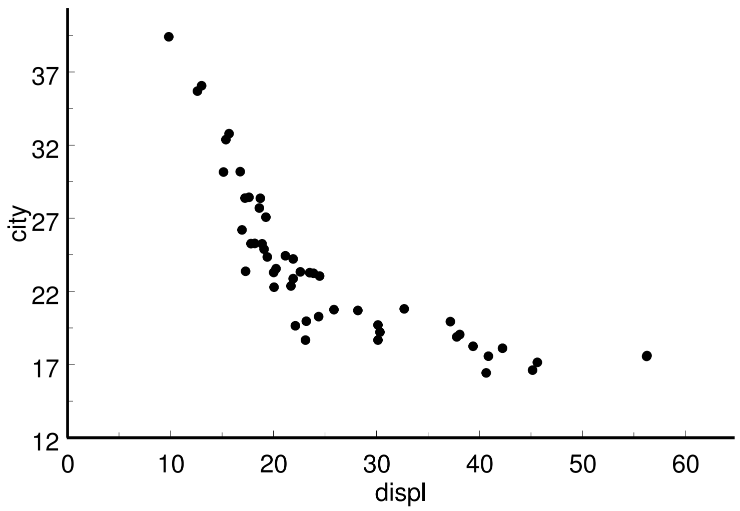
Since relatively few subcompact car models have engine displacements below 1.5 liters (only 9% of the subcompact car models with displacements no larger than 2.5 liters and only 6% of all 51 subcompact car models), we might argue that the three small displacement car models are unusual enough to justify their exclusion from our analysis of the relationship between city mileage and displacement.

Figure 17. Plot with fitted line and residual plot, for the 32 cars with displacements above 1 liter but no larger than 2.5 liters.



Our analysis of the relationship between subcompact car city mileage and engine displacement was initially restricted to car models with displacements no larger than 2.5 liters. You may have wondered why we imposed this restriction. It is instructive to reconsider the relationship between city mileage and displacement when all 51 of the subcompact car models are included. From the scatterplot of Figure 18, which is based on all 51 subcompact car models, we see that the relationship between city mileage and engine displacement is not linear when the car models with large engines are included.

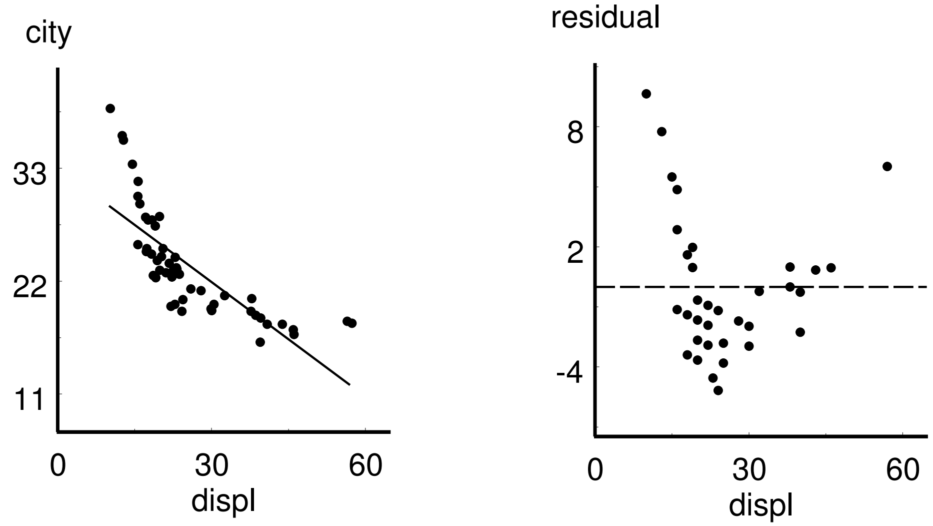
Figure 18. Plot of EPA city mileage versus displacement, for all 51 cars (excluding the 5 unusual cars).



It clearly does not make sense to fit a straight line to these data. However, we will do so to demonstrate what happens in such a situation. We would particularly like to see how this nonlinearity affects the residual plot. The residual plot for this example is given in Figure 19. There is an obvious pattern of systematic lack of fit of the least squares regression line to the data in this residual plot. For small values of X the residuals are positive and decrease as X gets larger. For middle values of X the residuals are negative and they first decrease and then increase as X gets larger. Finally, for large values of X the residuals are positive and they increase as X gets larger. This U–shaped pattern in the residual plot indicates that the least squares regression line is too low for small and large X values; and it is too high for middle X values.

In a situation like this one where there is nonlinear association we need to use more complicated regression methods that allow us to fit a curved line instead of a straight line. It is not particularly difficult to generalize the least squares approach to curved lines; however, this is beyond the scope of this chapter.

Figure 19. Plot with fitted line and residual plot, for all 51 cars (excluding the 5 unusual cars).


city
displ
residual
displ