

Unit 12

Regression Diagnostics

Objectives:

- To obtain the residuals resulting from a least squares line
- To identify potential outliers among residuals
- To identify patterns in residual plots that may suggest that the assumption that a relationship is linear is not correct

Although there are many different types of relationships which can exist between two quantitative variables, we are considering only linear relationships. When we use a least squares line to make predictions, we are accepting the *linearity assumption*, which is the assumption that whatever relationship exists must be a linear one. A scatter plot similar to one of Figure 9-1c, Figure 9-1d, Figure 9-1e, or Figure 9-1f would lead us to believe that the linearity assumption is correct. One might also say that a scatter plot similar to either Figure 9-1g or Figure 9-1h suggests that the linearity assumption is reasonable, even though both of these scatter plots appear to suggest that no strong relationship exists; this is because even if a very weak relationship were to exist, the linearity assumption would be reasonable. A scatter plot, such as the one displayed in Figure 9-1i, suggests that the relationship is not linear, but if one were interested only in the upper half of the range of values of X, it appears that the linearity assumption is not unreasonable at all.

We are not going to discuss any very sophisticated techniques for deciding when the linearity assumption is or is not reasonable; we shall simply make our best judgment by looking at plots. In the past, we have looked only considered scatter plots; however, when attempting to decide whether or not the linearity assumption is reasonable, it can also be helpful to examine *residuals*. A residual is an observed value of the response variable Y minus the corresponding predicted value of Y from the least squares line; that is,

$$residual = observed\ Y - predicted\ Y.$$

(The order of subtraction is easy to remember by just realizing that the letter “o” which is the first letter of the word “observed” comes in the alphabet before the letter “p” which is the first letter of the word “predicted”.) The residuals provide information about the variation of the observed data points around the least squares line similar to the way in which deviations from the mean provide information about variation in the distribution of a quantitative variable. In addition to using residuals to decide whether or not the linearity assumption is reasonable, we can also obtain the five-number summary for the residuals in order to identify potential outliers.

To illustrate, let us return to the Dosage and Reaction Time Data of Table 10-2. Previously, you used the calculations done in Table 10-3 to find that the least squares line for predicting reaction time (in seconds) from dosage (in grams) is $rct = 10.225 - 0.875(dsg)$. Figure 10-1 is a scatter plot of the data together with the least squares line you graphed. The

Table 12-1			
Residuals with the Data of Table 10-1,			
where Y = "Reaction Time" and X =			
"Drug Dosage"			
	<i>observed</i>	<i>predicted</i>	<i>residual</i>
<i>x</i>	<i>y</i>	<i>y</i>	<i>obs - pre</i>
4	7.5		
4	6.8		
6	4.0		
6	4.4		
8	3.9		
8	3.1		
10	1.4		
10	1.7		

vertical distances between the data points and the line are the residuals. A residual is positive if the data point lies above the line and is negative if the data point lies below the line. The residuals always sum to zero, just as the deviations from the mean always sum to zero; however, we will not attempt to prove this fact here.

Table 12-1 is designed to obtain the residuals from the Dosage and Reaction Time Data of Table 10-2. The first two columns list the observed values for $X = \text{"age"}$ and $Y = \text{"grip strength"}$ respectively. The third column is for the predicted values of Y calculated from the least squares line $rct = 10.225 - 0.875(dsg)$. When $X = 4$, the predicted value of Y is $10.225 - 0.875(dsg) = 6.725$. Since both of the first two values of X are 4 in the first column, 6.725 should be the first and second entries of the third column. Make these entries, and use the least squares line to find and enter the remaining predicted values in the third column. (You should find the remaining predicted values to be 4.975, 4.975, 3.225, 3.225, 1.475, and 1.475.)

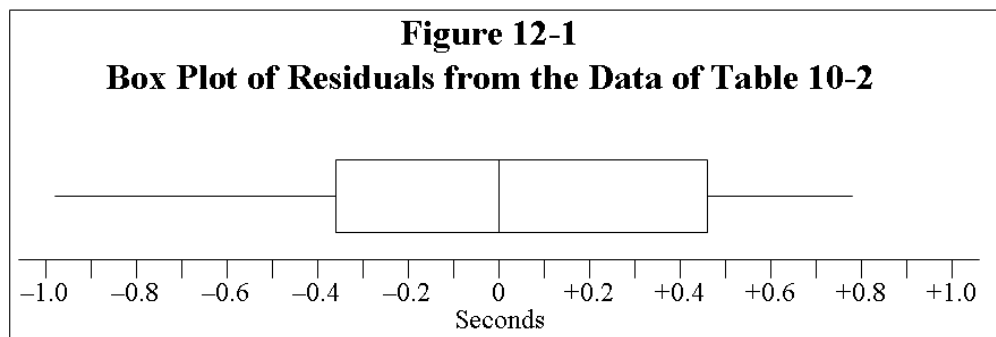
The fourth column of Table 12-1 is for the residuals, calculated by subtracting the predicted values of Y from the observed values of Y . The first residual in this fourth column should be $7.5 - 6.725 = +0.775$, the second residual should be $6.8 - 6.725 = +0.075$, and the third residual should be $4.0 - 6.725 = -0.975$. Make these entries, and find and enter the remaining residuals in this fourth column. (You should find the remaining residuals to be -0.575 , $+0.675$, -0.125 , -0.075 , and $+0.225$; you can easily verify that these residuals sum to zero.) We shall consider how to use the residuals to identify potential outliers and to decide whether or not the linearity assumption seems reasonable.

From Table 12-1, it is easy to obtain the following ordered array for the residuals:

$$-0.975 \quad -0.575 \quad -0.125 \quad -0.075 \quad +0.075 \quad +0.225 \quad +0.675 \quad +0.775 \quad .$$

From this ordered array, verify that the five-number summary is $-0.975, -0.350, 0, +0.45, 0.775$. The interquartile range is $+0.45 - (-0.35) = 0.8$. The first quartile exceeds the smallest residual of -0.975 by 0.625 , and the largest residual of $+0.775$ exceeds the third quartile by 0.325 . Since neither 0.625 nor 0.325 is greater than $(1.5)(0.8) = 1.2$, there do not appear to be any candidates for outliers. Figure 12-1 is a box plot of the residuals.

Examination of the distribution of residuals indicates to us whether or not there are any data points that might be potential outliers. In order to help us decide whether or not the linearity assumption is reasonable, we can look at



a *residual plot*. A residual plot is constructed the same way as a scatter plot, except that the vertical axis represents the residuals instead of the Y variable. Also, since the residuals always sum to zero, it is customary draw a horizontal line across the plot at zero on the vertical axis. Figure 12-2 is a residual plot constructed from Table 12-1; verify that this residual plot has been constructed correctly.

When the linearity assumption is reasonable, then we would expect to see the data points vary randomly around the least squares line; this would result in a residual plot which looks like a random scattering of dots. If the linearity assumption is not reasonable, then we would expect to see the data points vary around the least squares line in a pattern which suggests that some type of curve is more appropriate than a straight line; this would result in a residual plot which does not look random.

Look again at Figure 9-1i. This scatter plot illustrates a nonlinear relationship. Try to imagine (or go ahead and actually try to draw) the graph of the least squares line in Figure 9-1i. You should realize that for the lower values of X and for the higher values of X , the data points will almost always be below the least squares line, whereas for values of X in the middle of the range, the data points will almost always be above the least squares line. This type of nonrandom variation is often more pronounced in a residual plot than in a scatter plot, which is why a residual plot can often provide additional insight into whether or not the linearity assumption is reasonable. A residual plot corresponding to Figure 9-1i would show mostly negative residuals for the lower

values of X and for the higher values of X , whereas residuals would show mostly positive for values of X in the middle of the range.

Now look at Figure 9-1j, where you should see that the least squares line will be one with a slope close to zero (that is, an almost horizontal line). A residual plot corresponding to Figure 9-1j would show mostly positive residuals for the lower values of X and for the higher values of X , whereas residuals would show mostly negative for values of X in the middle of the range. It is these kinds of nonrandom patterns which signal to us that the linearity assumption may not be reasonable.

Not surprisingly, it is easier to decide whether or not the linearity assumption is appropriate from a large number of data points than from a small number of data points. Returning to the residual plot of Figure 12-2, it might appear that the points are not random, but we must always be careful about the conclusions we make with data based on a small number of data points such as when $n = 8$.

Figure 12-2

Residual Plot for the Data of Table 10-2

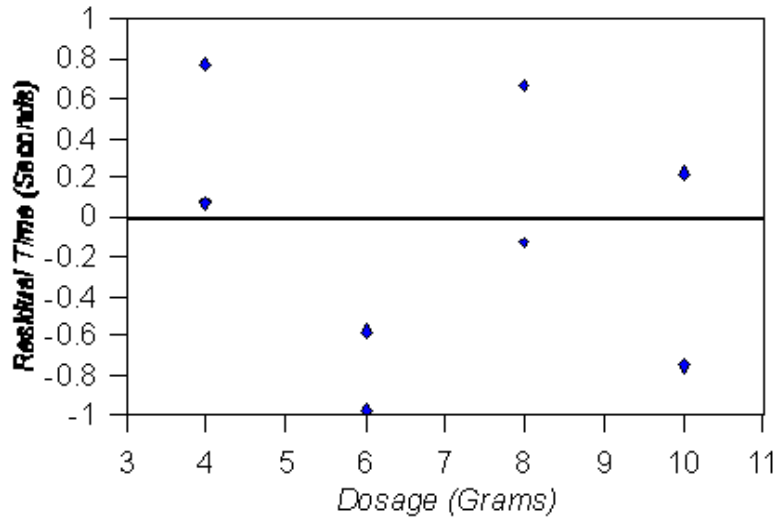
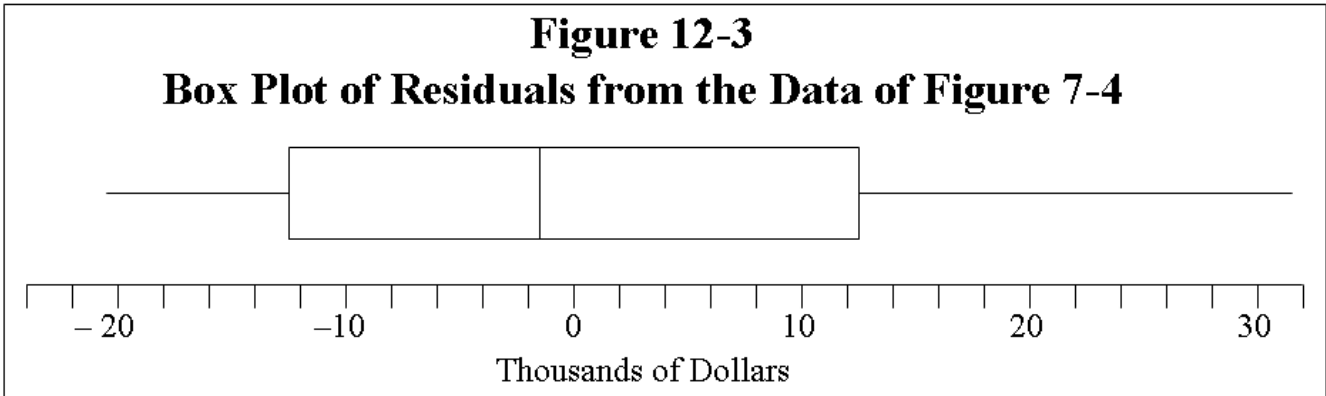


Figure 12-3

Box Plot of Residuals from the Data of Figure 7-4



Let us now return to Figure 7-4, which is a scatter plot of the ages and yearly incomes from the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1; the graph of the least squares line, which we previously found to be $\text{inc} = 16.2 + 0.661(\text{age})$ (where inc is an abbreviation for yearly income in thousands of dollars), is included in the scatter plot. Figure 12-3 is a box plot of the residuals, and Figure 12-4 is a residual plot. The five number summary for the residuals is $-20.6, -12.2, -1.8, +12.7, +31.6$, and you can easily verify that there are no outliers. Since the points in Figure 12-4 show no identifiable, nonrandom pattern among the residuals, we have no reason to doubt the linearity assumption.

Figure 12-4

Residual Plot for Figure 7-4

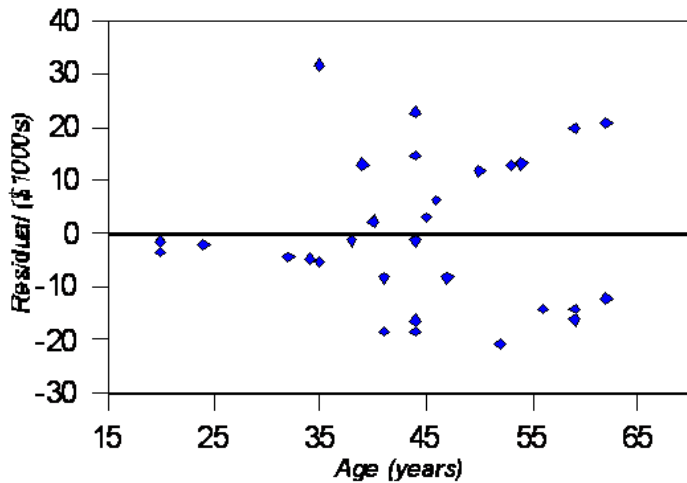


Table 12-2

Fertilizer and Corn Yield Data

One of six different amounts of a fertilizer (pounds per acre) is used on each of 24 identical plots, and corn yield (bushels per acre) are recorded for each plot.

Fertilizer Amount (lbs./acre)	Corn Yield (bushels/acre)
10	15 19
20	22 33 25 31 34
30	57 42 46 50 44
40	58 61 55 62 63
50	67 60 64 69 62
60	72 65

Self-Test Problem 12-1. In Self-Test Problem 10-1 and 11-1, the Age and Grip Strength Data, displayed in Table 10-4, is used in a study of the relationship between age and grip strength among right-handed males, with interest in the prediction of grip strength from age. The least squares line was found to be $grp = 26 + 2(age)$, where grp represents grip strength in lbs.

- Find the residuals by constructing a table similar to Table 12-1.
- Find the five-number summary for the residuals, find the interquartile range for the residuals, and construct a modified box plot of the residuals.
- Do there appear to be any candidates for outliers? Why or why not?
- Construct a residual plot.
- Decide whether or not the linearity assumption appears to be reasonable, and state why or not.

Self-Test Problem 12-2. The Fertilizer and Corn Yield Data, displayed as Table 12-2, is used in a study of the relationship between amounts of a fertilizer (pounds per acre) and corn yield (bushels per acre), with regard to the prediction of corn yield from amount of fertilizer.

- Identify the response variable Y and the explanatory variable X .
- Verify that the least squares line to predict corn yield (crn) from fertilizer amount (frt) is $crn = 10.290 + 1.106(frtr)$. (A programmable calculator or some appropriate statistical software will be helpful.)
- Find the five-number summary for the residuals, find the interquartile range for the residuals, and construct a modified box plot of the residuals.
- Do there appear to be any candidates for outliers? Why or why not?
- Construct a scatter plot of the data, and graph the least squares line on the scatter plot.
- From the scatter plot constructed in part (e), decide whether or not you think the residual plot will look random.
- Construct a residual plot.
- Decide whether or not the linearity assumption appears to be reasonable, and state why or not.

Answers to Self-Test Problems

- 12-1** (a) See Table 12-3. (b) The five-number summary is $-16, -5, 0, +4, +16$; the interquartile range is $4 - (-5) = 9$; Figure 12-5 is a box plot of the residuals. (c) The first quartile exceeds the smallest residual of -16 by 11 , and the largest residual of $+16$ exceeds the third quartile by 12 . Since neither 11 nor 12 is greater than $(1.5)(9) = 12$, there do not appear to be any candidates for outliers. (d) See Figure 12-6. (e) There does not appear to be any identifiable nonrandom pattern among the residuals in Figure 12-6; consequently, we have no reason to doubt the linearity assumption.
- 12-2** (a) Corn yield is the response variable and amount of fertilizer is the explanatory variable. (c) The residuals are $-6.35, -2.35, -10.41, +0.59, -7.41, -1.41, +1.59, +13.53, -1.47, +2.53, +6.53, +0.53, +3.47, +6.47, +0.47, +7.47, +8.47, +1.41, -5.59, -1.59, +3.41, -3.59, -4.65, -11.65$; the five-number summary is $-11.65, -4.12, +0.50, +3.44, +13.53$; the interquartile range is $+3.44 - (-4.125) = 7.56$; Figure 12-7 is a box plot of the residuals. (d) The first quartile exceeds the smallest residual of -11.65 by 7.53 , and the largest residual of $+13.53$ exceeds the third quartile by 10.09 . Since neither 7.53 nor 10.09 is greater than $(1.5)(7.56) = 11.34$, there do not appear to be any candidates for outliers. (e) See Figure 12-8. (f) Since the scatter plot appears to show a somewhat nonlinear trend, the residual plot will probably not look random. (g) See Figure 12-9. (h) The residuals tend to be negative for the lowest amounts of fertilizer used and for the largest amounts of fertilizer used, but the residuals tend to be positive for amounts of fertilizer in the middle of the range; this nonrandom pattern in the residuals suggests that the linearity assumption may not be reasonable.

Figure 12-5
Box Plot of Residuals from the Data of Table 10-4

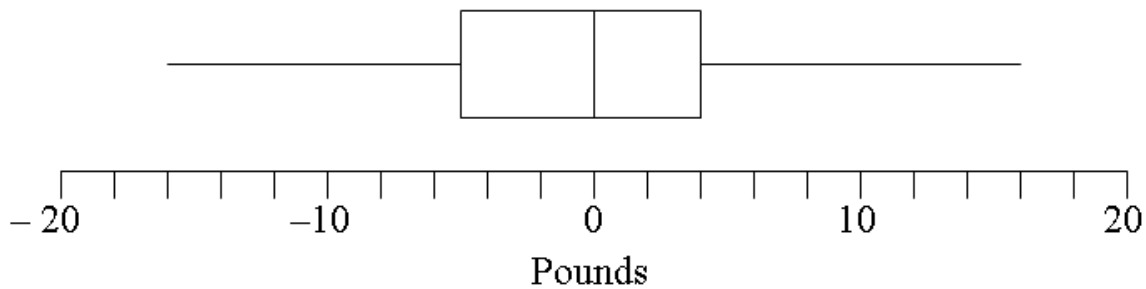


Table 12-3
Residuals with the Data of
Table 10-4 where $Y = \text{"Grip Strength"}$ and $X = \text{"Age"}$

<i>x</i>	<i>y</i>	<i>predicted y</i>	<i>residual obs - pre</i>
15	50	$26 + 2(15) = 56$	-6
17	54	$26 + 2(17) = 60$	-6
19	66	$26 + 2(19) = 64$	+2
11	46	$26 + 2(11) = 48$	-2
16	58	$26 + 2(16) = 58$	0
22	54	$26 + 2(22) = 70$	-16
17	64	$26 + 2(17) = 60$	+4
25	80	$26 + 2(25) = 76$	+4
12	46	$26 + 2(12) = 50$	-4
14	70	$26 + 2(14) = 54$	+16
25	76	$26 + 2(25) = 76$	0
23	80	$26 + 2(23) = 72$	+8

Figure 12-6

Residual Plot for the Data of Table 10-4

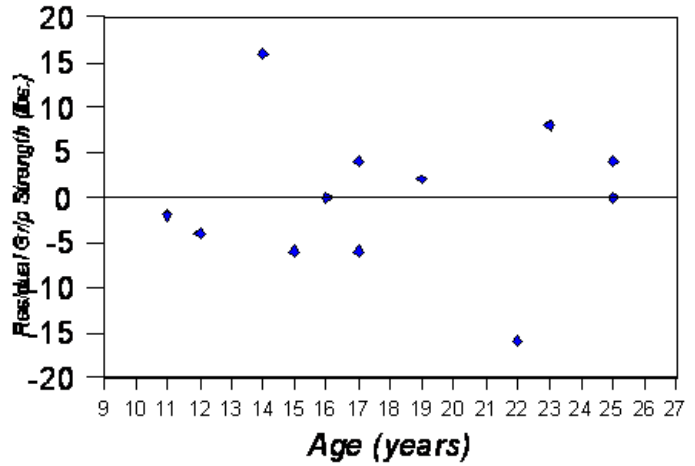


Figure 12-7

Box Plot of Residuals from the Data of Table 12-2

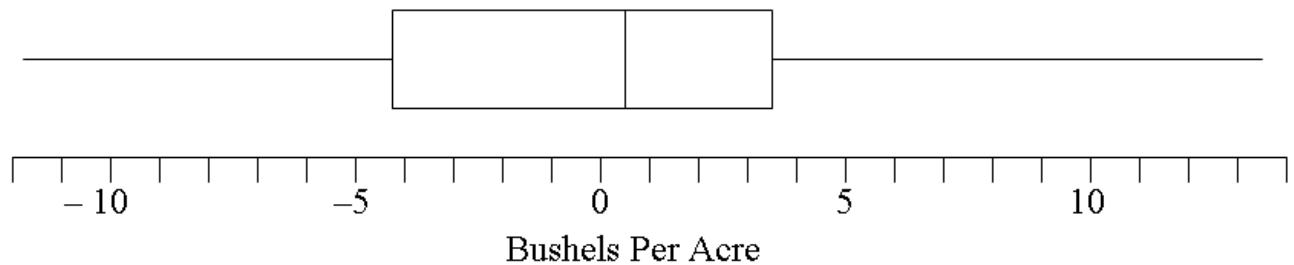


Figure 12-8

Scatterplot & Least Squares Line for Table 12-2

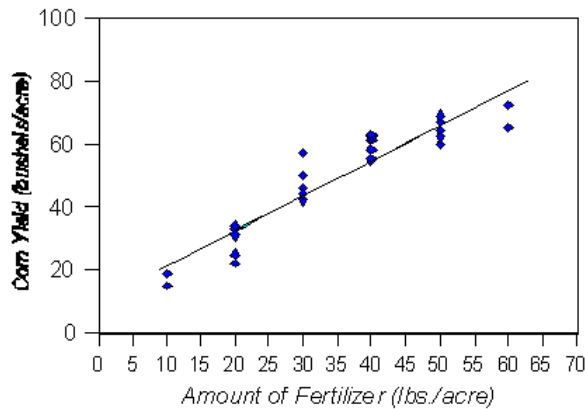
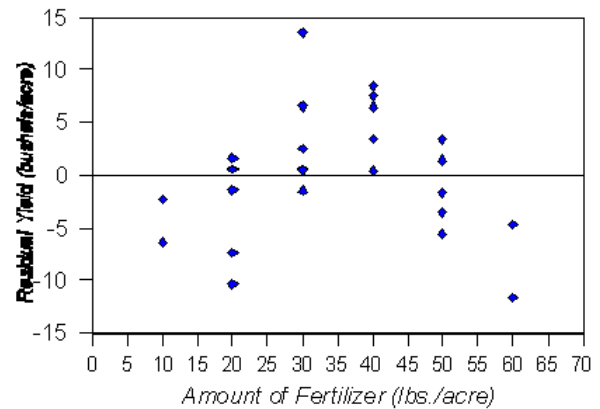


Figure 12-9

Residual Plot for Table 12-2



Summary

When we use a least squares line to make predictions, we are accepting the *linearity assumption*, which is the assumption that the relationship between two quantitative variables is a linear one. In attempting to decide whether or not the linearity assumption is reasonable, it can also be helpful to examine *residuals*. A residual is an observed value of the response variable Y minus the corresponding predicted value of Y from the least squares line; that is,

$$\text{residual} = \text{observed } Y - \text{predicted } Y.$$

Examination of the distribution of residuals indicates to us whether or not there are any data points that might be potential outliers. In order to help us decide whether or not the linearity assumption is reasonable, we can look at a *residual plot*. A residual plot is constructed the same way as a scatter plot, except that the vertical axis represents the residuals instead of the Y variable, and it is customary draw a horizontal line across the plot at zero on the vertical axis, since the residuals always sum to zero. If the linearity assumption is reasonable, then we would expect to see a residual plot which looks like a random scattering of dots (since the data points would vary randomly around the least squares line). If the linearity assumption is not reasonable, then we would expect to see a residual plot which does not look random (since the data points would vary around the least squares line in a pattern suggesting that some type of curve is more appropriate than a straight line).