RESIDUAL PLOTS AND CORRELATION 6.2.1, 6.2.2, and 6.2.4

A **residual plot** is created in order to analyze the appropriateness of a best-fit model. See the Math Notes box in Lesson 6.2.3 for more on residual plots.

The **correlation coefficient**, r, is a measure of how much or how little data is scattered around the LSRL; it is a measure of the strength of a linear association. R-squared can be interpreted as the percentage of the change in the dependent variable (y) which is accounted for or explained by the change in the independent variable (x). See the Math Notes box in Lesson 6.2.4.

For additional information, see the narrative in Checkpoint 8 at the back of the textbook. For additional examples and more practice on the topics from this chapter, see the Checkpoint 8 problems.

CALCULATORS

Statistical calculators or software can make statistical computations with ease. See details about using calculators in the Introduction section of this *Parent Guide with Extra Practice*. See the previous section for additional information about using a TI-83/84+ calculator. More detailed instructions for using a TI-83+/84+ calculator can be found under "Technology Resources" at www.cpm.org.

When you calculate the LSRL, the calculator reports the correlation coefficient on the same screen as it reports the slope and y-intercept. If your TI calculator does not calculate r, press [2nd] [CATALOG] "DiagnosticOn" [ENTER] [ENTER] and try again.

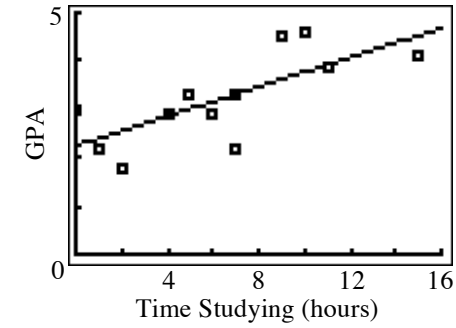
Example

It seems reasonable that there would be a relationship between the amount of time a student spends studying and their GPA. Suppose you were interested in predicting a student's GPA based on the hours they study per week. You were able to randomly select 12 students and obtain this information from each student.

Hours	4	5	11	1	15	2	10	6	7	0	7	9
GPA	2.9	3.3	3.9	2.2	4.1	1.8	4.6	2.9	2.2	3	3.3	4.5

a. Graph the data and the LSRL.

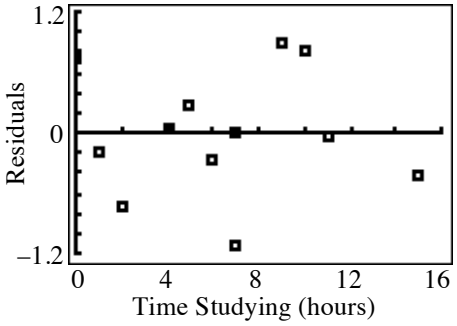
See scatterplot at right.



b. Create and interpret a residual plot.

See residual plot below right.

There is no apparent pattern in the residual plot—it looks like randomly scattered points—which means a linear model (instead of a curve) is the most appropriate way to model the relationship.



c. Find the coefficient of correlation r and R-squared. Interpret the meaning of R-squared in context.

From the calculator, $r \approx 0.7338$, R-squared $\approx 53.8\%$. About 54% of the variability in GPA can be explained by a linear relationship with study hours per week.

d. Fully describe the association between GPA and hours studied.

For a description of how to fully describe an association, see the narrative in Checkpoint 8 at the back of the textbook. *When describing an association, the form direction, strength, and outliers should be described.*

From the calculator, the least squares regression line is $y = 2.249 + 0.152x$, where y is the predicted GPA, and x is the number of hours studied.

The *form* in the scatterplot appears to be linear; it does not appear to be curved nor simply a collection of randomly scattered points. The residual plot shows random scatter, confirming that a linear model was an appropriate choice.

The *direction* is positive; as we find students who study more, we also find higher GPAs. From the slope, a student's GPA is expected to increase by 0.15 points for every additional hour of studying per week.

Example continues on next page →

Example continued from previous page.

The *strength* is moderate: it is strong enough to easily see its form, but there is scatter about the line. The correlation coefficient is 0.73, confirming a moderately strong linear association. From *R*-squared, 54% of the variability in GPA can be explained by the variability in study hours per week.

There are no apparent *outliers*.

Problems

1. It seems reasonable that the horsepower of a car is related to its gas mileage. Suppose a random sample of 9 car models is selected and the engine horsepower and city gas mileage is recorded for each one.

HP	197	170	166	230	381	170	326	451	290
MPG	16	24	19	15	13	21	11	10	15

   a. Graph the data and the LSRL.

   b. Create and interpret a residual plot.

   c. Find the coefficient of correlation *r* and *R*-squared. Interpret the meaning of *R*-squared in context.

   d. Fully describe the association.

2. Many people believe that students who are strong in music are also strong in mathematics. But the principal at University High School wonders if that same connection exists between music students and English students. The principal went through the records for the past year and found 10 students who were enrolled in both Advanced Placement Music and Advanced Placement English. He compared their final exam scores.

Final Exam Scores	
AP Music	AP English
88	63
74	96
82	86
64	90
97	68
90	90
82	78
72	74
78	96
62	79
checksum 789	*checksum 820*

   a. Graph the data and the LSRL.

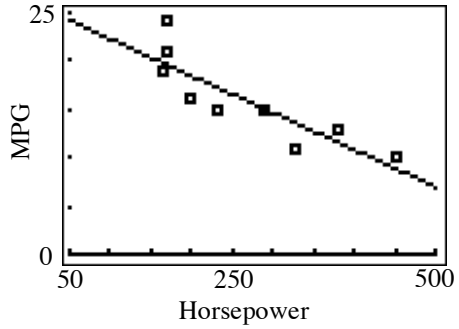
   b. Create and interpret a residual plot.

   c. Find the coefficient of correlation *r* and *R*-squared. Interpret the meaning of *R*-squared in context.

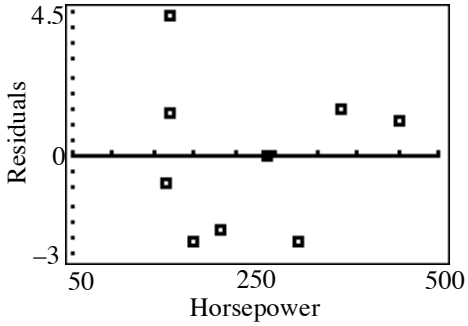
   d. Fully describe the association.

Answers

1. a. See scatterplot at right.



b. See residual plot below right.

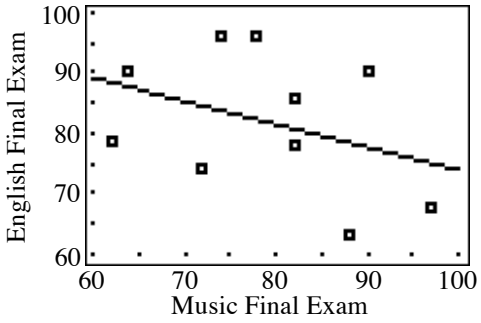


There is an apparent U-shaped pattern in the residual plot, which suggests a curved model (instead of a line) would be the most appropriate way to model the relationship. Nonetheless, by visual examination of the scatterplot and LSRL, a linear model is not completely inappropriate. We will continue with the analysis of the linear model.

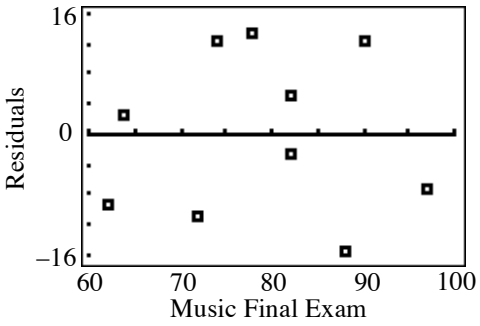
c. $r \approx -0.8602$, R-squared ≈ 0.7400. About 74% of the variability in gasoline mileage can be explained by a linear relationship with horsepower.

d. From the calculator, the least squares regression line is $y = 26.11 - 0.0382x$, where y is the predicted mileage (in mpg) and x is the horsepower. The residual plot indicates a curved model might fit the data better, but by observation of the scatterplot, the LSRL models the data well enough to proceed. The association is negative: the model predicts that for every increase of one horsepower, the mileage is expected to decrease by 0.04 mpg. The association is fairly strong with a correlation coefficient of nearly –0.9. About 74% of the variability in gasoline mileage can be explained by a linear relationship with horsepower. There are no apparent outliers. (However, if you have been following this problem since Lesson 6.1.1 in this *Parent Guide with Extra Practice*, you will recall that the data set initially had an apparent outlier at 438 horsepower and 20 mpg. That point was not considered in creating the best-fit line, and subsequently the data point was dropped from the analysis. If that data point is included, the correlation coefficient will be closer to 0.)

2. a. See scatterplot at right.



b. See residual plot below right.



There is no apparent pattern in the residual plot—it looks like randomly scattered points—which means a linear model (instead of a curve) is the most appropriate way to model the relationship.

c. From the calculator, $r \approx -0.3733$, R-squared $\approx 13.9\%$. About 14% of the variability in AP English scores can be explained by a linear relationship with AP Music scores.

d. The LSRL is $y = 112.1 - 0.3813x$, where y is the predicted AP English score, and x is the AP Music score. The residual plot confirms that a linear model is appropriate. The association is negative: the model predicts that for every increase of one in the AP Music score, the AP English score will drop by 0.38 points. The association is very weak, with a correlation coefficient of approximately –0.37. Since R-squared is approximately 14%, about 14% of the variability in AP English scores can be explained by a linear relationship with AP Music scores. There are no apparent outliers in the data collected.