Chapter 9 – Finding Equations of Linear Models
Section 3 – Linear Regression Model

Objectives
1. Compute and interpret residuals.
2. Compute and interpret the sum of squared residuals.
3. Find an equation of a linear regression model and use it to make predictions.
4. Interpret residual plots.
5. Use a residual plot to help determine whether a regression line is an appropriate model.
6. Identify influential points.
7. Compute and interpret the coefficient of determination.

Vocabulary
1. observed/predicted value
2. residual
3. sum of squared residuals
4. linear regression equation/function
5. regression line
6. influential point
7. coefficient of determination

Lesson/Activity
OBJECTIVE 1 − Compute and interpret residuals.

1. The ages and asking prices for 8 Honda Accords at dealerships in the Boston area are shown in the following table.

Age (in years)	Price (thousands of dollars)
8	13.0
7	9.2
12	7.0
9	9.9
5	13.0
11	10.0
3	17.0
5	14.2

Source: Edmunds.com

Let **x** be the age (in years) and **y** be the price (in thousands of dollars), both for a Honda Accord.

a. Describe the four characteristics of the association.
   Compute and interpret r as part of your analysis.
b. By viewing a scatterplot, we can determine that the line that contains the data points (9, 9.9) and (5, 13.0) comes close to the data points. By using the technique discussed in Section 9.2, we can find that an equation of the line is y = -0.78x + 16.88 (try it).
   Verify that the line comes close to the data points.
c. Use the model to predict the price of an 8-year-old Honda Accord.
d. Find the difference of the actual price and the predicted price for an 8-year-old Honda Accord.
e. Find the difference of the actual price and the predicted price for a 7-year-old Honda Accord.

For a data point (x, y), the **observed value** of y is y and the **predicted value** of y ($\hat{y}$ or $\widehat{f(x)}$ is the value obtained by using a model to predict y.

Definition: Residual

For a given data point (x, y), the **residual** is the difference of the observed value of y and the predicted value of y:

Residual = Observed value of y – Predicted value of y = y – $\hat{y}$

Residuals for Data Points Above, Below, or on a Line

Suppose some data points are modeled by a line.
A data point on the line has residual equal to 0.
A data point above the line has positive residual.
A data point below the line has negative residual.

OBJECTIVE 2 − Compute and interpret the sum of squared residuals.

Sum of Squared Residuals

We measure how well a line fits some data points by calculating the sum of the squared residuals:

$$\sum (y_i - \hat{y}_i)^2$$

The smaller the sum of squared residuals, the better the line will fit the data.
If the sum of squared residuals is 0, then there is an exact linear association.

2. Here we continue to work with the Honda Accord data (see the following table). Let **x** be the age (in years) and **y** be the price (in thousands of dollars), both for a Honda Accord. Find the sum of squared residuals for the linear model we worked with in Example 1, which is $\hat{y}$ = -0.78x + 16.88.

x	y	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$
8	13.0	10.64	2.36	5.57
7	9.2	11.42	-2.22	4.93
12	7.0	7.52	-0.52	0.27
9	9.9	9.86	0.04	0.00
5	13.0	12.98	0.02	0.00
11	10.0	8.3	1.7	2.89
3	17.0	14.54	2.46	6.05
5	14.2	12.98	1.22	1.49

Source: Edmunds.com Sum of Squares = 21.20

OBJECTIVE 3 − Find an equation of a linear regression model and use it to make predictions.

Definition: Linear regression function, line, equation, and model

For a group of points, the **linear regression function** is the linear function with the least sum of squared residuals. Its graph is called the **regression line** and its equation is called the **linear regression equation**, written $\hat{y} = b_1x + b_0$ where b_1 is the slope and $(0, b_0)$ is the **y**-intercept. The **linear regression model** is the linear regression function for a group of data points.

3. a. Use technology to find the linear regression equation for the Honda Accord association.
   b. Predict the price of a 10-year-old Honda Accord.
   c. What is the slope? What does it mean in this situation?
   d. What is the y-intercept? What does it mean in this situation?

Let **x** be the age (in years) and **y** be the price (in thousands of dollars), both for a Honda Accord. Find the sum of squared residuals for the linear regression model $\hat{y}$ = -0.90x + 18.42.

x	y	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$
8	13.0	11.22	1.78	3.17
7	9.2	12.12	-2.92	8.53
12	7.0	7.62	-0.62	0.38
9	9.9	10.32	-0.42	0.18
5	13.0	13.92	-0.92	0.85
11	10.0	8.52	1.48	2.19
3	17.0	15.72	1.28	1.64
5	14.2	13.92	0.28	0.08

Source: Edmunds.com Sum of Squares = 17.02

OBJECTIVE 4 – Interpret residual plots.

A **residual plot** is a graph that compares data values of the explanatory variable with the data points' residuals.

4. a. Construct a residual plot for the Honda Accord linear regression model.
   b. How many dots are above the zero residual line of the residual plot? What do they mean in this situation?
   c. How many points are below the zero residual line of the residual plot? What do they mean in this situation?
   d. Which point is farthest from the zero residual line? What does that mean in this situation?
   e. Does the residual plot support that the regression line is a reasonable model? Explain.

OBJECTIVE 5 – Use a residual plot to help determine whether a regression line is an appropriate model.

Using a Residual Plot to Help Determine Whether a Regression Line Is an Appropriate Model

The following statements apply to a residual plot for a regression line.

- If the residual plot has a pattern where the dots do not lie close to the zero residual line, then there is either a nonlinear association between the explanatory and response variables or there is no association.
- If a dot lies much farther away from the zero residual line than most or all the other dots, then the dot corresponds to an outlier. If the outlier is neither adjusted nor removed, the regression line may not be an appropriate model.
- The **vertical** spread of the residual plot should be about the same for each value of the explanatory variable.

OBJECTIVE 6 – Identify influential points.

If the slope of a regression line is greatly affected by the removal of a data point, we say the data point is an **influential point**.

5. a. The scatterplot and the regression line in Fig. 2.3 compare the hand lengths and heights of 75 women. The residual plot of the observations is shown in Fig. 2.4. Identify all outliers for the given situation.

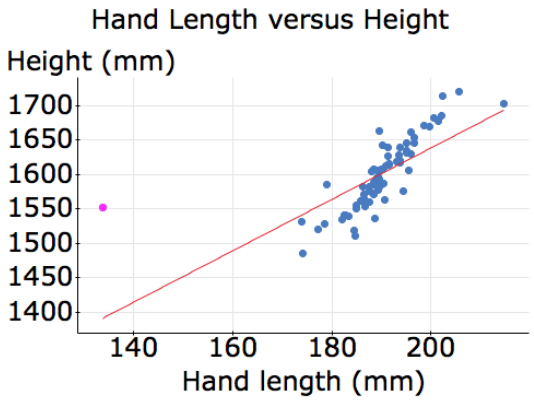

Figure 2.3: Scatterplot for hand-height data (Source: S.G. Sani, E.D. Kizilkanat, N. Boyan, et al. (2005). "Stature Estimation Based on Hand Length and Foot Length," Clinical Anatomy, Vol. 18, pp. 589-596.)

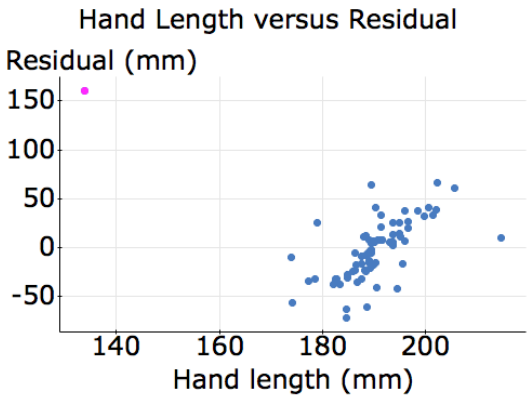

Figure 2.4: Residual plot for hand-height data

b. The outlier you likely found in Part (a) has been removed and the remaining data points are described by the scatterplot in Fig. 2.5. Is the removed outlier an influential point? Explain.

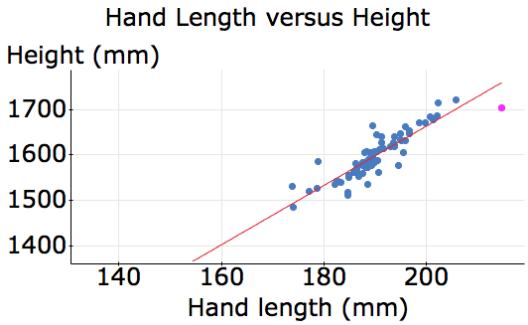

Figure 2.5: Scatterplot with outlier removed

6. a. The lengths of cruise ships and the sizes of the crews are compared by the scatterplot and the regression line shown in Fig. 2.6. The residual plot of the observations is shown in Fig. 2.7. Identify all outliers for the given situation.

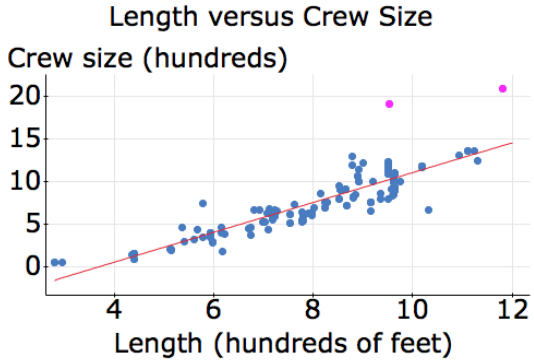

Figure 2.6: Scatterplot for cruise data (Source: True Cruise)

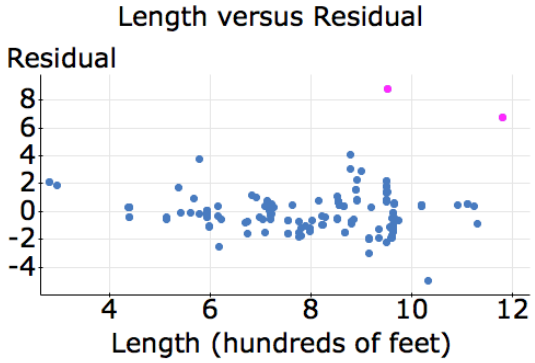

Figure 2.7: Residual plot for cruise data

b. The outliers you likely found in Part (a) have been removed and the remaining data points are described by the scatterplot in Fig. 2.8. Are the outliers that have been removed influential points? Explain.

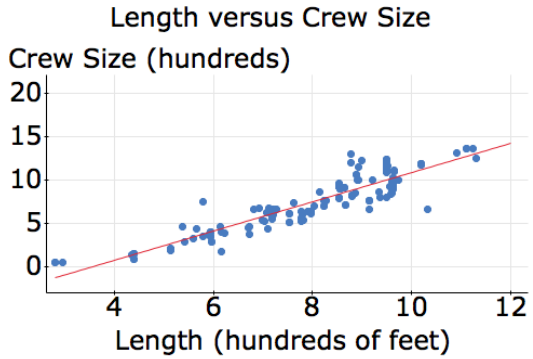

Figure 2.8: Scatterplot with outliers removed

Outliers tend to be influential points when they are horizontally far from the other data points.

OBJECTIVE 7 – Compute and interpret the coefficient of determination.

Coefficient of Determination

The coefficient of determination, r^2, is the proportion of the variation in the response variable that is explained by the regression line.

7. The temperatures and relative humidities in Phoenix, Arizona, are shown in the following table for various times on June 1, 2014.

Temperature (Fahrenheit degrees)	Relative Humidity (percent)	Temperature (Fahrenheit degrees)	Relative Humidity (percent)
89.06	11	102.92	7
84.92	14	100.94	7
82.04	16	105.98	5
80.06	20	104	5
78.98	22	102.92	5
80.96	21	102.2	6
84.02	19	100.4	5
87.98	16	98.06	6
93.02	12	91.94	9
98.06	8	91.94	9
100.04	8	87.98	3

Source: Weatherbase

a. Construct a scatterplot.
b. Find the linear regression model. Graph it on the scatterplot.
c. Compute the coefficient of determination. What does it mean in this situation?

Homework/Assessment

1, 3, 5, 9, 13, 17, 21, 23, 25, 29, 33, 35, 43, 45, 47, 49, 51, 61