The graph for the prediction interval shown in Figure 13.3.3 has drastically wider confidence bands around the regression line than the graph for the confidence interval for the mean value in Figure 13.3.2. A high price has been paid in order to account for individual variability.

Using the model for prediction outside the range of the x-values used to create the model can be very inaccurate. The nature of the relationship may not be linear outside of the range of the x's used to define the model. In the Jeep Cherokee example, the range of x-values spans from 1 to 6 years. Notice in Figure 13.3.3 that as you approach the edges of the data, the confidence interval widens somewhat. Using the model to predict the price of a 10-year-old Jeep Cherokee would no doubt have sizable error. Inferential methods are not valid outside the range of the data used to estimate the model.

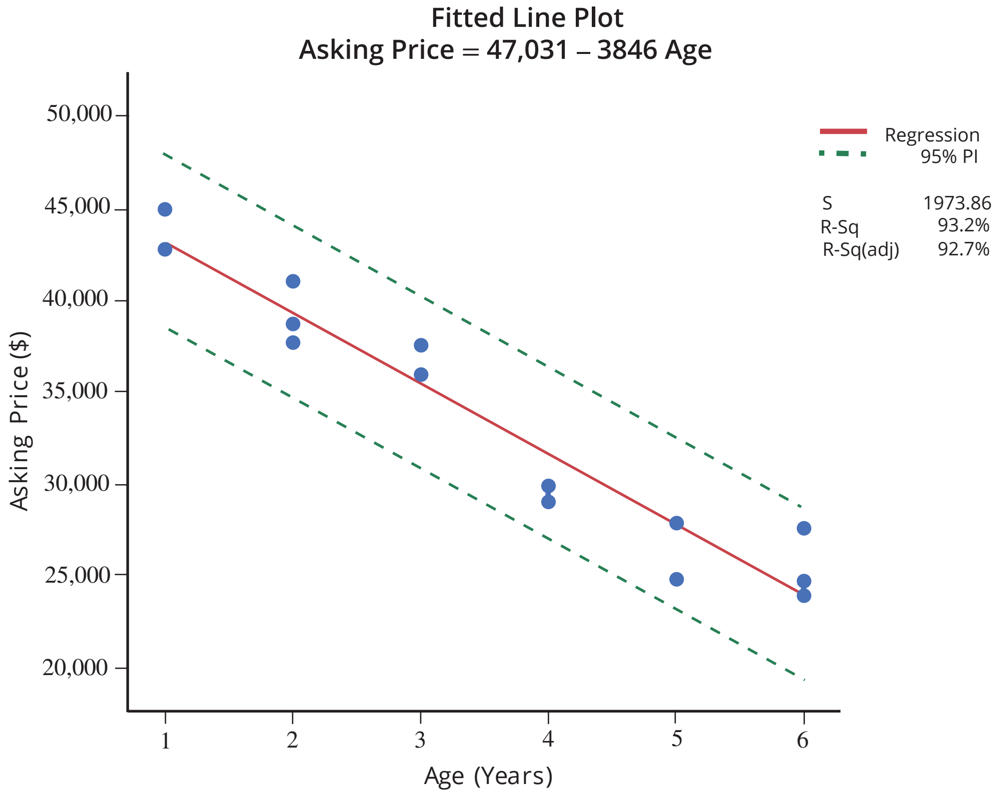

Figure 13.3.3

Technology

In order to graph the regression line, along with the prediction interval bands, using Minitab, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Regression > Linear Regression Fitted Line Plot with Prediction Interval**.

13.3 Exercises

Basic Concepts

1. Describe the difference in the interpretation of confidence intervals for the mean value of y given x and the predicted value of y given x.
2. Given a confidence interval and a prediction interval, which interval is wider? Explain why.
3. Why should you be cautious when using a regression model to predict outside the range of the x-values used to create the model?

Exercises

4. In Nevada, many forms of gambling are legal and very profitable. Sports betting amounts to billions of dollars annually. In football, a customer will bet on one of the teams to win the contest. However, in an attempt to even the game (from a betting point of view) one of the teams is selected as the favorite. The favorite's score in the game is reduced by an amount called the line. For example, if the Cowboys are favored over the Falcons by four points, then four points are subtracted from the Cowboys' score to determine the outcome of the game for betting purposes. Thus, if the Cowboys defeat the Falcons 32 to 30, in so far as settling any bets, the Cowboys score would be reduced by the spread and the Cowboys would be the loser 32 – 4 = 28 to 30. Where does the betting line come from? The line is created by a betting market. If too many people are betting on the Cowboys before the game starts, the bookmaker will try to make the game more attractive to potential Falcon bettors by increasing the spread say from four points to five points. On the other hand, if too many people are betting on the Falcons, the spread will diminish from four to perhaps three points. How accurate is the betting spread at predicting the actual spread, which is the actual difference in points between the favorite and the underdog? In the example of the Cowboys and the Falcons, the actual spread was +2 (32 – 30). To examine this question, we want to build the following model:

$$\textit{Actual Point Spread} = \beta_0 + \beta_1 \textit{Betting Spread} + \varepsilon_i.$$

If the betting spread is a good predictor of the actual spread, it should be able to account for a substantial portion of the variation in the actual spreads. The following table contains betting and actual spreads from 15 randomly selected football games.

Betting vs. Actual Spreads

Betting	4	1	3	2	1	2	5	5	3	4	2	3	5	7	6
Actual	12	−2	6	7	3	1	14	3	−7	5	14	9	2	21	8

a. Draw a scatterplot of the data. Describe the relationship you observe between actual point spread and the betting spread.

b. Estimate the parameters of the model using statistical software.

c. Is there evidence at the 0.05 level of a linear relationship between the betting spread and the actual spread?

d. What proportion of the variation in the actual point spread is explained by the betting spread?

e. Interpret the coefficient of the betting spread in the model (β_1).

f. Construct and interpret a 95% confidence interval for β_1.

g. If the betting spread is five, what is the predicted actual spread?

h. Construct and interpret a 95% prediction interval for a betting spread of five.

i. Construct a 95% confidence interval for the average value of the actual spread when the betting spread is five.

5. Net income is the level of actual profit that a company reports for the year. Net sales is the total sales less adjustment for returns. What is the relationship between net income and net sales for large corporations? Suppose a random sample of 27

large corporations has been selected, and the net income and net sales have been recorded. A regression analysis has been performed to estimate the model, and the output is given.

$$Net\ Income = \beta_0 + \beta_1 Net\ Sales + \varepsilon_i$$

Regression Analysis: Income versus Sales

The regression equation is
Income = 84 + 18.4 Sales

Predictor	Coef	SE Coef	T	P
Constant	83.6	118.1	0.71	0.486
Sales	18.434	4.446	4.15	0.000

S = 372.478 R-Sq = 40.7% R-Sq(adj) = 38.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2384660	2384660	17.19	0.000
Residual Error	25	3468497	138740		
Total	26	5853157			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	1005.3	147.1	(702.4, 1308.2)	(180.5, 1830.0)

Values of Predictors for New Observations

New Obs	Sales
1	50.0

a. Find and interpret the standard deviation of the error terms in the output.

b. Interpret the slope coefficient. (The data used to estimate the model was in millions of dollars.)

c. What proportion of the variation in net income is explained by net sales?

d. Is there evidence of a linear relationship between net income and net sales? Test at the 0.05 level.

e. Construct and interpret a 95% confidence interval for β_1, the slope of the line.

f. The output also contains a predicted value for net income when sales are \$50,000,000. Find the predicted value of net income when sales are \$50,000,000. (Note that in the original data all observations were measured in millions of dollars. Thus, a predicted value of 10,000,000 would be displayed in the output as 10.)

g. Find and interpret the 95% confidence interval for the average value of net income given that sales are \$50,000,000.

h. Suppose your firm generated \$50,000,000 in sales. What would be the 95% prediction interval for your firm's net income?

i. Use the model to predict net income for a company with \$60,000,000 in sales. (Note that you must compute this manually.)

6. The personnel director of a large hospital is interested in determining the relationship (if any) between an employee's age and the number of sick days the employee takes per year. The director randomly selects eight employees and records their age and the number of sick days which they took in the previous year.

Sick Days and Age								
Employee	1	2	3	4	5	6	7	8
Age	30	50	40	55	30	28	60	25
Sick Days	7	4	3	2	9	10	0	8

A regression analysis has been performed to estimate the model and the output is given.

$$\text{Sick Days} = \beta_0 + \beta_1 \text{Age} + \varepsilon_i$$

Regression Analysis: Sick Days versus Age

The regression equation is
Sick Days = 15.2 - 0.247 Age

Predictor	Coef	SE Coef	T	P
Constant	15.186	1.713	8.86	0.000
Age	-0.24681	0.04105	-6.01	0.001

S = 1.47652 R-Sq = 85.8% R-Sq(adj) = 83.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	78.794	78.794	36.14	0.001
Residual Error	6	13.081	2.180		
Total	7	91.875			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	6.547	0.557	(5.184, 7.911)	(2.686, 10.409)

Values of Predictors for New Observations

New Obs	Age
1	35.0

a. Draw a scatterplot of the data. Describe the relationship you observe between the number of sick days and age.

b. Find and interpret the standard deviation of the error terms in the output.

c. Interpret the slope coefficient.

d. What proportion of the variation in the number of sick days an employee takes per year is explained by age?

e. Is there evidence of a linear relationship between the number of sick days an employee takes per year and age? Test at the 0.05 level.

f. Construct and interpret a 95% confidence interval for β_1, the slope of the line.

g. Find the predicted value of the number of sick days an employee will take per year if the employee is 35 years old.

h. Find and interpret the 95% confidence interval for the average number of sick days an employee will take per year, given the employee is 35.

i. Suppose a new employee is 35. Find a 95% prediction interval for the number of sick days this employee will take this year.

j. Use the model to predict the number of sick days per year for an employee who is 45 years old. Round to the nearest whole number.

7. A manufacturing company that produces automobiles is interested in studying the relationship between the number of hours of training that an automotive painter receives and the number of paint defects per auto produced. Ten employees are randomly selected. The number of hours of training each automotive painter has received is recorded and the number of paint defects on the most recent auto that they painted is determined. The results are as follows.

Training Hours and Paint Defects										
Hours of Training	1	4	7	3	2	2	5	5	1	6
Defects per Auto	1	4	0	3	5	4	3	2	5	1

A regression analysis has been performed to estimate the model, and the following output is produced.

$$Paint\ Defects\ per\ Auto = \beta_0 + \beta_1 Hours\ of\ Training + \varepsilon_i$$

Regression Analysis: Paint Defects per Auto versus Hours of Training

The regression equation is
Paint Defects per Auto = 4.65 - 0.515 Hours of Training

Predictor	Coef	SE Coef	T	P
Constant	4.6535	0.9426	4.94	0.001
Hours of Training	-0.5149	0.2286	-2.25	0.054

S = 1.45306   R-Sq = 38.8%   R-Sq(adj) = 31.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10.709	10.709	5.07	0.054
Residual Error	8	16.891	2.111		
Total	9	27.600			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	2.594	0.469	(1.514, 3.674)	(-0.927, 6.115)

Values of Predictors for New Observations

New Obs	Hours of Training
1	4.00

a. Draw a scatterplot of the data. Describe the relationship you observe between the number of paint defects per auto and hours of training. Are there any unusual observations?

b. Find and interpret the standard deviation of the error terms in the output.

c. Interpret the slope coefficient.

d. What proportion of the variation in the number of paint defects per auto is explained by the hours of training? What other factors might affect the number of paint defects?

e. Is there evidence of a linear relationship between the number of hours of training and the number of paint defects per auto? Test at the 0.05 level and the 0.10 level.

f. Construct and interpret a 95% confidence interval for β_1, the slope coefficient.

g. Find the predicted value of the number of paint defects per auto for an automotive painter who has had 4 hours of training.

h. Find and interpret the 95% confidence interval for the average number of paint defects per auto for automotive painters who have had 4 hours of training.

i. Suppose a new automotive painter has had 4 hours of training. What would be the 95% prediction interval for the number of paint defects per auto?

j. Use the model to predict the number of paint defects per auto for an automotive painter who has had 7 hours of training. Round your answer to the nearest whole number.

Data

The data set is available on stat.hawkeslearning.com under **Data Sets > Global Statistics by Country**.

8. Using the Global Statistics by Country data set, answer the following questions regarding birth rate and female literacy rate.

a. Find the predicted value of birth rate for a female literacy rate of 90.

b. Find and interpret the 99% confidence interval for the average birth rate for a female literacy rate of 90.

c. Find and interpret the 99% prediction interval for a female literacy rate of 90.

d. How do these two intervals compare?

CR Chapter Review

Key Terms and Ideas

- Simple Linear Regression Model
- Population Regression Line
- Sample Regression Line
- Assumptions about the Error Term
- Residual
- Confidence Interval for β_1
- Testing a Hypothesis Concerning β_1
- Mean Square Error
- Standard Error
- Confidence Interval for the Mean Value of y Given x
- Confidence Interval for the Predicted Value of y Given x

Key Formulas		
Section		
13.1	**Simple Linear Regression Model** $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	**Estimated Simple Linear Regression Equation** $\hat{y}_i = b_0 + b_1 x_i$
	Sum of Squared Errors $\text{SSE} = \sum\left(y_i - \hat{y}_i\right)^2 = \sum\left(y_i - \left(b_0 + b_1 x_i\right)\right)^2$	**Slope of the Least Squares Line** $b_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}.$