Section 3.2 Exercises

35. **What's my line?** You use the same bar of soap to shower each morning. The bar weighs 80 grams when it is new. Its weight goes down by 6 grams per day on average. What is the equation of the regression line for predicting weight from days of use?

36. **What's my line?** An eccentric professor believes that a child with IQ 100 should have a reading test score of 50 and predicts that reading score should increase by 1 point for every additional point of IQ. What is the equation of the professor's regression line for predicting reading score from IQ?

37. **Gas mileage** We expect a car's highway gas mileage to be related to its city gas mileage. Data for all 1198 vehicles in the government's recent *Fuel Economy Guide* give the regression line: predicted highway mpg = 4.62 + 1.109 (city mpg).

(a) What's the slope of this line? Interpret this value in context.

(b) What's the y intercept? Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted highway mileage for a car that gets 16 miles per gallon in the city.

38. **IQ and reading scores** Data on the IQ test scores and reading test scores for a group of fifth-grade children give the following regression line: predicted reading score = −33.4 + 0.882(IQ score).

(a) What's the slope of this line? Interpret this value in context.

(b) What's the y intercept? Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted reading score for a child with an IQ score of 90.

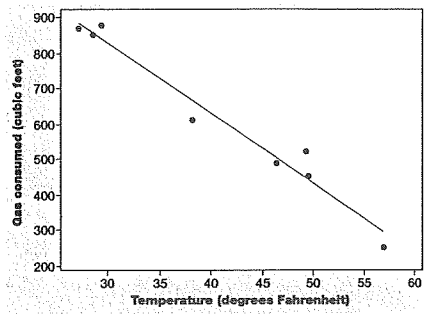
39. **Acid rain** Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH values show higher acidity. The researchers observed a linear pattern over time. They reported that the regression line $\widehat{\text{pH}} = 5.43 - 0.0053(\text{weeks})$ fit the data well.[19]

(a) Identify the slope of the line and explain what it means in this setting.

(b) Identify the y intercept of the line and explain what it means in this setting.

(c) According to the regression line, what was the pH at the end of this study?

40. **How much gas?** In Exercise 4 (page 159), we examined the relationship between the average monthly temperature and the amount of natural gas consumed in Joan's midwestern home. The figure below shows the original scatterplot with the least-squares line added. The equation of the least-squares line is $\hat{y} = 1425 - 19.87x$.



(a) Identify the slope of the line and explain what it means in this setting.

(b) Identify the y intercept of the line. Explain why it's risky to use this value as a prediction.

(c) Use the regression line to predict the amount of natural gas Joan will use in a month with an average temperature of 30°F.

41. **Acid rain** Refer to Exercise 39. Would it be appropriate to use the regression line to predict pH after 1000 months? Justify your answer.

42. **How much gas?** Refer to Exercise 40. Would it be appropriate to use the regression line to predict Joan's natural-gas consumption in a future month with an average temperature of 65°F? Justify your answer.

43. **Least-squares idea** The table below gives a small set of data. Which of the following two lines fits the data better: $\hat{y} = 1 - x$ or $\hat{y} = 3 - 2x$? Use the least-squares criterion to justify your answer. (*Note:* Neither of these two lines is the least-squares regression line for these data.)

x:	−1	1	1	3	5
y:	2	0	1	−1	−5

44. **Least-squares idea** In Exercise 40, the line drawn on the scatterplot is the least-squares regression line. Explain the meaning of the phrase "least-squares" to Joan, who knows very little about statistics.

45. **Acid rain** In the acid rain study of Exercise 39, the actual pH measurement for Week 50 was 5.08. Find and interpret the residual for this week.

46. **How much gas?** Refer to Exercise 40. During March, the average temperature was 46.4°F and Joan used 490 cubic feet of gas per day. Find and interpret the residual for this month.

47. **Bird colonies** Exercise 6 (page 159) examined the relationship between the number of new birds y and percent of returning birds x for 13 sparrowhawk colonies. Here are the data once again.

Percent return:	74	66	81	52	73	62	52	45	62	46	60	46	38
New adults:	5	6	8	11	12	15	16	17	18	18	19	20	20

(a) Use your calculator to help make a scatterplot.

(b) Use your calculator's regression function to find the equation of the least-squares regression line. Add this line to your scatterplot from (a).

(c) Explain in words what the slope of the regression line tells us.

(d) Calculate and interpret the residual for the colony that had 52% of the sparrowhawks return and 11 new adults.

48. **Do heavier people burn more energy?** Exercise 10 (page 160) presented data on the lean body mass and resting metabolic rate for 12 women who were subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate, in calories burned per 24 hours, is the rate at which the body consumes energy. Here are the data again.

Mass:	36.1	54.6	48.5	42.0	50.6	42.0	40.3	33.1	42.4	34.5	51.1	41.2
Rate:	995	1425	1396	1418	1502	1256	1189	913	1124	1052	1347	1204

(a) Use your calculator to help make a scatterplot.

(b) Use your calculator's regression function to find the equation of the least-squares regression line. Add this line to your scatterplot from part (a).

(c) Explain in words what the slope of the regression line tells us.

(d) Calculate and interpret the residual for the woman who had a lean body mass of 50.6 kg and a metabolic rate of 1502.

49. **Bird colonies** Refer to Exercise 47.

(a) Use your calculator to make a residual plot. Describe what this graph tells you about the appropriateness of using a linear model.

(b) Which point has the largest residual? Explain what this residual means in context.

50. **Do heavier people burn more energy?** Refer to Exercise 48.

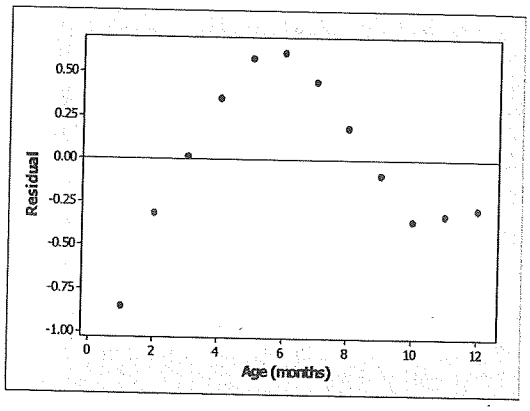
(a) Use your calculator to make a residual plot. Describe what this graph tells you about the appropriateness of using a linear model.

(b) Which point has the largest residual? Explain what the value of that residual means in context.

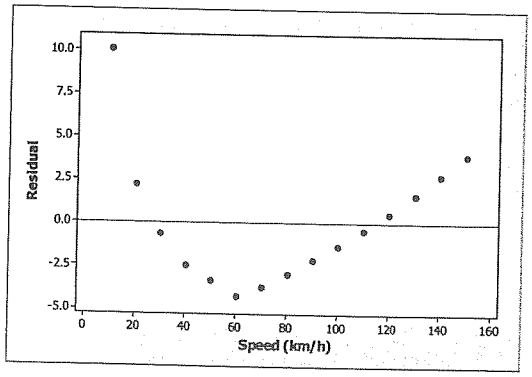
51. **Nahya infant weights** A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Here are the mean weights (in kilograms) for 170 infants in Nahya who were weighed each month during their first year of life:

Age (months):	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg):	4.3	5.1	5.7	6.3	6.8	7.1	7.2	7.2	7.2	7.2	7.5	7.8

A hasty user of statistics enters the data into software and computes the least-squares line without plotting the data. The result is $\widehat{\text{weight}} = 4.88 + 0.267\,(\text{age})$. A residual plot is shown below. Would it be appropriate to use this regression line to predict y from x? Justify your answer.

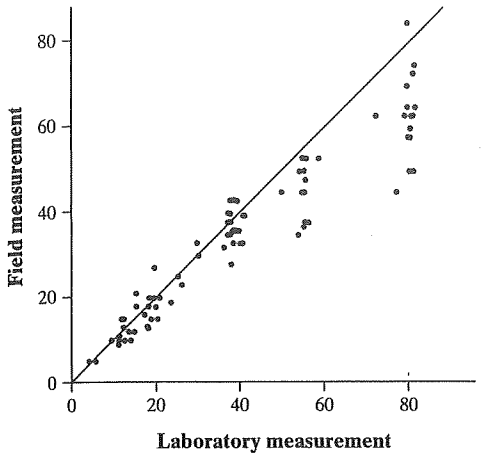


52. **Driving speed and fuel consumption** Exercise 9 (page 160) gives data on the fuel consumption y of a car at various speeds x. Fuel consumption is measured in liters of gasoline per 100 kilometers driven and speed is measured in kilometers per hour. A statistical software package gives the least-squares regression line and the residual plot shown below. The regression line is $\hat{y} = 11.058 - 0.01466x$. Would it be appropriate to use the regression line to predict y from x? Justify your answer.



53. **Oil and residuals** The Trans-Alaska Oil Pipeline is a tube that is formed from 1/2-inch-thick steel and that carries oil across 800 miles of sensitive arctic and subarctic terrain. The pipe segments and the welds that join them were carefully examined before installation. How accurate are field measurements of the depth of small defects? The figure below compares the results of measurements on 100 defects made in the field with

measurements of the same defects made in the laboratory.[20] The line $y = x$ is drawn on the scatterplot.

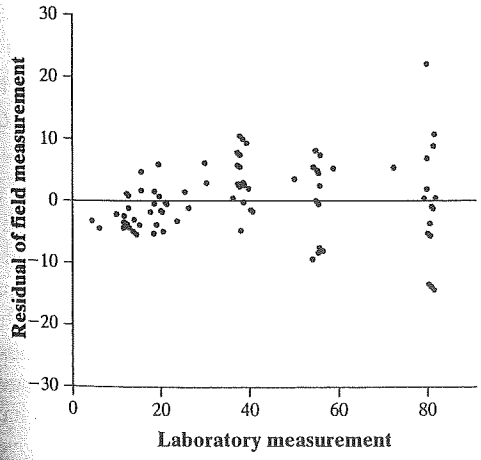


(a) Describe the overall pattern you see in the scatterplot, as well as any deviations from that pattern.

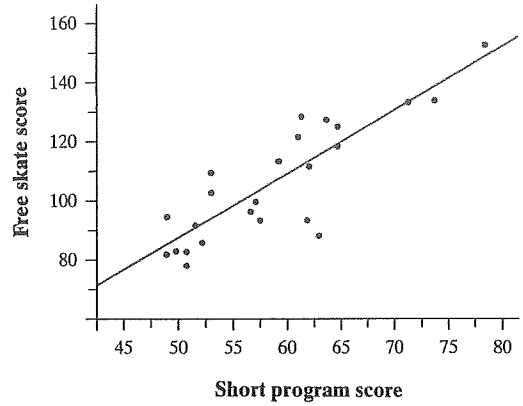
(b) If field and laboratory measurements all agree, then the points should fall on the $y = x$ line drawn on the plot, except for small variations in the measurements. Is this the case? Explain.

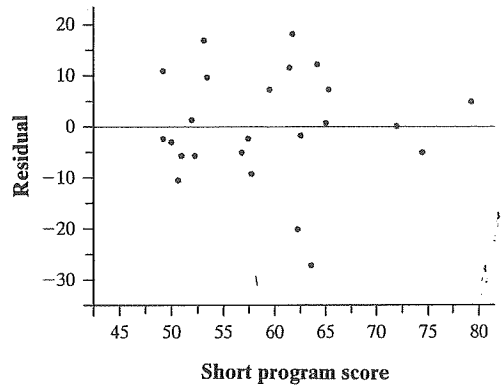
(c) The line drawn on the scatterplot ($y = x$) is *not* the least-squares regression line. How would the slope and y intercept of the least-squares line compare? Justify your answer.

54. **Oil and residuals** Refer to Exercise 53. The following figure shows a residual plot for the least-squares regression line. Discuss what the residual plot tells you about the appropriateness of using a linear model.



55. **Olympic figure skating** For many people, the women's figure skating competition is the highlight of the Olympic Winter Games. Scores in the short program x and scores in the free skate y were recorded for each of the 24 skaters who competed in both rounds during the 2010 Winter Olympics in Vancouver, Canada.[21] A regression analysis was performed using these data. The scatterplot and residual plot follow. The equation of the least-squares regression line is $\hat{y} = -16.2 + 2.07x$. Also, $s = 10.2$ and $r^2 = 0.736$.





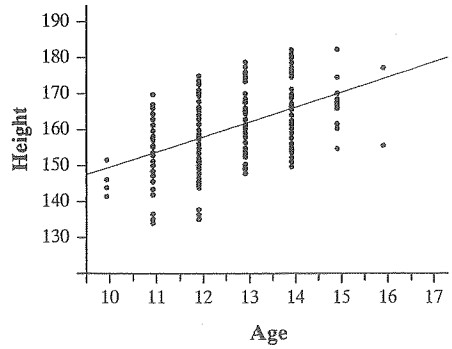
(a) Calculate and interpret the residual for the gold medal winner, Yu-Na Kim, who scored 78.50 in the short program and 150.06 in the free skate.

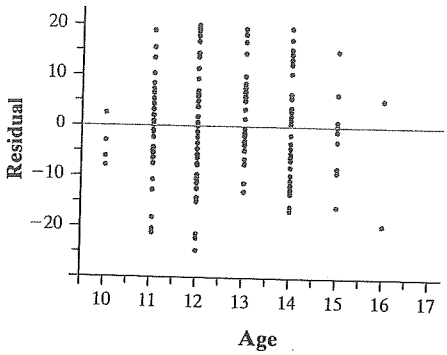
(b) Is a linear model appropriate for these data? Explain.

(c) Interpret the value of s.

(d) Interpret the value of r^2.

56. **Age and height** A random sample of 195 students was selected from the United Kingdom using the CensusAtSchool data selector. The age (in years) x and height (in centimeters) y was recorded for each of the students. A regression analysis was performed using these data. The scatterplot and residual plot are shown below. The equation of the least-squares regression line is $\hat{y} = 106.1 + 4.21x$. Also, $s = 8.61$ and $r^2 = 0.274$.





(a) Calculate and interpret the residual for the student who was 141 cm tall at age 10.

(b) Is a linear model appropriate for these data? Explain.

(c) Interpret the value of s.

(d) Interpret the value of r^2.

57. **Bird colonies** Refer to Exercises 47 and 49. For the regression you performed earlier, $r^2 = 0.56$ and $s = 3.67$. Explain what each of these values means in this setting.

58. **Do heavier people burn more energy?** Refer to Exercises 48 and 50. For the regression you performed earlier, $r^2 = 0.768$ and $s = 95.08$. Explain what each of these values means in this setting.

59. **Merlins breeding** Exercise 13 (page 160) gives data on the number of breeding pairs of merlins in an isolated area in each of seven years and the percent of males who returned the next year. The data show that the percent returning is lower after successful breeding seasons and that the relationship is roughly linear. The figure below shows Minitab regression output for these data.

pg 181

```
Session
Regression Analysis: Percent return versus Breeding pairs

Predictor        Coef   SE Coef      T      P
Constant       266.07     52.15   5.10  0.004
Breeding pairs -6.650     1.736  -3.83  0.012

S = 7.76227   R-Sq = 74.6%   R-Sq(adj) = 69.5%
```

(a) What is the equation of the least-squares regression line for predicting the percent of males that return from the number of breeding pairs? Use the equation to predict the percent of returning males after a season with 30 breeding pairs.

(b) What percent of the year-to-year variation in percent of returning males is accounted for by the straight-line relationship with number of breeding pairs the previous year?

(c) Use the information in the figure to find the correlation r between percent of males that return and number of breeding pairs. How do you know whether the sign of r is + or −?

(d) Interpret the value of s in this setting.

60. **Does social rejection hurt?** Exercise 14 (page 161) gives data from a study that shows that social exclusion causes "real pain." That is, activity in an area of the brain that responds to physical pain goes up as distress from social exclusion goes up. A scatterplot shows a moderately strong, linear relationship. The figure below shows Minitab regression output for these data.

```
Session
Regression Analysis: Brain versus Distress

Predictor        Coef    SE Coef      T      P
Constant      -0.12608   0.02465  -5.12  0.000
distress      0.060782   0.009979   6.09  0.000

S = 0.0250896   R-Sq = 77.1%   R-Sq (adj) = 75.1%
```

(a) What is the equation of the least-squares regression line for predicting brain activity from social distress score? Use the equation to predict brain activity for social distress score 2.0.

(b) What percent of the variation in brain activity among these subjects is accounted for by the straight-line relationship with social distress score?

(c) Use the information in the figure to find the correlation r between social distress score and brain activity. How do you know whether the sign of r is + or −?

(d) Interpret the value of s in this setting.

61. **Husbands and wives** The mean height of married American women in their early twenties is 64.5 inches and the standard deviation is 2.5 inches. The mean height of married men the same age is 68.5 inches, with standard deviation 2.7 inches. The correlation between the heights of husbands and wives is about $r = 0.5$.

pg 183

(a) Find the equation of the least-squares regression line for predicting a husband's height from his wife's height for married couples in their early 20s. Show your work.

(b) Suppose that the height of a randomly selected wife was 1 standard deviation below average. Predict the height of her husband *without* using the least-squares line. Show your work.

62. **The stock market** Some people think that the behavior of the stock market in January predicts its behavior for the rest of the year. Take the explanatory variable x to be the percent change in a stock market index in January and the response variable y to be the change in the index for the entire year. We expect a positive correlation between x and y because the change during January contributes to the full year's change. Calculation from data for an 18-year period gives

$$\bar{x} = 1.75\% \quad s_x = 5.36\% \quad \bar{y} = 9.07\%$$
$$s_y = 15.35\% \quad r = 0.596$$

(a) Find the equation of the least-squares line for predicting full-year change from January change. Show your work.

(b) Suppose that the percent change in a particular January was 2 standard deviations above average. Predict the percent change for the entire year, *without* using the least-squares line. Show your work.

63. **Husbands and wives** Refer to Exercise 61.

(a) Find r^2 and interpret this value in context.

(b) For these data, $s = 1.2$. Interpret this value.

64. **The stock market** Refer to Exercise 62.

(a) Find r^2 and interpret this value in context.

(b) For these data, $s = 8.3$. Interpret this value.

65. **Will I bomb the final?** We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period.[22] Assume that both the midterm and final exam were scored out of 100 points.

(a) State the equation of the least-squares regression line if each student scored the same on the midterm and the final.

(b) The actual least-squares line for predicting final-exam score y from midterm-exam score x was $\hat{y} = 46.6 + 0.41x$. Predict the score of a student who scored 50 on the midterm and a student who scored 100 on the midterm.

(c) Explain how your answers to part (b) illustrate regression to the mean.

66. **It's still early** We expect that a baseball player who has a high batting average in the first month of the season will also have a high batting average the rest of the season. Using 66 Major League Baseball players from the 2010 season,[23] a least-squares regression line was calculated to predict rest-of-season batting average y from first-month batting average x. *Note:* A player's batting average is the proportion of times at bat that he gets a hit. A batting average over 0.300 is considered very good in Major League Baseball.

(a) State the equation of the least-squares regression line if each player had the same batting average the rest of the season as he did in the first month of the season.

(b) The actual equation of the least-squares regression line is $\hat{y} = 0.245 + 0.109x$. Predict the rest-of-season batting average for a player who had a 0.200 batting average the first month of the season and for a player who had a 0.400 batting average the first month of the season.

(c) Explain how your answers to part (b) illustrate regression to the mean.

67. STEP 4 **Beavers and beetles** Do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them. If so, more stumps should produce more beetle larvae. Here are the data:[24]

Stumps:	2	2	1	3	3	4	3	1	2	5	1	3
Beetle larvae:	10	30	12	24	36	40	43	11	27	56	18	40
Stumps:	2	1	2	2	1	1	4	1	2	1	4	
Beetle larvae:	25	8	21	14	16	6	54	9	13	14	50	

Can we use a linear model to predict the number of beetle larvae from the number of stumps? If so, how accurate will our predictions be? Follow the four-step process.

68. STEP 4 **Fat and calories** The number of calories in a food item depends on many factors, including the amount of fat in the item. The data below show the amount of fat (in grams) and the number of calories in 7 beef sandwiches at McDonalds.[25]

Sandwich	Fat	Calories
Big Mac®	29	550
Quarter Pounder® with Cheese	26	520
Double Quarter Pounder® with Cheese	42	750
Hamburger	9	250
Cheeseburger	12	300
Double Cheeseburger	23	440
McDouble	19	390

Can we use a linear model to predict the number of calories from the amount of fat? If so, how accurate will our predictions be? Follow the four-step process.

69. **Managing diabetes** People with diabetes measure their fasting plasma glucose (FPG; measured in units of milligrams per milliliter) after fasting for at least 8 hours. Another measurement, made at regular medical checkups, is called HbA. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months. The table below gives data on both HbA and FPG for 18 diabetics five months after they had completed a diabetes education class.[27]

Subject	HbA (%)	FPG (mg/mL)	Subject	HbA (%)	FPG (mg/mL)
1	6.1	141	10	8.7	172
2	6.3	158	11	9.4	200
3	6.4	112	12	10.4	271
4	6.8	153	13	10.6	103
5	7.0	134	14	10.7	172
6	7.1	95	15	10.7	359
7	7.5	96	16	11.2	145
8	7.7	78	17	13.7	147
9	7.9	148	18	19.3	255

(a) Make a scatterplot with HbA as the explanatory variable. Describe what you see.

(b) Subject 18 is an outlier in the x direction. What effect do you think this subject has on the correlation? What effect do you think this subject has on the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this subject to confirm your answer.

(c) Subject 15 is an outlier in the y direction. What effect do you think this subject has on the correlation? What effect do you think this subject has on the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this subject to confirm your answer.

70. **Rushing for points** What is the relationship between rushing yards and points scored in the 2011 National Football League? The table below gives the number of rushing yards and the number of points scored for each of the 16 games played by the 2011 Jacksonville Jaguars.[26]

Game	Rushing yards	Points scored
1	163	16
2	112	3
3	128	10
4	104	10
5	96	20
6	133	13
7	132	12
8	84	14
9	141	17
10	108	10
11	105	13
12	129	14
13	116	41
14	116	14
15	113	17
16	190	19

(a) Make a scatterplot with rushing yards as the explanatory variable. Describe what you see.

(b) The number of rushing yards in Game 16 is an outlier in the x direction. What effect do you think this game has on the correlation? On the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this game to confirm your answers.

(c) The number of points scored in Game 13 is an outlier in the y direction. What effect do you think this game has on the correlation? On the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this game to confirm your answers.

Multiple choice: Select the best answer for Exercises 71 to 78.

71. Which of the following is *not* a characteristic of the least-squares regression line?

(a) The slope of the least-squares regression line is always between -1 and 1.

(b) The least-squares regression line always goes through the point $(\bar{x}, \bar{y})$.

(c) The least-squares regression line minimizes the sum of squared residuals.

(d) The slope of the least-squares regression line will always have the same sign as the correlation.

(e) The least-squares regression line is not resistant to outliers.

72. Each year, students in an elementary school take a standardized math test at the end of the school year. For a class of fourth-graders, the average score was 55.1 with a standard deviation of 12.3. In the third grade, these same students had an average score of 61.7 with a standard deviation of 14.0. The correlation between the two sets of scores is $r = 0.95$. Calculate the equation of the least-squares regression line for predicting a fourth-grade score from a third-grade score.

(a) $\hat{y} = 3.60 + 0.835x$

(b) $\hat{y} = 15.69 + 0.835x$

(c) $\hat{y} = 2.19 + 1.08x$

(d) $\hat{y} = -11.54 + 1.08x$

(e) Cannot be calculated without the data.

73. Using data from the 2009 LPGA tour, a regression analysis was performed using x = average driving distance and y = scoring average. Using the output from the regression analysis shown below, determine the equation of the least-squares regression line.

```
Predictor          Coef        SE Coef    T       P
Constant           87.974      2.391      36.78   0.000
Driving Distance   -0.060934   0.009536   -6.39   0.000

S = 1.01216    R-Sq = 22.1%    R-Sq(adj) = 21.6%
```

(a) $\hat{y} = 87.947 + 2.391x$

(b) $\hat{y} = 87.947 + 1.01216x$

(c) $\hat{y} = 87.947 - 0.060934x$

(d) $\hat{y} = -0.060934 + 1.01216x$

(e) $\hat{y} = -0.060934 + 87.947x$

Exercises 74 to 78 refer to the following setting. Measurements on young children in Mumbai, India, found this least-squares line for predicting height y from arm span x:[28]

$$\hat{y} = 6.4 + 0.93x$$

Measurements are in centimeters (cm).

74. By looking at the equation of the least-squares regression line, you can see that the correlation between height and arm span is

(a) greater than zero.

(b) less than zero.

(c) 0.93.
(d) 6.4.
(e) Can't tell without seeing the data.

75. In addition to the regression line, the report on the Mumbai measurements says that $r^2 = 0.95$. This suggests that

(a) although arm span and height are correlated, arm span does not predict height very accurately.
(b) height increases by $\sqrt{0.95} = 0.97$ cm for each additional centimeter of arm span.
(c) 95% of the relationship between height and arm span is accounted for by the regression line.
(d) 95% of the variation in height is accounted for by the regression line.
(e) 95% of the height measurements are accounted for by the regression line.

76. One child in the Mumbai study had height 59 cm and arm span 60 cm. This child's residual is

(a) −3.2 cm. (c) −1.3 cm. (e) 62.2 cm.
(b) −2.2 cm. (d) 3.2 cm.

77. Suppose that a tall child with arm span 120 cm and height 118 cm was added to the sample used in this study. What effect will adding this child have on the correlation and the slope of the least-squares regression line?

(a) Correlation will increase, slope will increase.
(b) Correlation will increase, slope will stay the same.
(c) Correlation will increase, slope will decrease.
(d) Correlation will stay the same, slope will stay the same.
(e) Correlation will stay the same, slope will increase.

78. Suppose that the measurements of arm span and height were converted from centimeters to meters by dividing each measurement by 100. How will this conversion affect the values of r^2 and s?

(a) r^2 will increase, s will increase.
(b) r^2 will increase, s will stay the same.
(c) r^2 will increase, s will decrease.
(d) r^2 will stay the same, s will stay the same.
(e) r^2 will stay the same, s will decrease.

Exercises 79 and 80 refer to the following setting.

In its recent *Fuel Economy Guide*, the Environmental Protection Agency gives data on 1152 vehicles. There are a number of outliers, mainly vehicles with very poor gas mileage. If we ignore the outliers, however, the combined city and highway gas mileage of the other 1120 or so vehicles is approximately Normal with mean 18.7 miles per gallon (mpg) and standard deviation 4.3 mpg.

79. **In my Chevrolet** (2.2) The Chevrolet Malibu with a four-cylinder engine has a combined gas mileage of 25 mpg. What percent of all vehicles have worse gas mileage than the Malibu?

80. **The top 10%** (2.2) How high must a vehicle's gas mileage be in order to fall in the top 10% of all vehicles? (The distribution omits a few high outliers, mainly hybrid gas-electric vehicles.)

81. **Marijuana and traffic accidents** (1.1) Researchers in New Zealand interviewed 907 drivers at age 21. They had data on traffic accidents and they asked the drivers about marijuana use. Here are data on the numbers of accidents caused by these drivers at age 19, broken down by marijuana use at the same age:[29]

	Marijuana use per year			
	Never	1–10 times	11–50 times	51 + times
Drivers	452	229	70	156
Accidents caused	59	36	15	50

(a) Make a graph that displays the accident rate for each class. Is there evidence of an association between marijuana use and traffic accidents?
(b) Explain why we can't conclude that marijuana use *causes* accidents.

HAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Two statistics students went to a flower shop and randomly selected 12 carnations. When they got home, the students prepared 12 identical vases with exactly the same amount of water in each vase. They put one tablespoon of sugar in 3 vases, two tablespoons of sugar in 3 vases, and three tablespoons of sugar in 3 vases. In the remaining 3 vases, they put no sugar. After the vases were prepared, the students randomly assigned 1 carnation to each vase