**1.** 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.

The movement patterns of animals are believed to be affected by their physiological states (such as hunger), predation risk, and other factors such as whether they are searching for food for their young. Optimal foraging theory predicts that foragers (animals searching for food) with a central home location should move faster the farther they are from home.

American beavers are foragers with a central home location that make foraging trips from their lodge and then return. To test the optimal foraging theory, researchers trapped and radio-tagged a random sample of 67 American beavers from the population in a certain region and recorded their movements over the course of about a year. The researchers fit a least-squares regression line to the data, where the explanatory variable was distance to lodge, in meters, and the response variable was speed, in meters per hour. The resulting residual plot of the regression analysis is shown.
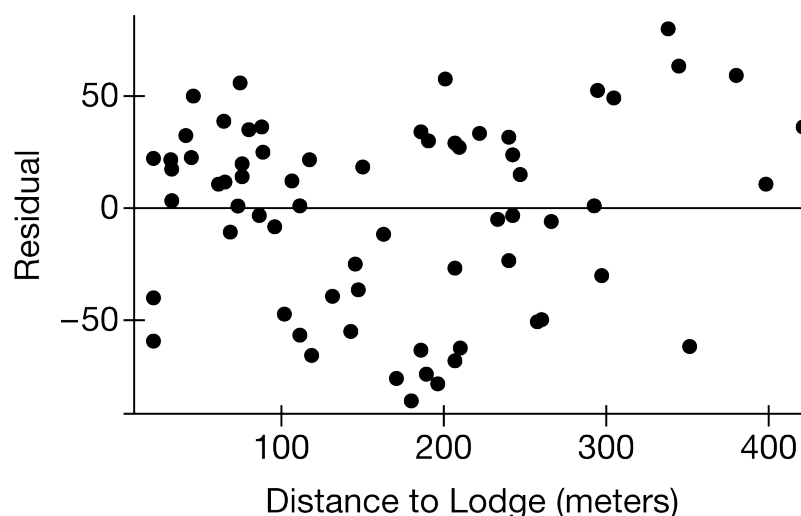
(a) Is a linear model appropriate to use for these data? Justify your response by referencing the residual plot.

The computer output from the regression analysis and the boxplot of the residuals are shown.

	Estimate	Error	T	P
Intercept	-16.462	9.726	-1.693	0.0953
Distance	0.849	0.049	17.177	0.0000
$S = 41.025$		R-sq $= 0.819$		
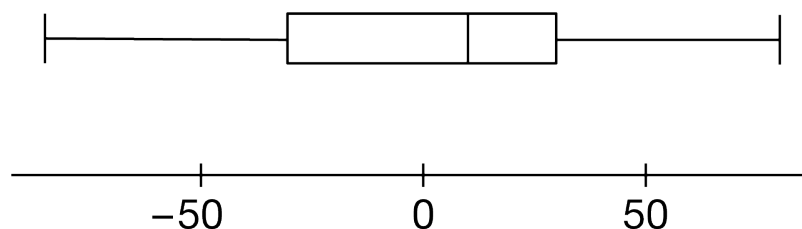
(b) Identify the intercept of the least-squares regression line and interpret it in the context of the data.

(c) The researchers plan to calculate a 95 percent confidence interval for the population slope. Identify and check the conditions for the appropriate confidence interval for the population slope.

(d) The 95% confidence interval for the population slope was calculated as $(0.751, 0.947)$. Based on the confidence interval, do the data provide convincing statistical evidence that foragers with a central home location move faster the farther they are from home? Justify your response.

Part A, B, C, and D

The primary goals of this question are to (1) assess a student's ability to identify whether a linear model is appropriate to use for a given set of data; (2) identify and interpret the intercept of a best fit line in context; (3) identify the conditions that must be checked before calculating a confidence interval for the population slope and determine whether the conditions have been met; and (4) use a confidence interval to assess a claim about a population slope.

Each essentially correct (E) part counts as 1 point.

Each partially correct (P) part counts as $1/2$ point.

<u>Scoring</u>

Parts (a), (b), (c), and (d) are scored as essentially correct (E), partially correct (P), or incorrect (I).

If a response is between two scores (for example, $2^1/_2$ points), use a holistic approach to decide whether to score up or down, depending on the overall strength of the response and communication.

Reasons to score up:

· *All notation is correct and clearly marked*

· *All explanations are clear*

· *No wrong information is included that was not part of the scoring (for example, saying sample size must be greater than 30 when that has nothing to do with the problem)*

· *No minor calculation errors are made, if they are not part of the scoring*

· *Interpretation parts are especially strong*

Reasons to score down:

· *Notation is not wrong, but is spotty and not clearly marked*

· *Explanations are not wrong, but are hard to follow*

· *Wrong or extraneous information is included but not part of scoring*

· *Minor calculation errors that are not part of the scoring are made*

· *Interpretation parts are scored an E but are considered a weak E*

✓

0	1	2	3	4

Parts (a) through (d) sum to 4 points

OR

Parts (a) through (d) sum to $3\frac{1}{2}$ points AND a holistic approach is used to decide to score up

- ☐ Part (a) essentially correct
- ☐ Part (a) partially correct
- ☐ Part (a) incorrect
- ☐ Part (b) essentially correct
- ☐ Part (b) partially correct
- ☐ Part (b) incorrect
- ☐ Part (c) essentially correct
- ☐ Part (c) partially correct
- ☐ Part (c) incorrect
- ☐ Part (d) essentially correct
- ☐ Part (d) partially correct
- ☐ Part (d) incorrect

Solution

Part (a):

Yes. The residual plot does not show any curvature, which indicates that the data points would exhibit no curvature about the best fit line, and thus that a linear model is appropriate.

Solution

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following two components.

· Correctly says "Yes"

· Justifies the choice by referencing the lack of curvature in the residual plot

Partially correct (P) if the response is "Yes" but does not reference the lack of curvature

Incorrect (I) if the response does not satisfy the criteria for E or P

Solution

Part (b):

The intercept of the line of best fit is -16.462 meters per hour. This is the predicted speed for a beaver that is 0 meters from the lodge (at or in the lodge).

Scoring

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

· Provides the correct value of the intercept with units

· Correctly interprets the intercept as an expected/predicted/average value for a distance from the lodge of 0 meters

Partially correct (P) if the response includes only one of the two components

Incorrect (I) if the response does not satisfy the criteria for E or P

Note:

A response that indicates that the intercept does not have a meaningful interpretation in context satisfies component 2.

Solution

Part (c):

The conditions are as follows:

1. The true relationship between the speed and distance from the lodge is linear.

· It is reasonable to assume that this condition is met, because the residual plot does not show any curvature and there is apparent randomness in the plot.

2. The standard deviation of errors must be constant across values of x

· It is reasonable to assume that this condition is met, because the residual plot does not exhibit a funnel shape (i.e., the variability of y decreasing or increasing as x increases).

3. The data points must be independent.

· It is reasonable to assume that this condition is met, because the problem states that the sample can be regarded as random.

4. For a particular value of the explanatory variable, the responses are normally distributed.

· It is reasonable to assume that this condition is met, because the boxplot of the residuals appears to be roughly symmetric and free of outliers.

Scoring

Part (c) is scored as follows:

Essentially correct (E) if three or four of the conditions are identified and justified

Partially correct (P) if two of the four conditions are identified and justified

Incorrect (I) if the response does not satisfy the criteria for E or P

Solution

Part (d):

Yes; Because 0 is not contained in the confidence interval, there is convincing statistical evidence that there is a relationship between distance to lodge and speed for American beavers.

Scoring

Part (d) is scored as follows:

Essentially correct (E) if the response satisfies the following three components.

· Responds "Yes"

· Indicates that 0 is not contained in the interval

· Provides context

Partially correct (P) if the response includes two of the three components.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Notes:

· A response that states that foragers with a central home location move faster the farther they are from home satisfies component 3.

· A response that indicates that the relationship is linear is considered a positive for holistic scoring.

2. A zoologist selected 12 black bears in a Canadian habitat at random to examine the relationship between the age in years, x, and the weight in tens of pounds, y.

The 95 percent confidence interval for estimating the population slope of the linear regression line predicting weight in tens of pounds based on the age in years is given by 1.272 ± 0.570.

Assume that the conditions for inference for the slope of the regression equation are met. Which of the following is the correct interpretation of the interval?

| (A) | We are 95 percent confident that the mean increase in the weight of a black bear for each one-year increase in the age of the bear is between 7.0 and 18.4 pounds. | ✓ |

(B) We are 95 percent confident that an increase of one year in the age of an individual black bear will result in an increase in the black bear's weight of between 7.0 and 18.4 pounds.

(C) We are 95 percent confident that for every one-year increase in the age of black bears in the sample, the average increase in the weights of those black bears is between 7.0 and 18.4 pounds.

(D) We are 95 percent confident that the mean increase in the age of a black bear for each one-pound increase in the weight of the black bear is between 7.0 and 18.4 years.

(E) We are 95 percent confident that any sample of 12 black bears will produce a slope of the regression line between 7.0 and 18.4.

Answer A

Correct. The 95 confidence interval given by 1.272 ± 0.570, or $(0.702, 1.842)$, estimates the relationship between weight in tens of pounds and age in years. So we are 95 percent confident that the mean increase in the weight of a black bear will be between 7.0 and 18.4 pounds for each one-year increase in the age of the bear.

3. Biologists are interested in how temperature changes might affect the frequency of mating calls of frogs. Twenty gray tree frogs are randomly chosen for a study. For each frog, the temperature of its habitat (in degrees Celsius) and the frequency of its mating call (in tones per second) are recorded.

The 96 percent confidence interval for estimating the population slope of the linear regression line predicting mating call frequency based on habitat temperature is given by 2.341 ± 0.768.

Assume that the conditions for inference for the slope of the regression line have been met. Which of the following is the correct interpretation of the calculated confidence interval?

(A) We are 96 percent confident that the increase in mating call frequency of an individual frog when its habitat temperature increases by 1 degree Celsius is between 1.573 and 3.109 tones per second.

(B) We are 96 percent confident that the average increase in mating call frequency in the sample of frogs when habitat temperature increases by 1 degree Celsius is between 1.573 and 3.109 tones per second.

| (C) | We are 96 percent confident that the average increase in mating call frequency in the population of frogs when habitat temperature increases by 1 degree Celsius is between 1.573 and 3.109 tones per second. | ✓ |

(D) We are 96 percent confident that the average increase in habitat temperature in the sample of frogs when mating call frequency increases by one tone per second is between 1.573 and 3.109 degrees Celsius.

(E) We are 96 percent confident that the average increase in habitat temperature in the population of frogs when mating call frequency increases by one tone per second is between 1.573 and 3.109 degrees Celsius.

> **Answer C**
>
> Correct. The confidence interval is a statement about how confident we are that we have captured the population parameter, which is the true slope of the regression line—that is, the average increase in mating call frequency among all gray tree frogs based on a 1-degree increase in habitat temperature.

4. A high school basketball coach wants to see whether there is a linear relationship between player height, x, and the number of points scored in a game by basketball players in the coach's state, y. The 96 percent confidence interval to estimate the slope of the linear regression line relating player height to points scored in a game is calculated to be $(-0.432, 1.844)$.

   Assume all conditions for inference for the slope of a regression line were met. Based on the confidence interval, which of the following claims is supported by the confidence interval?

   (A) Each player scored at most 1.844 points in a game.

   (B) The average difference between actual points scored and predicted points scored is between -0.432 and 1.844.

   (C) There is a positive linear relationship between player height and number of points scored for basketball players in the coach's state.

   (D) There is a negative linear relationship between player height and number of points scored for basketball players in the coach's state.

   (E) It cannot be determined whether the linear relationship between player height and number of points scored for basketball players in the coach's state is positive or negative. ✓

> **Answer E**
>
> Correct. A confidence interval contains plausible values for the population parameter (the slope of the regression line relating player height to number of points scored). Since the interval contains negative and positive values, it cannot be determined whether the slope of the regression line is positive or negative.

5. To investigate the relationship between the selling price of a house, y, in dollars, and the size of the house x, in square feet, a local builder collected data on a random sample of 120 houses from a certain region. Assume that the conditions for inference for the slope of a regression line are met. The resulting 95 percent confidence interval for the population slope of the regression line relating price and size is given by $(62, 99)$.

   The local builder claims that the selling price of houses from the region increases by $104 for every extra square foot of space in the house. Which of the following best describes the conclusion that can be reached about this claim based on the confidence interval?

(A) The claim is supported by the interval, since the interval does not contain the value 0.

(B) The claim is supported by the interval, since all values in the interval are positive.

(C) The claim is supported by the interval, since the interval does not contain the value 104.

(D) The claim is not supported by the interval, since the interval does not contain the value 0.

(E) The claim is not supported by the interval, since the interval does not contain the value 104. ✓

Answer E

Correct. A claim that the selling price increases by $104 for each square foot increase in space is <u>not</u> supported by the interval. The interval represents plausible values for the slope of the regression line, and the interval does not include the value 104.

6. An international organization is investigating the relationship between the life expectancies of men and women in nonindustrialized countries. A random sample of such countries was selected, and life expectancies, in years, were determined for both men and women. A check of the conditions necessary for inference on the slope of a regression line shows that they are met.

A 98 percent confidence interval for the slope of the regression line relating life expectancy for men, x, and women, y, is given by $(1.01, 1.34)$. Based on the interval, which of the following claims is supported?

(A) Since the interval does not contain 0, it can be concluded that there is no linear relationship between the life expectancies of men and women in nonindustrialized countries.

(B) Since the interval does not contain 1, it can be concluded that there is no linear relationship between the life expectancies of men and women in nonindustrialized countries.

(C) Since the values in the interval are greater than 0, it can be concluded that the life expectancies of women are greater than the life expectancies of men in nonindustrialized countries.

(D) Since the values in the interval are positive, it can be concluded that there is an increase, on average, in the life expectancies of women for each 1-year increase in the life expectancy of men in nonindustrialized countries. ✓

(E) Since the values in the interval are positive, it can be concluded that there is a decrease, on average, in the life expectancies of women for each 1-year increase in the life expectancy of men in nonindustrialized countries.

Answer D

Correct. The interval contains plausible values for the slope of the linear regression line relating the life expectancies of men and women in nonindustrialized countries. Since the values in the interval are positive, it can be concluded that a one-year increase in life expectancy for men is associated with an increase in life expectancy for women.

7. Peggy constructed the 95 percent confidence interval $(4.8, 5.2)$ to estimate the slope of a regression model for a set of bivariate data with 24 data values. Peggy claims that the width of the confidence interval will increase if a sample size of 30 is used, all other things remaining the same. Quincy claims that the width of the confidence interval will decrease if a sample size of 30 is used. Which statement is true about the claims made by Peggy and Quincy?

(A) Peggy's claim is correct.

(B) Quincy's claim is correct. ✓

(C) Both Peggy's claim and Quincy's claim are correct.

(D) Neither Peggy's claim nor Quincy's claim is correct.

(E) There is not enough information to determine whether the claims are correct.

Answer B

Correct. When all other things remain the same, a larger sample size results in a narrower confidence interval.

8. A 95 percent confidence interval for the slope of the regression line relating the number of grams of carbohydrates and the number of kilocalories per 100-gram sample of various raw foods is given by $(2.505, 6.696)$. The confidence interval is based on a random sample of n raw foods. A check of the conditions for inference on the slope shows they are reasonably met.

Which of the following is a correct interpretation of the interval?

(A) Ninety-five percent of all such samples of size n will produce a sample slope between 2.505 and 6.696 for the regression line relating grams of carbohydrates and kilocalories per 100-gram sample of various raw foods.

(B) The probability is 0.95 that the true slope of the regression line relating grams of carbohydrates and kilocalories per 100-gram sample of various raw foods is between 2.505 and 6.696.

(C) We are 95 percent confident that the slope of the regression line for the random sample of n raw foods is between 2.505 and 6.696.

(D) We are 95 percent confident that the predicted number of kilocalories per 100-gram sample will be between 2.505 and 6.696.

(E) We are 95 percent confident that the true slope of the regression line relating grams of carbohydrates and kilocalories per 100-gram sample of various raw foods is between 2.505 and 6.696. ✓

Answer E

Correct. The confidence interval is an interval in which we are 95 percent confident that the population parameter has been captured. So we are 95 percent confident that the true slope of the regression line relating grams of carbohydrates and kilocalories per 100-gram sample of raw foods is between 2.505 and

6.696.

9.   A statistician constructed the 95 percent confidence interval $(2.3, 3.7)$ to estimate the slope of a regression model for a set of bivariate data with 24 data values.

If the sample size n changes but all other things remain the same, which of the following claims about the confidence interval is true?

  (A)   The interval becomes narrower if $n = 20$.

  (B)   The interval width remains the same if $n = 20$.

  (C)   The interval becomes wider if $n = 28$.

  (D)   The interval becomes narrower if $n = 28$. ✓

  (E)   The interval width remains the same if $n = 28$.

Answer D

Correct. When all other things remain the same, a larger sample size will result in a decrease in the width of the confidence interval.

10.   A researcher was interested in the relationship between the number of texts sent in a day and the number of e-mails sent in a day by employees at a certain company. Using 15 data values, a 90 percent confidence interval for the slope of a regression model was found to be $(2.31, 3.47)$. The researcher claims that the interval would have been narrower with a different sample size if all other things remained the same.

Which of the following sample sizes would make the researcher's claim NOT true?

  (A)   14 ✓

  (B)   16

  (C)   20

  (D)   30

  (E)   100

Answer A

Correct. When all other things remain the same, a smaller sample size will result in a wider confidence interval, so the researcher's claim will not be true when a sample size smaller than 15 is used.