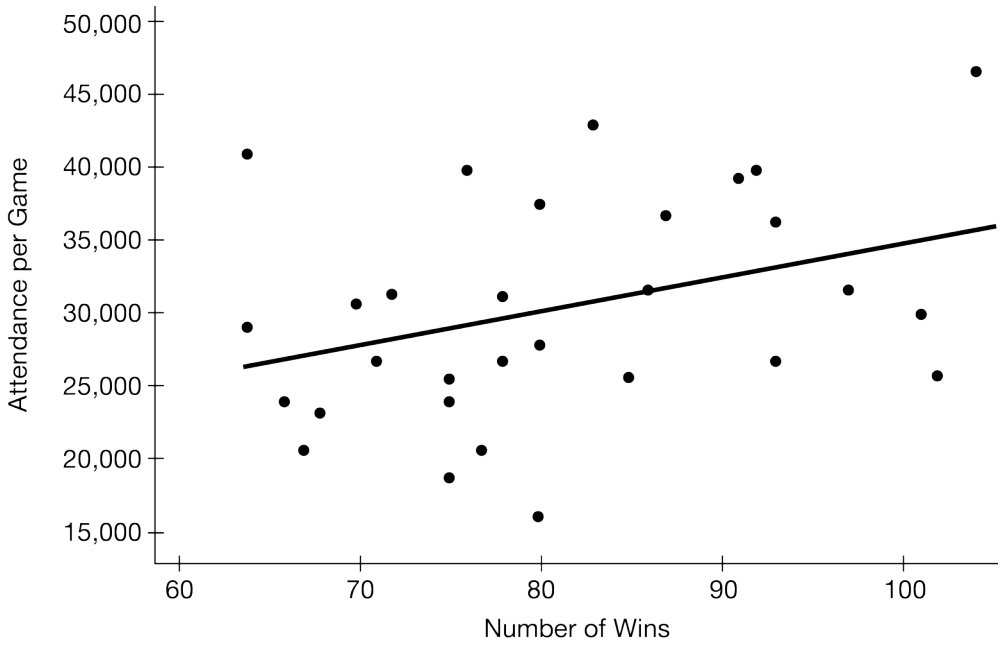
Analyzing Departures from Linearity Quiz

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 following scatterplot shows the number of wins and the attendance per game for 30 baseball teams in 2017. Also shown are the least-squares regression line and computer output.



Term	Coef	SE Coef	T-Value	P-Value
Constant	10834	9716	1.12	0.274
Wins	235	119	1.98	0.058
$S = 7{,}377$		$R - \text{sq} = 12.29\%$		$\text{Adj}\, R - \text{sq} = 9.16\%$

(a) Interpret the slope of the least-squares regression line in context.

(b) Explain why it is not reasonable to use the least-squares regression model to predict attendance per game for 0 wins.

(c) What is the value of the correlation coefficient for the sample?

(d) If the point representing 64 wins and attendance of $40{,}786$ people per game is removed from the set of data and a new regression analysis is conducted, how would the following be impacted? Explain your reasoning.

(i) The slope of the least-squares line:

Analyzing Departures from Linearity Quiz

(ii) The correlation coefficient:

Parts A, B, C, and D

Intent of Question

The primary goals of this question are to assess a student's ability to (1) identify the slope from computer output and interpret the slope; (2) explain why it is not reasonable to use extrapolation to predict a response variable; (3) calculate the correlation using computer output; and (4) discuss how a specific point influences the value of the slope and the correlation.

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

Each partially correct (P) part counts as ½ point.

Scoring

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, 2½ 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 ✓

Analyzing Departures from Linearity Quiz

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

OR

Parts (a) through (d) sum to 3½ 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): For each additional win, the predicted attendance per game increases by 235 people.

Scoring

Part (a) is scored as follows.

Essentially correct (E) if the response includes the following three components:

· The response correctly identifies the numerical value of the slope from the computer output.

· The response interprets the slope as the change in attendance per game for each additional win, in context.

· The interpretation of slope includes nondeterministic language (e.g., "predicted attendance per game").

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

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

Notes:

· A response that incorrectly identifies the numerical value of the slope can still satisfy components 2 and 3 using the incorrect value.

· Examples of nondeterministic language include "predicted attendance per game," "expected attendance per game," "estimated attendance per game," "average attendance per game," "…attendance per game, on average," and so on.

Analyzing Departures from Linearity Quiz

However, “about” and “approximately” do not satisfy component 3 as they could be referring to rounding.

· Responses that use “attendance” rather than “attendance per game” can still satisfy the context requirement of component 2.

Solution

Part (b): It is not reasonable to predict the value of attendance per game when the number of wins is equal to 0. The number of wins in the data set only includes values from 64 to 104, so we cannot be confident that the linear model is a good predictor of attendance per game if we were to extrapolate outside of this interval, including at $x = 0$.

Scoring

Part (b) is scored as follows.

Essentially correct (E) if the response includes the following three components:

· The response states that the *x* values used to create the model do not include $x = 0$.

· The response provides numerical evidence that the *x* values used to create the model do not include 0 (e.g., the *x* values go from about 64 wins to about 104 wins).

· The response uses at least one of the variable names (number of wins, attendance per game).

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

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

Notes:

· A response that says a prediction for $x = 0$ would be an extrapolation satisfies component 1.

· Responses that use “attendance” rather than “attendance per game” can still satisfy component 3.

Solution

Part (c): $r = \sqrt{0.1229} \approx 0.351$

Scoring

Part (c) is scored as follows.

Essentially correct (E) if the response has the correct value of and includes supporting work.

Partially correct (P) if the response

· has the correct value of $r = 0.351$ with no supporting work,

OR

· states that $r = \pm 0.351$ with supporting work,

OR

Analyzing Departures from Linearity Quiz

· incorrectly uses the value of adjusted r^2 to get $r = \sqrt{0.0916} = 0.303$ with supporting work.

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

Solution

Part (d-i): If the point were removed, the slope of the new least-squares regression line would be greater. Because the point is to the left of $\bar{x}$ and above the least-squares regression line, the slope of the least-squares regression with the point included is less steep (closer to 0). Because the slope of the regression line changes substantially, the point $(64,\ 40,786)$ is an influential point.

Part (d-ii): If the point were removed, the correlation of the new least-squares regression line would be greater (closer to 1). Because this point is outside the linear pattern of the other data points, the correlation of the least-squares regression with the point included is less (closer to 0) than that of the new least-squares regression line.

Scoring

Part (d) is scored as follows.

Essentially correct (E) if the response includes the following four components:

· In part (d-i), the response states that removing the identified point would cause the slope to be steeper.

· In part (d-i), the response includes an explanation based on the location of the identified point relative to the other points.

· In part (d-ii), the response states that removing the identified point would cause the correlation coefficient to be greater (closer to 1) or stronger.

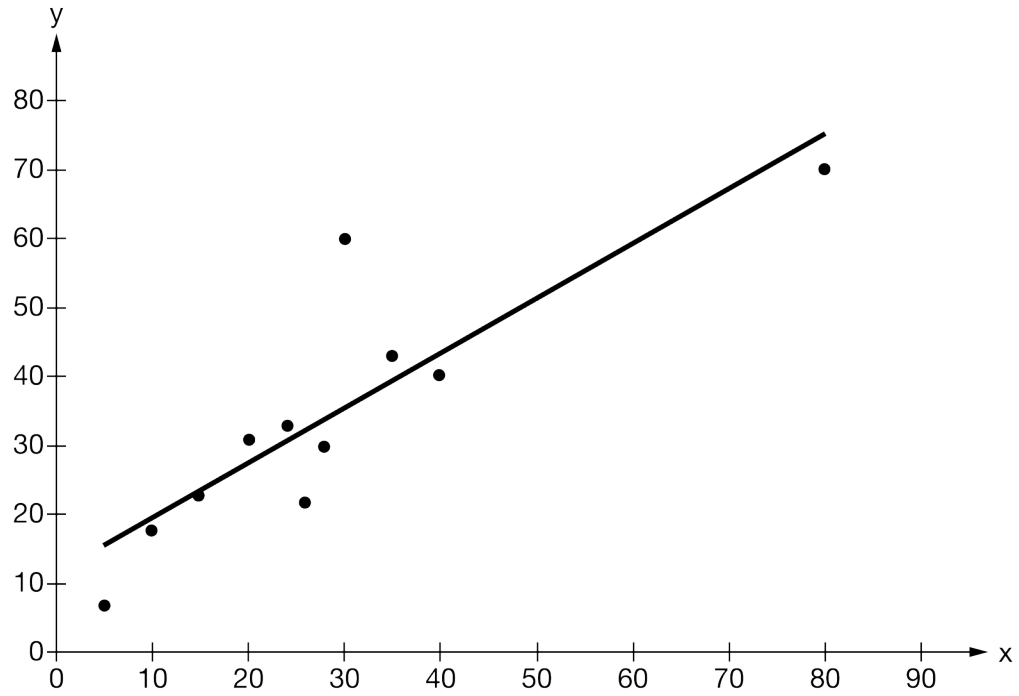
· In part (d-ii), the response includes an explanation based on the location of the identified point relative to the other points.

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

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

Analyzing Departures from Linearity Quiz

2. The following scatterplot shows two variables, x and y, along with a least-squares model.



Which of the following is a high leverage point with respect to the regression?

(A) $(5, 8)$

(B) $(20, 31)$

(C) $(27, 22)$

(D) $(30, 60)$

(E) $(80, 70)$ ✓

Answer E

Correct. A high leverage point is one that has a substantially larger or smaller x-value than the other observations. The x-value of 80 is substantially larger than the other x-values that occur between 5 and 40.

3. An exponential relationship exists between the explanatory variable and the response variable in a set of data. The common logarithm of each value of the response variable is taken, and the least-squares regression line has an equation of $\log(\hat{y}) = 7.3 - 1.5x$. Which of the following is closest to the predicted value of the response variable for $x = 4.8$?

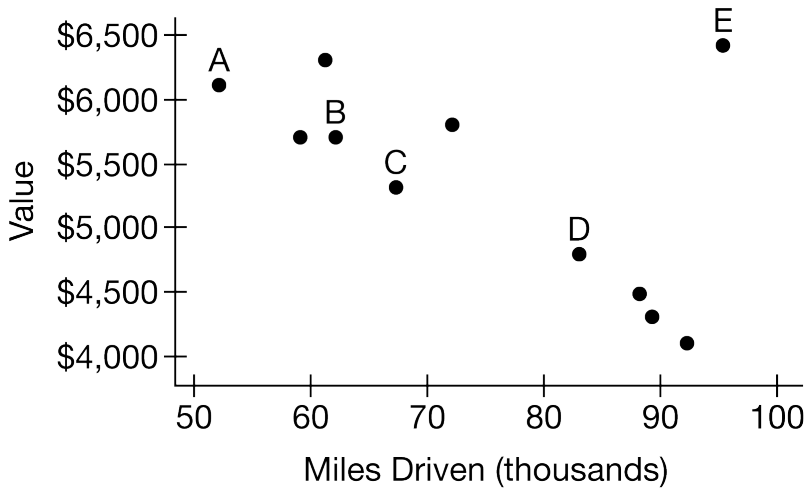
Analyzing Departures from Linearity Quiz

(A) 0.1
(B) 0.68
(C) 1.105
(D) 1.26 ✓
(E) 14.5

Answer D

Correct. Substituting $x = 4.8$ into the equation gives $\log(\hat{y}) = 7.3 - 1.5(4.8)$ or $\log(\hat{y}) = 0.1$. To solve for $\hat{y}$, raise 10 to the power of 0.1 to get 1.26.

4. In a study to determine whether miles driven is a good predictor of trade-in value, 11 cars of the same age, make, model, and condition were randomly selected. The following scatterplot shows trade-in value and mileage for those cars. Five of the points are labeled A, B, C, D, and E, respectively.



Which of the five labeled points is the most influential with respect to a regression of trade-in value versus miles driven?

(A) A
(B) B
(C) C
(D) D
(E) E ✓

Analyzing Departures from Linearity Quiz

Answer E

Correct. Point E does not follow the trend with respect to the other data and is probably an outlier. The value of the car is much higher than other cars with similar miles driven.

5. Data were collected on two variables, x and y, to create a model to predict y from x. A scatterplot of the collected data revealed a curved pattern with a possible cubic relationship ($y = ax^3$, where a is a constant) between the variables. Which of the following transformations would be most appropriate for creating linearity between the variables?

(A) Taking the cube of y

(B) Taking the cube root of y

(C) Taking the cube root of both y and x

(D) Taking the log of y

(E) Taking the log of both y and x ✓

Answer E

Correct. Variables related by a power, such as $y = x^3$, are best transformed by taking the log of both variables.

6. The relationship between carbon dioxide emissions and fuel efficiency of a certain car can be modeled by the least-squares regression equation $\ln(\hat{y}) = 7 - 0.045x$, where x represents the fuel efficiency, in miles per gallon, and $\hat{y}$ represents the predicted carbon dioxide emissions, in grams per mile.

Which of the following is closest to the predicted carbon dioxide emissions, in grams per mile, for a car of this type with a fuel efficiency of 20 miles per gallon?

(A) 1.8

(B) 6.1

(C) 446 ✓

(D) 2,697

(E) 1,250,000

Answer C

Correct. When 20 is substituted for x, the resulting value on the right side of the equation is 6.1. The

value of approximately 446 results from raising e to the power of 6.1 $\left(\text{that is}, e^{6.1}\right)$.

7. Which of the following statements about a least-squares regression analysis is true?

I. A point with a large residual is an outlier.
II. A point with high leverage has a y-value that is not consistent with the other y-values in the set.
III. The removal of an influential point from a data set could change the value of the correlation coefficient.

(A) I only
(B) II only
(C) I and III only ✓
(D) III only
(E) I, II, and III

Answer C

Correct. Statement I and III are both true. A point with a large residual is an outlier and a influential point is one for which its removal from the set can have a substantial effect on the correlation.