AP Stats: Lunchtime Review #3
Bivariate Data

Bivariate Data: Chapter 3

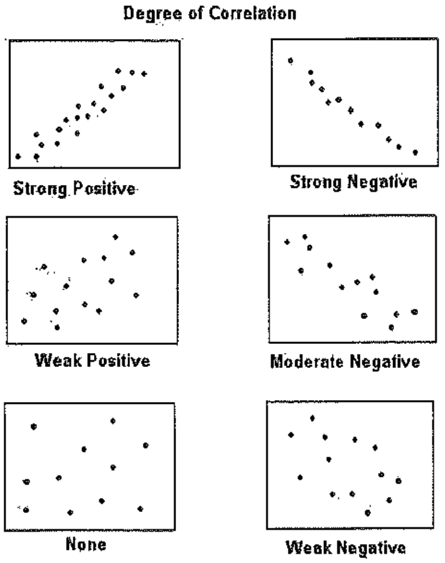
- Collecting data on two different variables
- Is there a relationship between these two variables?

Any time you're asked to comment on the relationship between two variables, mention form, strength, and direction of the two variables.

Form

- Linear: the best fitting line (by definition) minimizes the sum of the squares of the residuals.
  - $\hat{y} = a + bx$
  - always includes * $(\bar{x}, \bar{y})$
  - slope $b = \boxed{r}(S_y/S_X)$
- Non-linear
  - Can be transformed to linear models using logs and powers (Ch 12)

Direction



Strength:

- Correlation coefficient, r, quantifies the degree and direction of the linear relationship between two quantitative variables: $-1 \leq r \leq +1$

Closer to ±1 more of a line.

- Coefficient of determination $\boxed{r^2}$ the percent of variation in Y-values attributed to the linear relationship between X and Y.

Important terminology/concepts/displays:

- Y is the dependent variable, or response variable
- X is the independent variable, or explanatory variable
- We are truly examining the values of y as assessing the fit of the relationship between the two variables.
  - Residual: $\overset{y}{\text{Observed}} \overset{-}{-} \overset{\hat{y}}{\text{Expected}}$
  - Residual Plot: plot of explanatory values and corresponding residuals on the y-axis. It is used to assess if a line was the best choice to fit the data. If so, there should be no distinct pattern around the line residual = 0; if a line was not the best fit, there should be a pattern in the residuals.
- Slope: amount of change in y-units for every x unit.
- Y –intercept: $\boxed{\text{the value of y when x = 0}}$
- Outliers are points that are far away from the overall pattern of a scatterplot
- Influential points are observations that are extreme in the trend, and would sharply change the regression line equation.

variable. Use caution not to extrapolate when making predictions, though, as we do not know if the relationship between the variables extends far beyond the observed values of *x*!

Concept 2: Least-Squares Regression Line

Chances are any scatterplot you construct or encounter will not display a perfectly straight line. In most cases, the observed points will be, well, scattered. Since most of our observed relationships are not perfectly linear, predictions of *y* made from our regression line will often be different than observed *y* values, resulting in a prediction error. That is, there will be some amount of vertical distance between the regression line and the observed value. This vertical difference (observed *y* – predicted *y*) is called a residual. The regression line that "best fits" our observed data is the one that minimizes the squared residuals. This "line of best fit" that minimizes that prediction error is called the least-squares regression line.

Familiarize yourself with the formulas that can be used to determine the slope and intercept of the least-squares regression line. We will rely on technology to generate this equation, but you should recognize that we can construct the equation by hand given the mean and standard deviation of *x* and *y* as well as the correlation *r* between them.

Once you have the equation of the least-squares regression line, you should be able to interpret it and use it. The most important feature to note when interpreting is the slope. You should be able to explain what the slope means in the context of the variables you are analyzing. That is, the slope represents the expected change in the predicted *y* value for each one-unit increase of the *x* value. Be sure to get familiar with this interpretation as you may be asked to provide it on the AP Exam!

Check for Understanding: ____ *I can construct, interpret, and apply the least-squares regression line.*

Using the following data, determine the least-squares regression line to predict exam scores from anxiety scores. Note: Higher anxiety scores indicate higher levels of test anxiety.

Anxiety	23	14	14	0	7	20	20	15	21	4
Exam Score	43	59	48	77	50	52	46	60	51	70

a) What is the equation of the least-squares regression line?

b) Interpret the slope of the least-squares regression line in the context of the situation.

c) What exam score can we predict for an anxiety score of 15?

d) What is the residual for an anxiety score of 15?

e) Would you use your least-squares regression line to predict an exam score for a person who had an anxiety score of 35? Why or why not?

Concept 3: Assessing How Well the Least-Squares Regression Line Fits the Data

In Section 3.1, we learned that our eyes aren't always the best judge of linear relationships. While correlation *r* gives us a better understanding of the strength of the linear relationship, we still need to assess how well the least-squares regression line fits the observed data. If it fits well, it may be a useful prediction tool. If it doesn't fit well, we may want to search for a model that fits it better.

One way to assess how well the least-squares regression line fits our data is to make a residual plot. Plotting the residuals gives us more information about the relationship between quantitative variables and helps us assess how well a linear model fits the data. If the residual plot displays a pattern, a better (perhaps nonlinear) model might exist!

We can also assess the fit of the least-squares regression line by interpreting the coefficient of determination r^2. r^2 is a measure of how well the regression model explains the response. Specifically, it is interpreted as the fraction of variation in the values of *y* that is explained by the least-squares regression line of *y* on *x*. For example, if r^2= 0.82, we can say that 82% of the variation in y is due to the linear relationship between *y* and *x*. 18% is due to factors other than x.

Check for Understanding: ____ *I can assess how well the least-squares regression line fits the data.*

Consider the equation of the least squares regression line of exam score on anxiety.

1) Construct and interpret the residual plot for the least-squares regression line.

2) What is the value of *r*? What is the value of r^2? Interpret each of these in the context of the problem.

Concept 4: Interpreting Computer Regression Output
As noted already, we will often rely on technology to generate the equation of the least-squares regression line. You are probably familiar with using your calculator to produce the equation. Make sure you can also interpret computer output to identify the slope and intercept of the regression line as well as other important values such as correlation and the coefficient of determination. There is a strong possibility you will need to read computer output on the AP Exam!

Check for Understanding: ____ *I can construct or identify the equation of a least squares regression line.*

A study was performed to determine the effect of temperature on a pond's algae level. Temperature was measured in degrees F, and algae level was measured in parts per million. Consider the computer output below.

```
Predictor  Coef     Stdev    t-ratio   p
Constant  42.8477  5.750     77.40     0.000
Temp       0.47620  0.5911    13.70     0.000
s = 0.4224         R-sq= 91.7%   R-sq(adj)=91.2%
```

1) Write the equation of the least squares regression line. Identify any variables used.

2) Interpret the slope of the least-squares regression line.

3) Identify and interpret the correlation coefficient.

4) Identify and interpret the standard deviation of the residuals.

Chapter Summary: Modeling Distributions of Data

In this chapter, we expanded our toolbox for working with quantitative data. We learned how to analyze and describe the relationship between two quantitative variables. Using scatterplots, we can display the relationship and describe the direction, strength, and form of the overall pattern. Correlation provides a numerical summary of the strength of the linear relationship between the variables and the equation of the least-squares regression line provides a model that can be used to make predictions. Residual plots, the standard deviation of the residuals, and the coefficient of determination help us assess the fit of the least-squares regression line and may suggest whether or not a linear model is appropriate. Finally, we learned that outliers and influential points can affect our interpretations and regression results. Just like we did with a single quantitative variable, we should be able to identify departures from the overall pattern and explain their influence on our analysis.

Perhaps the most important note for this chapter, though, is that while we now have some tools to help us describe the relationship between two quantitative variables, correlation does not always imply causation!

After You Read: "How Can I Close the Gap?"

Complete the vocabulary puzzle, multiple choice questions, and FRAPPY. Check your answers and your performance on each of the targets.

Target	Got It!	Almost There	Needs Some Work
I can identify explanatory and response variables in bivariate situations			
I can construct and interpret a scatterplot to display a bivariate relationship			
I can describe the direction, form, and strength of the pattern in a scatterplot			
I can calculate and interpret correlation			
I can identify outliers in a scatterplot and explain their effects on correlation			
I can construct or identify the equation of a least-squares regression line			
I can interpret the slope and *y*-intercept of a least-squares regression line			
I can calculate and interpret residuals			
I can construct and interpret residual plots			
I can explain the dangers of extrapolation			
I can use the least-squares regression line to predict values of the response variable			
I can use the standard deviation of the residuals to assess how well the line fits the data			
I can use r^2 to assess how well the line fits the data			

Did you check "Needs Some Work" for any of the targets? If so, what will you do to address your needs for those targets?

Learning Plan:

Chapter 3 Multiple Choice Practice

Directions. *Identify the choice that best completes the statement or answers the question. Check your answers and note your performance when you are finished.*

1. A study is conducted to determine if one can predict the academic performance of a first year college student based on their high school grade point average. The explanatory variable in this study is

A. academic performance of the first year student.
B. grade point average.
C. the experimenter.
D. number of credits the student is taking.
E. the college.

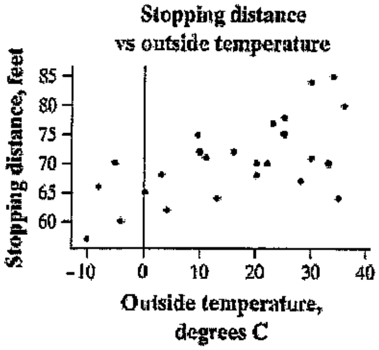
2. If two variables are positively associated, then

A. larger values of one variable are associated with larger values of the other.
B. larger values of one variable are associated with smaller values of the other.
C. smaller values of one variable are associated with larger values of the other.
D. smaller values of one variable are associated with both larger or smaller values of the other.
E. there is no pattern in the relationship between the two variables.

3. The correlation coefficient measures

A. whether there is a relationship between two variables.
B. the strength of the relationship between two quantitative variables.
C. whether or not a scatterplot shows an interesting pattern.
D. whether a cause and effect relation exists between two variables.
E. the strength of the linear relationship between two quantitative variables.

4. Consider the following scatterplot, which describes the relationship between stopping distance (in feet) and air temperature (in degrees Centigrade) for a certain 2,000-pound car travelling 40 mph.



Do these data provide strong evidence that warmer temperatures actually *cause* a greater stopping distance?

A. Yes. The strong straight-line association in the plot shows that temperature has a strong effect on stopping distance.
B. No. $r \neq +1$
C. No. We can't be sure the temperature is responsible for the difference in stopping distances.
D. No. The plot shows that differences among stopping distances are not large enough to be important.
E. No. The plot shows that stopping distances go down as temperature increases

5. If stopping distance was expressed in yards instead of feet, how would the correlation *r* between temperatures and stopping distance change?

A. *r* would be divided by 12.
B. *r* would be divided by 3.
C. *r* would not change.
D. *r* would be multiplied by 3.
E. *r* would be multiplied by 12.

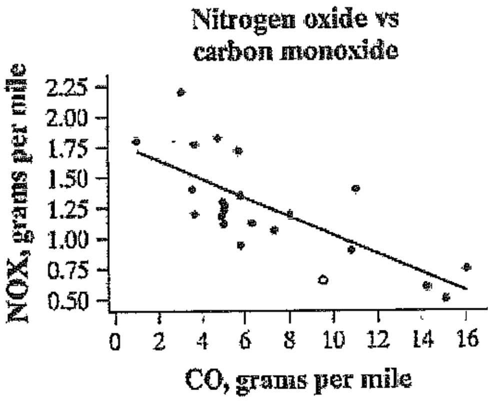
6. If another data point were added with an air temperature of 0° C and a stopping distance of 80 feet, the correlation would

A. decrease, since this new point is an outlier that does not follow the pattern in the data.
B. increase, since this new point is an outlier that does not follow the pattern in the data.
C. stay nearly the same, since correlation is resistant to outliers.
D. increase, since there would be more data points.
E. Whether this data point causes an increase or decrease cannot be determined without recalculating the correlation.

7. Which of the following is true of the correlation *r*?

A. It is a resistant measure of association.
B. $-1 \leq r \leq 1$.
C. If *r* is the correlation between *X* and *Y*, then -*r* is the correlation between *Y* and *X*.
D. Whenever all the data lie on a perfectly straight-line, the correlation *r* will always be equal to +1.0.
E. All of the above.

Consider the following scatterplot of amounts of CO (carbon monoxide) and NOX (nitrogen oxide) in grams per mile driven in the exhausts of cars. The least-squares regression line has been drawn in the plot.



8. Based on the scatterplot, the least-squares line would predict that a car that emits 2 grams of CO per mile driven would emit approximately how many grams of NOX per mile driven?

A. 4.0
B. 1.25
C. 2.0
D. 1.7
E. 0.7

9. In the scatterplot, the point indicated by the open circle

A. has a negative value for the residual.
B. has a positive value for the residual.
C. has a zero value for the residual.
D. has a zero value for the correlation.
E. is an outlier.

10. Which of the following is correct?

A. The correlation *r* is the slope of the least-squares regression line.

B. The square of the correlation is the slope of the least-squares regression line.

C. The square of the correlation is the proportion of the data lying on the least-squares regression line.

D. The coefficient of determination is the fraction of variability in *y* that can be explained by least-squares regression of *y* on *x*.

E. The sum of the squared residuals from the least-squares line is 0.

11. Which of the following statements concerning residuals from a LSRL is true?

A. The sum of the residuals is always 0.

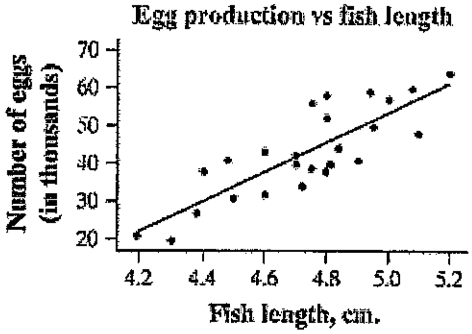
B. A plot of the residuals is useful for assessing the fit of the least-squares regression line.

C. The value of a residual is the observed value of the response minus the value of the response that one would predict from the least-squares regression line.

D. An influential point on a scatterplot is not necessarily the point with the largest residual.

E. All of the above.

A fisheries biologist studying whitefish in a Canadian Lake collected data on the length (in centimeters) and egg production for 25 female fish. A scatter plot of her results and computer regression analysis of egg production versus fish length are given below.
Note that Number of eggs is given in thousands (i.e., "40" means 40,000 eggs).



```
Predictor        Coef   SE Coef       T      P
Constant      -142.74     25.55   -5.59  0.000
Fish length    39.250     5.392    7.28  0.000

S = 6.75133   R-Sq = 69.7%   R-Sq(adj) = 68.4%
```

12. Which of the following statements is a correct interpretation of the slope of the regression line?

A. For each 1-cm increase in the fish length, the predicted number of eggs increases by 39.25.

B. For each 1-cm increase in the fish length, the predicted number of eggs decreases by 142.74.

C. For each 1-unit increase in the number of eggs, the predicted fish length increases by 39.25 cm.

D. For each 1-unit increase in the number of eggs, the predicted fish length decreases by 142.74cm.

E. For each 1-cm increase in the fish length, the predicted number of eggs increases by 39,250.

13. What percent of variability in the number of eggs is explained by the least-squares regression of *number of eggs* on *fish length*?

A. 25.55

B. 5.392

C. 6.75133

D. 69.7

E. Cannot be determined without the original data.

14. A study of the effects of television measured how many hours of television each of 125 grade school children watched per week during a school year and their reading scores. The study found that children who watch more television tend to have lower reading scores than children who watch fewer hours of television. The study report says that, "Hours of television watched explained 25% of the observed variation in the reading scores of the 125 subjects." The correlation between hours of TV and reading score must be

A. $r = 0.25$.
B. $r = -0.25$.
C. $r = -0.5$.
D. $r = 0.5$.
E. Can't tell from the information given.

15. A study gathers data on the outside temperature during the winter in degrees Fahrenheit and the amount of natural gas a household consumes in cubic feet per day. Call the temperature *x* and gas consumption *y*. The house is heated with gas, so *x* helps explain *y*. The least-squares regression line for predicting *y* from *x* is: $\hat{y} = 1344 - 19x$. When the temperature goes up 1 degree, what happens to the gas usage predicted by the regression line?

A. It goes up 19 cubic feet.
B. It goes down 19 cubic feet.
C. It goes up 1344 cubic feet.
D. It goes down 1344 cubic feet.
E. Can't tell without seeing the data.

Problem	Answer	Concept	Right	Wrong	Simple Mistake?	Need to Study More
1	B	Explanatory vs. Response				
2	A	Definition of Association				
3	E	Definition of Correlation				
4	C	Correlation vs. Causation				
5	C	Correlation				
6	A	Correlation				
7	B	Correlation				
8	D	Predicting with the LSRL				
9	A	Residuals				
10	D	Coefficient of Determination				
11	E	Residuals				
12	E	Slope of the LSRL				
13	D	Coefficient of Determination				
14	C	Coefficient of Determination				
15	B	Slope of the LSRL				

FRAPPY! Free Response AP Problem, Yay!

The following problem is modeled after actual Advanced Placement Statistics free response questions. Your task is to generate a complete, concise response in 15 minutes. After you generate your response, view two example solutions and determine whether or not you feel they are "complete," "substantial," "developing" or "minimal". If they are not "complete," what would you suggest to the student who wrote them to increase their score? Finally, you will be provided with a rubric. Score your response and note what, if anything, you would do differently to increase your own score.

A recent study was interested in determining the optimal location for fire stations in a suburban city. Ideally, fire stations should be placed so the distance between the station and residences is minimized. One component of the study examined the relationship between the amount of fire damage *y* (in thousands of dollars) and the distance between the fire station and the residence *x* (in miles). The results of the regression analysis are below.

```
Predictor    Coef     SE Coef       T        P
Constant     10.28    1.42        7.237    0.000
X             4.92    0.39       12.525    0.000

s = 2.232    R-Sq = 0.9235  R-Sq(adj) = 0.9176
```

(a) Write the equation of the least squares regression line. Define any variables used. Interpret the slope of the equation in context.

(b) A home located 3 miles from the fire station received $22,300 in damage. Use your equation in part (a) to calculate and interpret the residual for this observation.

(c) Identify and interpret the correlation coefficient.

Student Response 1:

a) $\hat{y} = 10.28 + 4.92x$
For each additional mile between the fire station and residence, we predict about $4920 additional dollars in damages.

b) $\hat{y} = 10.28 + 4.92(3) = 25.04$. Residual = 25.04 – 22.3 = 2.74. Our model overpredicted the amount of damage for this observation by $2740.

c) $r^2 = 0.9235$. There is a strong, positive linear relationship between the distance between a fire station and residence and the resulting damage in a fire.

How would you score this response? Is it substantial? Complete? Developing? Minimal? Is there anything this student could do to earn a better score?

Student Response 2:

a) $\widehat{\mathit{firedamage}} = 4.92\mathit{distance} + 10.28$
We predict about $4920 additional dollars in damage for each increase of one mile between the fire station and residence that is on fire.

b) $\widehat{\mathit{damage}} = 2.92(3) + 10.28 = 25.04$
residual = 22.3 – 25.04 = - 2.74. Our model overpredicts the damage amount by $2740.

c) r = 0.96. There is a very strong, positive, linear relationship between a residence's damage from a fire and its distance from a fire station.

How would you score this response? Is it substantial? Complete? Developing? Minimal? Is there anything this student could do to earn a better score?

Scoring Rubric

Use the following rubric to score your response. Each part receives a score of "Essentially Correct," "Partially Correct," or "Incorrect." When you have scored your response, reflect on your understanding of the concepts addressed in this problem. If necessary, note what you would do differently on future questions like this to increase your score.

Intent of the Question

The goal of this question is to determine your ability to interpret computer regression output and explain key concepts of linear regression.

Solution

(a) $\widehat{firedamage} = 10.28 + 4.92 distance$ OR $\hat{y} = 10.28 + 4.92x$ with x and y defined as distance and damage.

For each additional mile between the fire station and residence, we predict about $4920 additional dollars in damages.

(b) $\widehat{damage} = 2.92(3) + 10.28 = 25.04$

residual = 22.3 – 25.04 = - 2.74.

The model overpredicts the damage amount by $2740.

(c) Since $r^2 = 0.9325$, $r = 0.96$. There is a very strong, positive, linear relationship between a residence's damage from a fire and its distance from a fire station.

Scoring

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

Part (a) is essentially correct if the response (1) correctly identifies the least-squares regression equation in context or with variables defined and (2) correctly interprets the slope

Part (a) is partially correct if the response fails to define the variables in context or reverses the coefficients OR if the slope is not correctly defined in context (eg, predicts 4.92 dollars instead of $4920).

Part (b) is essentially correct if (1) the correct residual is calculated and (2) the interpretation is correct.

Part (b) is partially correct if only one of the above elements is correct.

Part (c) is essentially correct if the correlation coefficient is correctly identified and interpreted correctly with all three elements (strong, positive, linear).

Part (c) is partially correct if one of the elements (strong, positive, linear) is missing OR if r^2 is used instead of r.

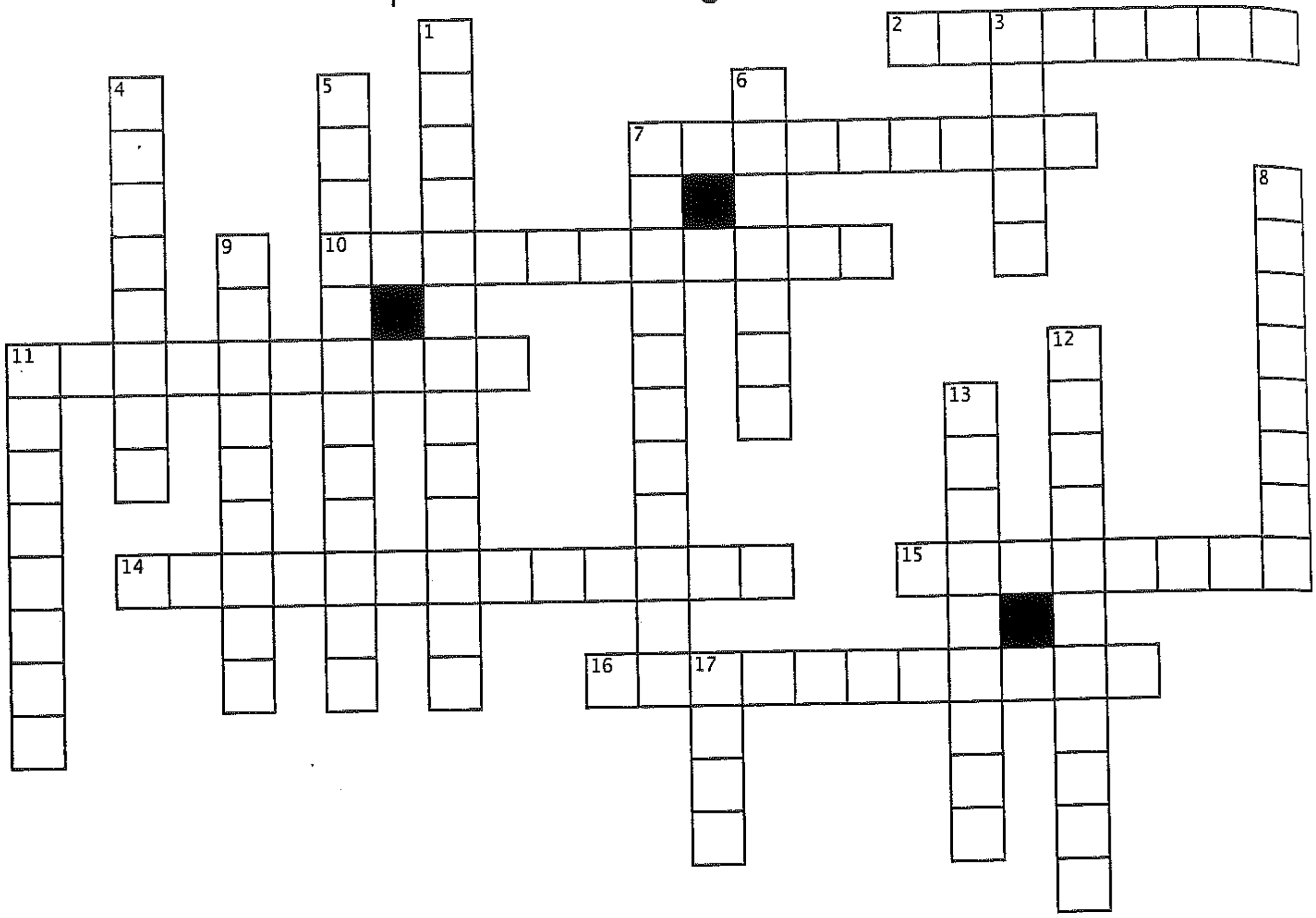
4 **Complete Response**
All three parts essentially correct

3 **Substantial Response**
Two parts essentially correct and one part partially correct

2 **Developing Response**
Two parts essentially correct and no parts partially correct
One part essentially correct and two parts partially correct
Three parts partially correct

1 **Minimal Response**
One part essentially correct and one part partially correct
One part essentially correct and no parts partially correct
No parts essentially correct and two parts partially correct

Chapter 3: Describing Relationships



Across

2. the difference between an observed value of the response and the value predicted by a regression line
7. Important note: Association does not imply ______.
10. graphical display of the relationship between two quantitative variables
11. line that describes the relationship between two quantitative variables
14. the coefficient of ______ describes the fraction of variability in y values that is explained by least squares regression on x.
15. A _____ association is defined when above average values of one variable are accompanied by below average values of the other.
16. individual points that substantially change the correlation or slope of the regression line

Down

1. the use of a regression line to make a prediction far outside the observed x values
3. the amount by which y is predicted to change when x increase by one unit
4. The ______ of a relationship in a scatterplot is determined by h closely the point follow a clear form.
5. the ____-_____ regression line is also known as the line of best (2 words)
6. an individual value that falls outside the overall pattern of the relationship
7. value that measures the strength of the linear relationship between two quantitative variables
8. A ______ association is defined when above average values of the explanatory are accompanied by above average values of response
9. y-hat is the ______ value of the y-variable for a given x
11. variable that measures the outcome of a study
12. variable that may help explain or influence changes in another variable
13. The _____ of a scatterplot indicates a positive or negative association between the variables.
17. The ____ of a scatterplot is usually linear or nonlinear.