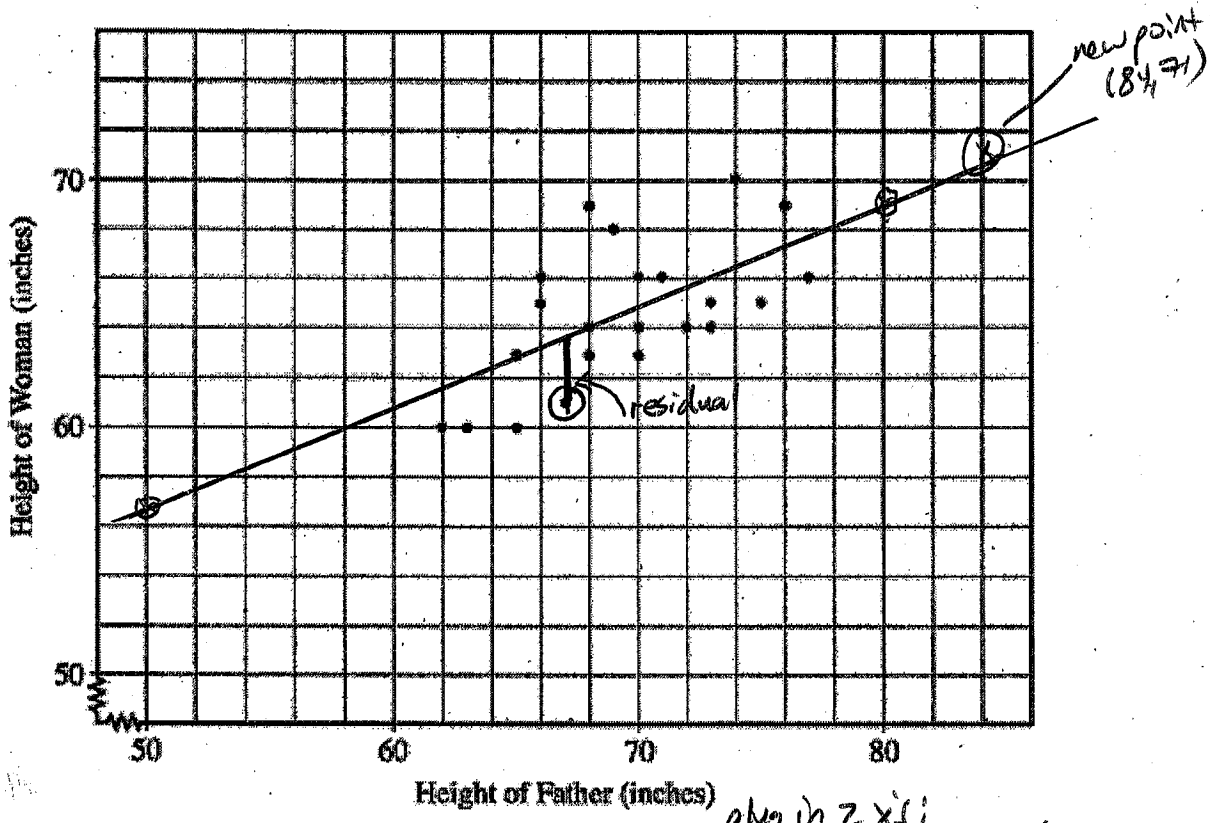
Solutions

AP Statistics – Unit 2 additional review

Free-Response Practice

Each of 25 adult women was asked to provide her own height (y), in inches, and the height (x), in inches, of her father. The scatterplot below displays the results. Only 22 of the 25 pairs are distinguishable because some of the (x,y) pairs were the same. The equation of the least squares regression line is $\hat{y} = 35.1 + 0.427x$.



(a) Draw the least squares regression line on the scatterplot above.

plug in 2 x's:
$\hat{y} = 35.1 + .427(50) = 56.45$
$\hat{y} = 35.1 + .427(80) = 69.26$

(b) One father's height was $x = 67$ inches and his daughter's height was $y = 61$ inches. Circle the point on the scatterplot above that represents this pair and draw the segment on the scatterplot that corresponds to the residual for it. Give a numerical value for the residual.

$\hat{y} = 35.1 + .427(67) = 63.709$

$y = 61$

$\text{residual} = y - \hat{y} = 61 - 63.709 = \boxed{-2.709 \text{ inches}}$

(actual − predicted)

(c) Suppose the point $x = 84$, $y = 71$ is added to the data set. Would the slope of the least squares regression line increase, decrease, or remain about the same? Explain.

(Note: No calculations are necessary to answer this question.)

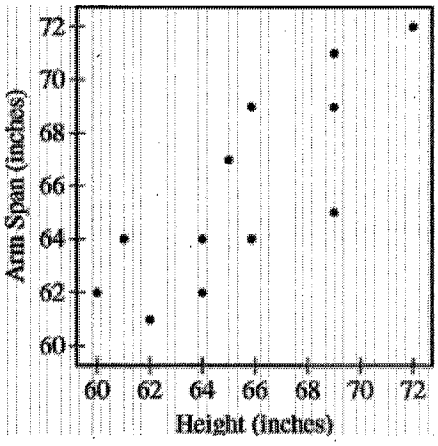
The slope would stay about the same. Although this point has high leverage, there is almost no residual.

Would the correlation increase, decrease, or remain about the same? Explain.

(Note: No calculations are necessary to answer this question.)

The correlation would become stronger (increase) because this point is close to the LSRL so points are now closer to the LSRL, on average.

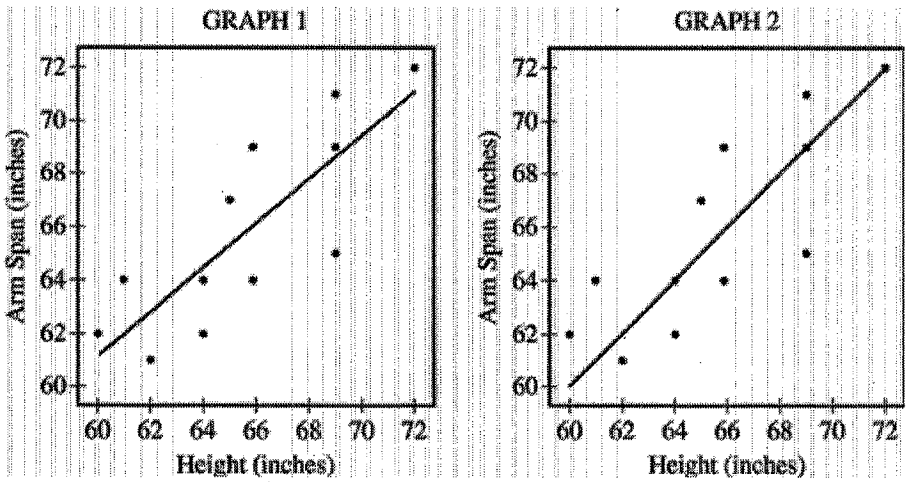
A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.



(a) Based on the scatterplot, describe the relationship between arm span and height for the sample of 12 seniors.

There is a medium-strength, positive, linear relationship between arm span and height

Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line $\hat{y} = 11.74 + 0.8247x$, and graph 2 shows the data with the line $y = x$.



(b) The criteria described in the table below can be used to classify people into one of three body shape categories: square, tall rectangle, or short rectangle.

Square	Tall Rectangle	Short Rectangle
Arm span is equal to height.	Arm span is less than height.	Arm span is greater than height.

(i) For which graph, 1 or 2, is the line helpful in classifying a student's body shape as square, tall rectangle, or short rectangle? Explain.

Graph 2 because points on line are 'square' points above the line are 'short', below are 'tall'

(ii) Complete the table of classifications for the 12 seniors.

Classification	Square	Tall Rectangle	Short Rectangle
Frequency	3	4	5

(c) Using the best model for prediction, calculate the predicted arm span for a senior with height 61 inches.

(Graph 1 is an LSRL) $\hat{y} = 11.74 + 0.8247(61) = 62.0467$ inches

(d) Interpret the slope for the model in graph 1 in the context of this problem.

slope $b = .8247 \frac{\text{arm inch}}{\text{height inch}}$

For every 1 additional inch in height, arm span increases by 0.8247 inches, on average.

(e) $r = 0.86$ Interpret r^2 (the coefficient of determination) in the context of this problem.

$r^2 = (.86)^2 = .7396$

About 74% of the variation in arm span is explained by the LSRL model which relates arm span to height.

Part 1: Multiple Choice. *Circle the letter corresponding to the best answer.*

(response) mileage

size (explanatory)

1. Other things being equal, larger automobile engines consume more fuel. You are planning an experiment to study the effect of engine size (in liters) on the gas mileage (in miles per gallon) of sport utility vehicles. In this study,
   (a) gas mileage is a response variable, and you expect to find a negative association.
   (b) gas mileage is a response variable, and you expect to find a positive association.
   (c) gas mileage is an explanatory variable, and you expect to find a strong negative association.
   (d) gas mileage is an explanatory variable, and you expect to find a strong positive association.
   (e) gas mileage is an explanatory variable, and you expect to find very little association.

2. In a statistics course, a linear regression equation was computed to predict the final-exam score from the score on the first test. The equation was $\hat{y} = 10 + 0.9x$ where y is the final-exam score and x is the score on the first test. Carla scored 95 on the first test. What is the predicted value of her score on the final exam?
   (a) 85.5 (b) 90 (c) 95 (d) 95.5 (e) none of these

$\hat{y} = 10 + 0.9(95)$

$\hat{y} = 95.5$

3. In the course described in #2, Bill scored a 90 on the first test and a 93 on the final exam. What is the value of his residual?
   (a) –2.0 (b) 2.0 (c) 3.0 (d) 93 (e) none of these

$\hat{y} = 10 + 0.9(90) = 91$

$y = 93$

$\text{resid} = y - \hat{y} = 93 - 91 = 2$

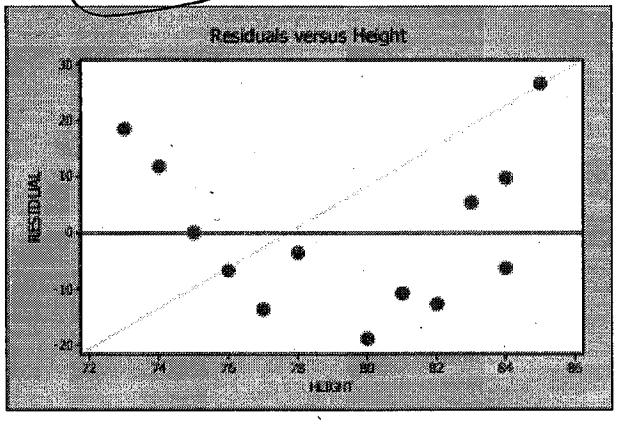
4. The correlation between the heights of fathers and the heights of their (fully grown) sons is $r = 0.52$. This value was based on both variables being measured in inches. If fathers' heights were measured in feet (one foot equals 12 inches), and sons' heights were measured in furlongs (one furlong equals 7920 inches), the correlation between heights of fathers and heights of sons would be
(a) much smaller than 0.52
(b) slightly smaller than 0.52
(c) unchanged: equal to 0.52
(d) slightly larger than 0.52
(e) much larger than 0.52

r is based on z-scores which are standardized so units don't matter (swapping x,y also doesn't) $\left(r = \sum \frac{z_x z_y}{n-1}\right)$

5. All but one of the following statements contains an error. Which statement could be correct?
(a) There is a correlation of 0.54 between the position a football player plays and his weight. ← can't have r w/ categorical variables
(b) We found a correlation of $r = -0.63$ between gender and political party preference. ←
(c) The correlation between the distance travelled by a hiker and the time spent hiking is $r = 0.9$ meters per second. r has no units
(d) We found a high correlation between the height and age of children: $r = 1.12$. r can only be −1 to 1
(e) The correlation between mid-August soil moisture and the per-acre yield of tomatoes is $r = 0.53$.

6. A set of data describes the relationship between the size of annual salary raises and the performance ratings for employees of a certain company. The least squares regression equation is $\hat{y} = 1400 + 2000x$ where y is the raise amount (in dollars) and x is the performance rating. Which of the following statements is *not necessarily* true?
✓(a) For each one-point increase in performance rating, the raise will increase on average by $2000.
(b) The actual relationship between salary raises and performance rating is linear. can't tell, this LSRL might have been fit to non-linear data
✓(c) A rating of 0 will yield a predicted raise of $1400.
✓(d) The correlation between salary raise and performance rating is positive. (yes, because slope is positive)
✓(e) If the average performance rating is 1.2, then the average raise is $3800. $\hat{y} = 1400 + 2000(1.2)$

7. A least-squares regression line for predicting weights of basketball players on the basis of their heights produced the residual plot below.



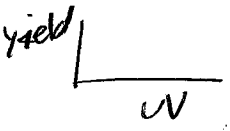
What does the residual plot tell you about the linear model?
✗(a) A residual plot is not an appropriate means for evaluating a linear model.
✗(b) The curved pattern in the residual plot suggests that there is no association between the weight and height of basketball players. could be a strong, but non-linear association
(c) The curved pattern in the residual plot suggests that the linear model is not appropriate.
✗(d) There are not enough data points to draw any conclusions from the residual plot. just a few are required
✗(e) The linear model is appropriate, because there are approximately the same number of points above and below the horizontal line in the residual plot. no, inappropriate due to pattern in residuals

Use the following to answer questions **8** and **9**.

One concern about the depletion of the ozone layer is that the increase in ultraviolet (UV) light will decrease crop yields. An experiment was conducted in a green house where soybean plants were exposed to varying levels of UV, measured in Dobson units. At the end of the experiment the yield (kg) was measured. A regression analysis was performed with the following results:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	3.9800118	0.053774	74.01	<.0001	3.8636398	4.0963838
uv	-0.046285	0.010741	* hidden *	0.0008	**** hidden ****	



8. The least-squares regression line is the line that

✗ (a) minimizes the sum of the distances between the actual UV values and the predicted UV values.

(b) minimizes the sum of the squared residuals between the actual yield and the predicted yield.

✗ (c) minimizes the sum of the distances between the actual yield and the predicted UV.

✗ (d) minimizes the sum of the squared residuals between the actual UV reading and the predicted UV values. × these are x values — residuals are differences in y values.

✗ (e) minimizes the perpendicular distance between the regression line and each data point.

residuals are vertical

9. Which of the following is correct?

✗ (a) If the UV value increases by 1 Dobson unit, the yield is expected to increase by 0.0463 kg.

✗ (b) If the yield increases by 1 kg, the UV value is expected to decrease by 0.0463 Dobson units.

(c) If the UV value increases by 1 Dobson unit, the yield is expected to decrease by 0.0463 kg.

✗ (d) The predicted yield is 4.3 kg when the UV value is 20 Dobson units. $\hat{y} = 3.9800118 - .046285(20)$ $= 3.0543$

(e) None of the above is correct.

10. Which statements below about least-squares regression are correct?

✗ I. Switching the explanatory and response variables will not change the least-squares regression line. Slope will change: $b = r\frac{s_x}{s_y}$ is not same as $b = r\frac{s_y}{s_x}$

✗ II. The slope of the line is very sensitive to outliers with large residuals.

✓ III. A value of r^2 close to 1 does not guarantee that the relationship between the variables is linear.

not always... only if they have high leverage

(a) Only I is correct.

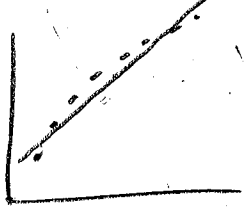
(b) Only II is correct.

(c) Only III is correct.

(d) Both II and III are correct.

(e) All three statements—I, II, and III—are correct.

true. You can have non-linear data that is only slightly curved...



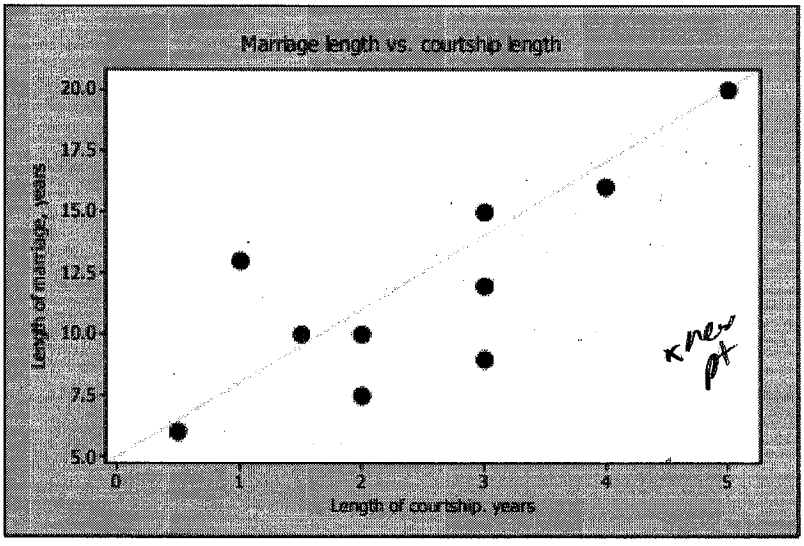
so the LSRL is close all points (r^2 close to 1) but data is still non-linear.

Part 2: Free Response

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

Questions 11-15 relate to the following.

A certain psychologist counsels people who are getting divorced. A random sample of ten of her patients provided the data in the following scatterplot, where x = number of years of courtship before marriage, and y = number of years of marriage before divorce.



11. Describe what the scatterplot reveals about the relationship between length of courtship and length of marriage.

there is a medium-strength, positive, linear relationship between length of courtship and length of marriage.

12. Suppose a new point at (4.5, 8), that is, years of courtship = 4.5 and years of marriage = 8, were added to the plot. What effect, if any, will this new point have on the correlation between courtship duration and marriage duration? Explain.

It would weaken the correlation (lower the r value) because points would now be further from the LSRL, on average.

The effect may be small, though, because the point is close to alignment with the centroid in the y-direction (z_y is close to 0) so it numerically doesn't contribute much in $r = \sum \frac{z_x z_y}{n-1}$

Below is the computer output for the regression of length of marriage *versus* length of courtship.

Predictor	Coef	SE Coef	T	P
Constant	5.710	1.880	3.04	0.016
courtship	2.4559	0.6669	3.68	0.006

S = 2.74982 R-Sq = 62.9% R-Sq(adj) = 58.3%

$\hat{y} = 5.710 + 2.4559x$
x: courtship (yrs)
y: marriage (yrs)

13. What is the slope of the regression line? Interpret the slope in the context of this problem.

$b = 2.4559 \frac{\text{marriage years}}{\text{courtship years}}$

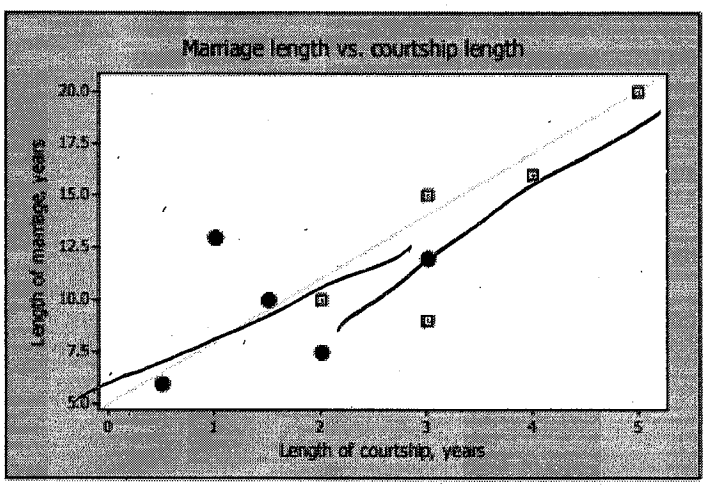
For every 1 additional year of courtship, marriage length increases by 2.4559 years, on average.

14. Explain what the quantity S = 2.74982 measures in the context of this problem.

S is the standard deviation of the residuals.

The difference between actual marriage length and predicted marriage length (for given courtship lengths) is 2.74982 years, on average.

15. The psychologist is curious about whether having children has an impact on this relationship. She draws a second scatterplot, with those couples who have children as open squares and couples without children as closed circles.



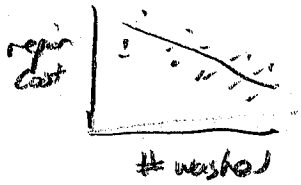
Comment on the impact that having children has on the relationship between length of courtship and length of marriage for these patients.

The marriages with children had higher lengths of both courtship and marriage than the marriages without children.

A rough sketch of separate LSRLs for these subgroups shows the slope for couples with children may be a little higher, but the correlation values still seem about the same (splitting these groups and analyzing separately probably wouldn't produce stronger predictive models)

One weekend, a statistician notices that some of the cars in his neighborhood are very clean and others are quite dirty. He decides to explore this phenomenon, and asks 15 of his neighbors how many times they wash their cars each year and how much they paid in car repair costs last year. His results are in the table below:

	Mean	Standard deviation
x = number of car washes per year	6.4	3.78
y = repairs costs for last year	$955.30	$323.50



The correlation for these to two variables is $r = -0.71$

16. Find the equation of the least-squares regression line (with y as the response variable).

$b = r\frac{s_y}{s_x}$

$b = (-.71)\frac{323.50}{3.78}$

$b = -60.7632$

$\hat{y} = a - 60.7632x$

centroid $(6.4, 955.30)$ on LSRL, so:

$(955.30) = a - 60.7632(6.4)$

$a = 1344.1845$

$\hat{y} = 1344.1845 - 60.7632x$

x: # car washes

y: repair costs ($)

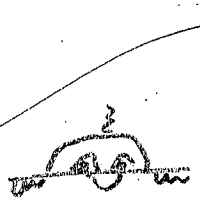
17. What percentage of the variation in repair costs can be explained by the number of times per year a car is washed?

$r^2 = (-.71)^2 = .5041$

About 50% of the variation in repair costs can be explained by the LSRL model which relates repair costs to number of times per year a car is washed.

18. Based on these data, can we conclude that washing your car frequently will reduce repair costs? Explain.

NO. Even though there is an association here, correlation does not mean causation. There may be lurking variables causing these factors to seem linked.



do this (and provide an example of what the lurking variable may be for extra credit