AP Statistics
Unit 02 – **FINAL REVIEW**

Name Key
Period____

Multiple Choice

1. A least-squares regression line for predicting performance on a college entrance exam based on high school grade point average (GPA) is determined to be $\widehat{score} = 273.5 + 91.2\,(GPA)$. One student in the study had a high school GPA of 3.0 and an exam score of 510. What is the residual for this student?

C

   a. 26.2
   b. 43.9
   c. -37.1
   d. -26.2
   e. 37.1

$\widehat{score} = 273.5 + 91.2(3.0) = 547.1$

residual $= y - \hat{y} = 510 - 547.1 = -37.1$

2. What is the regression equation for predicting weight from height in the following computer printout?

The regression equation is
$\widehat{weight}$ = ________ + ________ (height)

Predictor	Coeff	St Dev	*t*-ratio	*P*
Constant	-104.64	39.19	-2.67	0.037
Height	3.4715	0.5990	5.80	0.001
S = 7.936	R-sq = 84.8%		R-sq(adj) = 82.3%	

A

   a. $\widehat{weight}$ = -104.64 + 3.4715 (height)
   b. $\widehat{weight}$ = 3.4715 – 104.64 (height)
   c. $\widehat{weight}$ = -104.64 + 39.19 (height)
   d. $\widehat{weight}$ = 3.4715 + 0.5990 (height)
   e. $\widehat{weight}$ = 39.19 – 104.64 (height)

3. Using the computer printout from question #2, what is the value of r?

B

   a. r = 0.848
   b. r = 0.921
   c. r = -0.921
   d. r = -0.823
   e. r = 0.823

the slope is positive so r is positive
$\sqrt{R^2} = \sqrt{84.8\%}$
$r = +0.921$

4. Given a set of ordered pairs (x, y) so that $s_x = 1.6$, $s_y = 0.75$, r = 0.55. What is the slope of the least-squared regression line for these data.

D

   a. 1.82
   b. 1.17
   c. 2.18
   d. 0.26
   e. 0.78

$b = r\frac{s_y}{s_x} = 0.55 \cdot \frac{0.75}{1.6} = 0.258$

5. A study found a correlation of r = -0.58 between hours per week spent watching television and hours per week spent exercising. That is, the more hours spent watching television, the less hours spent exercising per week. Which of the following statements is most accurate? $R^2 = (-0.58)^2 = 0.3364$

A

a. About one-third of the variation in hours spent exercising can be explained by hours spent watching television. R^2 sentence frame
b. A person who watches less television will exercise more.
c. For each hour spent watching television, the predicted decrease in hours spent exercising is 0.58 hours.
d. There is a cause-and-effect relationship between hours spent watching television and a decline in hours spent exercising.
e. 58% of the hours spent exercising can be explained by the number of hours watching television.

6. A study was done on the relationship between high school grade point average (GPA) and scores on the SAT. The following 8 scores were from a random sample of students taking the exam:

X (GPA)	3.2	3.8	3.9	3.3	3.6	2.8	2.9	3.5
Y (SAT)	725	752	745	680	700	562	595	730

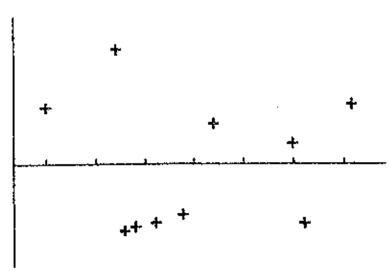
What percent of the variation in SAT scores is explained by the regression of SAT score on GPA? R^2 sentence frame

E

$R^2 = 0.7881$

a. 62.1%
b. 72.3%
c. 88.8%
d. 94.2%
e. 78.8%

7. Consider the following residual plot:



random scatter/no pattern

C

Which of the following statements is (are) true?
I. The residual plot indicates that a line is a reasonable model for the data.
~~II.~~ The residual plot indicates that there is no relationship between the data.
III. The correlation between the variables is probably non-zero.

a. I only
b. II only
c. I and III only
d. II and III only
e. I and II only

8. Consider the following data set:

x	45	73	82	91
y	15	7.9	5.8	3.5

Given that the LSRL for these data is y = 26.211 – 0.25x, what is the value of the residual for 73?

D

$y = 26.211 - 0.25(73) = 7.961$

a. 7.961
b. 0.061
c. 36.561
d. -0.061
e. -7.961

residual $= y - \hat{y} = 7.9 - 7.961 = -0.061$

9. Using the data table and regression line from question #8, which of the following are true about the point (91, 3.5)?

residual $= 3.5 - 3.461 = 0.039$

A

a. It is above the regression line because the residual is positive.
b. It is above the regression line because the residual is negative.
c. It is below the regression line because the residual is positive.
d. It is below the regression line because the residual is negative.
e. It is on the regression line because the residual is zero.

$\hat{y} = 26.211 - 0.25(91) = 3.461$

10. Measurements are made of the number of cockroaches present, on average, every 3 days, beginning on the second day, after apartments in one part of town are vacated. The data are as follows:

Days	2	5	8	11	14
# Roaches	5.69	11.96	18.23	24.5	30.77

enter into L1 and L2 stat > calc > linreg

How many cockroaches would you expect to be present after 9 days?

B

a. 20.23
b. 20.32
c. 21.33
d. 21.38
e. 23.41

$\hat{y} = 1.51 + 2.09x$

$\hat{y} = 1.51 + 2.09(9) = 20.32$

11. Mrs. De Marre is studying how the length of rides at her favorite theme park affects wait times for the rides. Her findings result in a two-variable data set such that $\bar{x} = 2.5$ minutes, $\bar{y} = 4.2$ minutes, $s_x = 0.82$ minutes, $s_y = 1.54$ minutes, and $r = 0.89$.

a. Determine the equation of the LSRL of wait times on ride length.

$\hat{y} = a + bx$

$a = \bar{y} - b\bar{x}$

$a = 4.2 - 1.671 \cdot 2.5$

$a = 0.02134$

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

$b = 0.89 \cdot \frac{1.54}{0.82}$

$b = 1.671$

$\boxed{\widehat{\text{wait time}} = 0.02134 + 1.671(\text{length of ride})}$

b. What is the value of R^2 and what does it mean in context?

$r = 0.89$

$R^2 = 0.7921$

79.21% of the variation in wait time is accounted for by the LSRL of wait time on ride length.

c. One of Mrs. De Marre's favorite rides is 3.4 minutes long with a wait time of 5.2 minutes. What is the predicted wait time for this ride?

x y

$\hat{y} = 0.02134 + 1.671(3.4) = 5.704$

The predicted wait time for this 3.4-minute ride is 5.704 minutes.

d. Calculate the value of the residual from part (c).

residual $= y - \hat{y} = 5.2 - 5.704 = -0.504$ minutes.

The wait time for this 3.4-minute ride is about half a minute less than predicted.

12. The regional champion in 10 and under 100 m backstroke has had the following winning times (in seconds) over the past 8 years:

Year	1	2	3	4	5	6	7	8
Time	77.3	80.2	77.1	76.4	75.5	75.9	75.1	74.3

How many years until you expect the winning time to be one minute or less? What's wrong with this estimate?

60 seconds

$\widehat{time} = 79.207 - 0.6071(year)$

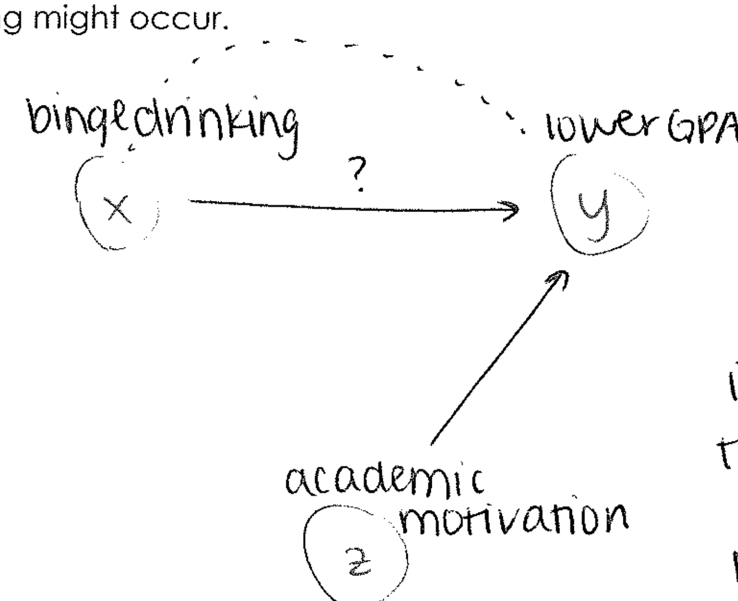
$60 = 79.207 - 0.6071(year)$

$\boxed{year = 31.64}$

The estimated # of years until the predicted winning time will be one minute or less is 31.64 years.

However, this is <u>extrapolation</u> because it is not within our domain of x-values and therefore cannot be predicted by our model.

13. Effects of binge drinking A common definition of "binge drinking" is 5 or more drinks at one sitting for men and 4 or more for women. An observational study finds that students who binge drink have lower average GPA than those who don't. Identify a variable that may be confounded with the effects of binge drinking. Explain how confounding might occur.



Academic motivation may be a confounding variable. If a student is a binge drinker but also has little or no academic motivation, it's impossible to separate the effects of each on the student's GPA. No academic motivation may cause students to do poorly in school because they just don't care.

14. The table of data shows the consumer price of a product compared to production cost of that product.

Production cost $	**$0.31**	**$0.32**	**$0.35**	**$0.41**	**$0.58**	**$0.87**	**$0.96**	**$1.21**	**$1.59**
Consumer cost $	$9.80	$9.85	$10.01	$10.33	$11.31	$13.19	$13.83	$15.79	$19.31
Production cost $	**$2.31**	**$2.40**	**$4.51**	**4.52**	**$4.80**	**$4.89**	**$5.00**	**$6.20**	**$7.75**
Consumer cost $	$28.27	$29.65	$90.64	$91.12	$105.68	$110.84	$117.49	$221.82	$504.08

a. Does a ~~linear model~~ the scatterplot of production cost vs. consumer cost seem appropriate for this data? Why or why not?

no! it is very curved. we should look at other models:

linear: curved r=0.845

exponential: straight r= 0.9999

power: curved r=0.9355

An exponential model is the best fit because the transformed data appears linear and r is nearly=1.

b. Find an equation of the LSRL for consumer cost on production cost.

$\widehat{\log consumer cost} = 0.91998 + 0.23001(production cost)$

$\boxed{\widehat{consumer cost} = 8.3173 \cdot 1.698^{production cost}}$

c. Express what R^2 is in context.

99.99% of the variation in consumer cost is accounted for by the LSRL of consumer cost on production cost.

d. Estimate the consumer cost of a product whose production cost is $4.51?

$\widehat{consumer cost} = 8.3173 \cdot 1.698^{4.51} = \$90.57$

The expected consumer cost (for a product whose production cost is $4.51) is $ 90.57.

e. Calculate the residual for the product in part d.

residual= $y - \hat{y} = 90.64 - 90.57 = \$0.07$

The actual consumer cost of a product (whose production cost is $4.51) is $0.07 higher than predicted.