Name: _____

PART 2 PRACTICE QUESTIONS

MUTIPLE CHOICE

Recommended Time – 45 minutes

Number of multiple choice questions – 20

Directions: Solve each of the following problems, using the available space for scratch work. Decide which is the best of the choices given and circle the corresponding letter on the answer sheet. No credit will be given for anything written in the test book. Do not spend too much time on any one problem.

1. Consider n pairs of numbers (x_1, y_1), (x_2, y_2), …, and (x_n, y_n). The mean and standard deviation of the x-values are $\bar{x} = 9$ and $s_x = 9$, respectively. The mean and standard deviaton of the y-values are $\bar{y} = 5$ and $s_y = 2$, respectively. Of the following, which could be the least squares regression line?

(A) $\hat{y} = -35.5 + 4.5x$
(B) $\hat{y} = 4.5x$
(C) $\hat{y} = 3.2 + 0.2x$
(D) $\hat{y} = 5 + 0.2x$
(E) $\hat{y} = 45 + 18x$

2. A least squares regression line was fitted to the weekly cost of groceries in dollars (cost) versus the number of household members (number) for a group of families. The resulting equation is

$$\widehat{cost} = -33.22 + 44.77(number).$$

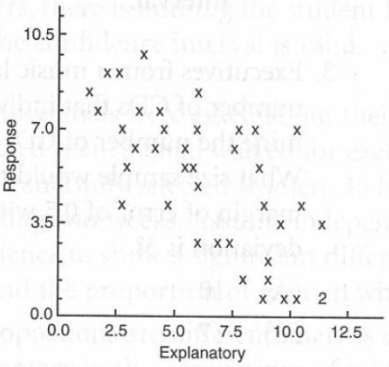
A randomly selected family of four spends $135 on groceries in an average week. What is the difference between this family's actual cost and the predicted average family cost?

(A) -$10.86
(B) -$0.24
(C) $0.24
(D) $10.86
(E) $145.86

3. The correlation between the depth (in feet) and length (in feet) for a sample of caves is found to be -0.346. If the measurement of depth is converted to meters. What will be the resulting correlation? (1 ft. = 0.3048 m)

(A) -0.627
(B) -0.346
(C) -0.105
(D) 0.105
(E) 0.346

4. For the given scatterplot, what is the correct regression output?



(A)

Predictor	Coef	SE Coef	T	P
Constant	9.1104	0.7745	11.76	0.000
Explanatory	0.6181	0.1085	−5.70	0.000
S = 2.02551	R-Sq = 46.1%	R-Sq(adj) = 44.7%		

(B)

Predictor	Coef	SE Coef	T	P
Constant	9.1104	0.7745	11.76	0.000
Explanatory	−0.6181	0.1085	−5.70	0.000
S = 2.02551	R-Sq = 67.9%	R-Sq(adj) = 66.3%		

(C)

Predictor	Coef	SE Coef	T	P
Constant	9.1104	0.7745	11.76	0.000
Explanatory	0.6181	0.1085	−5.70	0.000
S = 2.02551	R-Sq = 67.9%	R-Sq(adj) = 66.3%		

(D)

Predictor	Coef	SE Coef	T	P
Constant	9.1104	0.7745	11.76	0.000
Explanatory	−0.6181	0.1085	−5.70	0.000
S = 2.02551	R-Sq = 46.1%	R-Sq(adj) = 44.7%		

(E)

Predictor	Coef	SE Coef	T	P
Constant	−0.6181	0.7745	11.76	0.000
Explanatory	9.1104	0.1085	−5.70	0.000
S = 2.02551	R-Sq = 46.1%	R-Sq(adj) = 44.7%		

5. The coefficient of determination between two variables is computed to be 81%. Which of the following must be true?

(A) Large values of the explanatory variable correspond with large values of the response variable.
(B) Large values of the explanatory variable correspond with small values of the response variable.
(C) A cause-and-effect relationship exists between the explanatory and response variables.
(D) There is a strong, positive, linear relationship between the explanatory and response variables.
(E) Approximately 81% of the variability in the response variable is explained by linear regression on the explanatory variable.

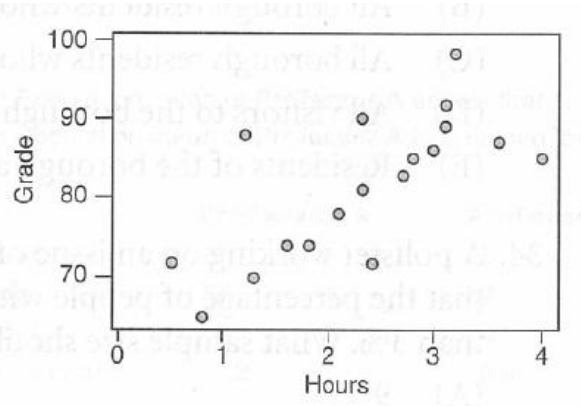
6. A regression equation is given as $\log \hat{y} = 0.214 - 1.28x$. What is the (approximate) predicted value for y when x = 2?

(A) -2.346
(B) -0.171
(C) 0.005
(D) 0.167
(E) Cannot be determined

7. Two measures, x and y, are taken on numerous subjects, and a least squares regression equation si computed.
The resulting equation is $\hat{y} = 382.1 - 12.25x$. A correct interpretation for the slope and intercept is

(A) For every increase of 100 units in x, y increases by approximately 1225 units; when x = 0, y is predicted to be 382.1.
(B) For every increase of 100 units in x, y increases approximately 1225 units; when x = 0, y is predicted to be 382.1.
(C) For every increase of 100 units in x, y decreases approximately 1225 units; when x = 0, y is predicted to be 382.1.
(D) For every increase of 100 units in x, y increases approximately 32,810 units; when x = 0, y is predicted to be 12.25.
(E) For every increase of 100 units in x, y increases approximately 32, 810 units; when x = 0, y is predicted to be -12.25.

8. Statistics students decided to analyze the relationship between hours spent studying per night and average grades on tests. Their data display is shown below.



The correlation coefficient for the least squares regression line is most nearly

(A) 0.95
(B) 0.70
(C) -0.55
(D) -0.30
(E) 0.10

9. Which of the statements listed below is/are correct?

I. The slope of a regression line can be calculated from the formula $b = \frac{r \cdot s_y}{s_x}$.
II. Residual = actual value – fitted value.
III. Causation is demonstrated by the correlation coefficient.

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

10.

Regression analysis is performed on two variables, X and Y. A residual plot is graphed, and summary statistics are found for the residuals.

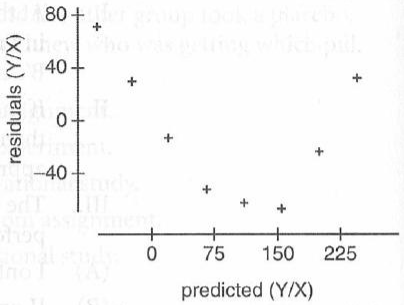
Dependent variable is: Y
No Selector
R squared = 83.0% R squared (adjusted) = 80.5%
s = 59.67 with 9 − 2 = 7 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	121320	1	121320	34.1
Residual	24921.5	7	3560.21	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	−113.056	43.35	−2.61	0.0350
X	44.9667	7.703	5.84	0.0006

Summary of residuals(Y/X)
No Selector

Count	9
Mean	−42.6326e−15
Median	−11.8444
MidRange	5.80556
StdDev	55.8139
Range	145.1
IntQRange	97.3167



What tells you there is a better fit for the data?

(A) The mean residual value is 0.

(B) A significant relationship does not exist for X and Y.

(C) There is a low correlation between X and Y.

(D) There is a pattern in the residual plot.

(E) The linear regression model for X and Y is appropriate; a better fit does not exist.

11.

The following information resulted from regression analysis for the percentage of white children under the age of 18 versus the percentage of black children under the age of 18 who live in poverty in several large cities.

Dependent variable is: Black
No Selector
R squared = 79.5% R squared (adjusted) = 77.8%
s = 1.072 with 14 − 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	53.5882	1	53.5882	46.6
Residual	13.7889	12	1.14907	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	30.4954	1.942	15.7	≤ 0.0001
White	0.911941	0.1335	6.83	≤ 0.0001

If the data were reevaluated using the percentage of white children as the dependent variable, the correlation coefficient would be which of the following?

(A) 0.778

(B) 0.795

(C) 0.882

(D) 0.892

(E) Not enough information is given to calculate the correlation coefficient.

12.

The residuals for a complete data set are shown below, and r^2 for the least squares regression line that resulted in these residuals is 88.6%.

x	25	30	35	40	45	50	55	60	65	70
Residual	0	1.2	2.0	0.97	0.9	−1.1	−4.6	−0.6	0.25	0.98

Which of the following is/are true?

I. The linear model is a good model for the data.
II. The sum of the squares of the residuals is zero.
III. The correlation is either $\pm\sqrt{0.886}$.

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

13. A researcher was interested in determining the relationship between pulse rate (in beats/minute) and the time (in minutes) it took to swim a fixed distance. Based on 25 trials in the pool, the correlation coefficient between time and pulse rate was found to be -0.654 (that is, large times – going slowly – were associated with slower pulses). Prior to publication, the researcher decided to change the time measurement to seconds (each of the 25 times was multiplied by 60). What would this conversion do to the correlation between the two variables?

(A) Since the units on only one of the variables was changed, the correlation between the two variables would decrease.
(B) The correlation would change proportional to the change in the units for time.
(C) The correlation would change, but there is now ay, based on the information given, to know by how much.
(D) Changing the units of measurement has no effect on the correlation coefficient. Hence, the correlation would be the same.
(E) Since changing from minutes to seconds would result in larger times, the correlation would actually increase.

14. Which of the following is *not* a property of the correlation coefficient

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)?$$

(A) r is not a function of the unites used for the variables.
(B) r can be calculated from either categorical or numerical values.
(C) R is not affected by which variable is called x and which variable is called y.
(D) $|r| > 1$
(E) r is positive when the slope of the regression line is positive and negative when the slope of the regression line is negative.

15. For the graph given above, which of the following statements is (are) true?

I. The point marked with the "X" is better described as an outlier than as an influential point.
II. Removing the point "X" would cause the correlation to increase.
III. Removing the point "X" would have a significant effect on the slope of the regression line.

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

16.

The relation between the selling price of a car (in $1,000) and its age (in years) is estimated from a random sample of cars of a specific model. The relation is given by the following formula:

$$\text{Selling price} = 15.9 - 0.983\ (\text{age})$$

Which of the following can we conclude from this equation?

(A) For every year the car gets older, the selling price goes down by approximately 9.83 percent.
(B) A new car costs on the average $9,830.
(C) For every year the car gets older, the selling price drops by approximately $1,590.
(D) A new car costs $16,883.
(E) For every year the car gets older, the selling price drops by approximately $983.

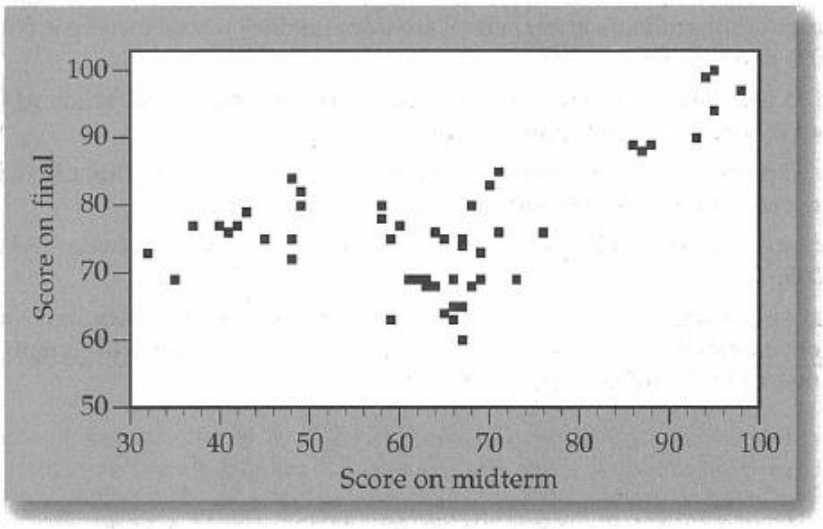
17.

Data was collected on two variables X and Y and a least squares regression line was fitted to the data. The estimated equation for this data is $y = -2.29 + 1.70x$. What is the residual for point (5, 6)?

(A) 7.91
(B) 6.21
(C) 0.21
(D) –0.21
(E) –2.91

18.

Extra study sessions were offered to students after the midterm to help improve their understanding of statistics. Student scores on the midterm and the final exam were recorded. The following scatterplot shows final test scores against the midterm test scores.



Which of the following statements correctly interprets the scatterplot?

(A) All students have shown significant improvement in the final exam scores as a result of the extra study sessions.

(B) The extra study sessions were of no help. Each student's final exam score was about the same as his or her score on the midterm.

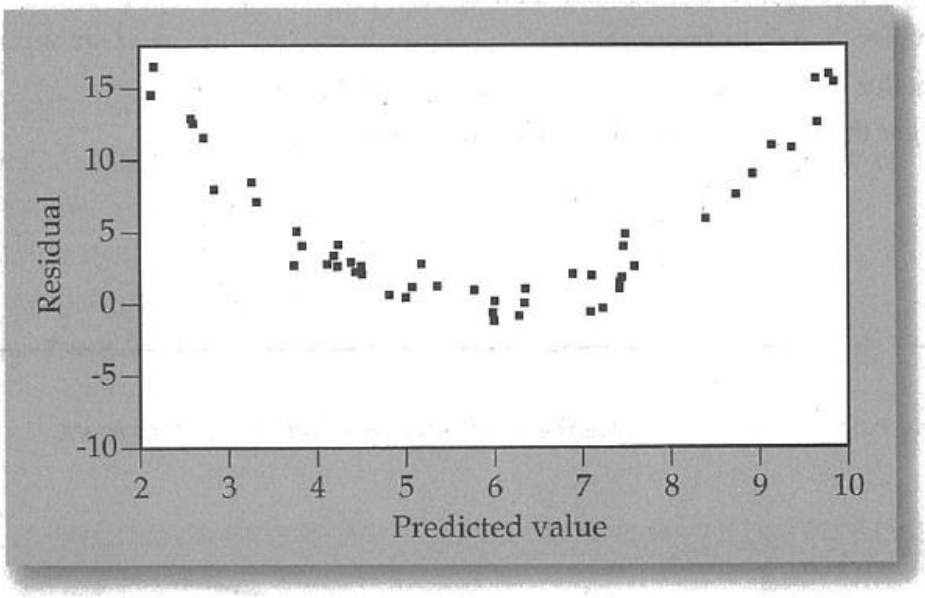
(C) The extra study sessions further confused students. All student scores decreased from midterm to final exam.

(D) Students that scored below 55 on the midterm showed considerable improvement on the final exam; those who scored between 55 and 80 on the midterm showed minimal improvement on the final exam; and those who scored above 80 on the midterm showed almost no improvement on the final exam.

(E) Students that scored below 55 on the midterm showed minimal improvement on the final exam; those who scored between 55 and 80 on the midterm showed moderate improvement on the final exam; and those who scored above 80 on the midterm showed considerable improvement on the final exam.

19.

Sixty pairs of measurements were taken at random to estimate the relation between variables X and Y. A least squares regression line was fitted to the collected data. The resulting residual plot is as follows:



Which of the following conclusions is appropriate?

(A) A line is an appropriate model to describe the relation between X and Y.
(B) A line is not an appropriate model to describe the relation between X and Y.
(C) The assumption of normality of errors has been violated.
(D) The assumption of constant sample standard deviations has been violated.
(E) The variables X and Y are not related at all.

20.

The relation between car speed (in miles per hour) and gas consumption (in miles per gallon) was estimated from a random sample to be

$$\text{gas consumption} = 39.43 - 0.18\ (\text{speed})$$

How will gas consumption change with a 10 mile per hour increase in the car's speed?

(A) Gas consumption will increase by 37.63 mpg.
(B) Gas consumption will increase by 1.8 mpg.
(C) Gas consumption will remain unchanged.
(D) Gas consumption will decrease by 1.8 mpg.
(E) Gas consumption will decrease by 37.63 mpg.

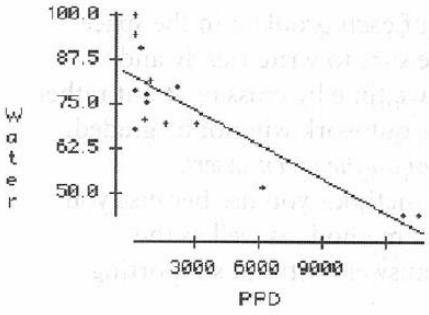
FREE RESPONSE

Questions 1-3

Spend about 45 minutes on this part of the exam.

Directions: 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.

1. The graph shows the percentage of the population with access to safe drinking water (Water) and the persons per doctor (PPD) for randomly selected regions of the world. Also given is the regression analysis for these variables.



Dependent variable is: **Water**
No Selector
R squared = 70.4% R squared (adjusted) = 68.3%
s = 9.58 with 16 - 2 = 14 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3060.62	1	3060.62	33.3
Residual	1284.82	14	91.773	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	83.4851	2.838	29.4	≤ 0.0001
Doctors	-0.00323837	0.0005608	-5.77	≤ 0.0001

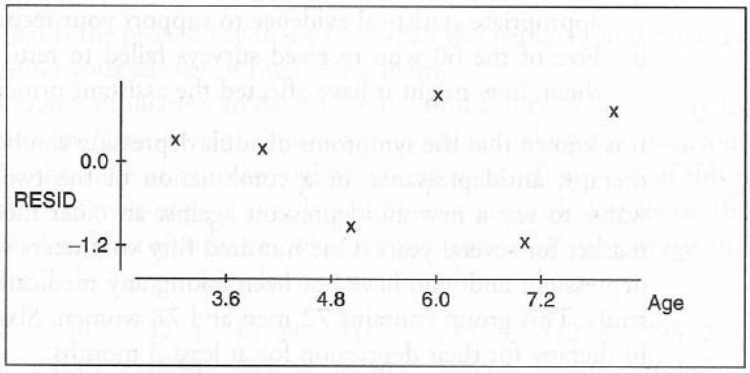
a. What is the equation of the least squares regression line that describes the relationship between the percentage with access to safe drinking water and the persons per doctor? Define any variables used in this equation.

b. Record *and* interpret the value of the correlation coefficient for percentage with access to safe drinking water and persons per doctor.

c. Record *and* interpret the value of the coefficient of determination for percentage with access to safe drinking water and persons per doctor.

2.

David was comparing the number of vocabulary words children know about transportation at various ages. He fit a least-squares regression line to the data. The residual plot and part of the computer output for the regression are given below.



Predictor	Coef	St Dev	*t* ratio	*P*
Constant	3.371	1.337	2.52	.065
Age	2.1143	0.2321	9.11	.001

s = 0.9710 R-sq = 95.4% R-sq(adj) = 94.3%

a. Is a line an appropriate model for these data? Explain.

b. What is the equation of the least-square regression line for predicting the number of words from age?

c. What is the predicted number of words for a child of 7.5 years of age?

d. Interpret the slope of the regression line in the context of the problem.

e. Would it be appropriate to use the model to predict the number of words a 12-year-old would know?

3.

A store that sells cellular telephones opened for business twelve years ago. Total sales of cellular telephones for each year of business are given.

Year	1	2	3	4	5	6	7	8	9	10	11	12
Units Sold	375	566	671	1106	1311	1283	2136	2967	4094	4572	5157	6621

a. Record and interpret the slope of the least squares regression line that describes the relationship between the number of cellular telephones sold and the business year.

b. Record and interpret the value of the correlation coefficient for the year of business and the number of cell phones sold.

c. Use this model to predict the number of cell phones the store will sell in the next business year.

d. Describe any shortcomings you see in this model.

e. Find a better model for predicting the number of units sold for a given business year. Justify your choice, and use three decimal places for your slope and intercept coefficients.

f. Use your model from part (e) to predict the number of cell phones the store will sell in the next business year.