13.11: Practice

13. The Correlation Coefficient r

1. In order to have a correlation coefficient between traits A and B, it is necessary to have:

a. one group of subjects, some of whom possess characteristics of trait A, the remainder possessing those of trait B
b. measures of trait A on one group of subjects and of trait B on another group
c. two groups of subjects, one which could be classified as A or not A, the other as B or not B
d. two groups of subjects, one which could be classified as A or not A, the other as B or not B

2. Define the Correlation Coefficient and give a unique example of its use.
3. If the correlation between age of an auto and money spent for repairs is +.90

a. 81% of the variation in the money spent for repairs is explained by the age of the auto
b. 81% of money spent for repairs is unexplained by the age of the auto
c. 90% of the money spent for repairs is explained by the age of the auto
d. none of the above

4. Suppose that college grade-point average and verbal portion of an IQ test had a correlation of .40. What percentage of the variance do these two have in common?

a. 20
b. 16
c. 40
d. 80

5. True or false? If false, explain why: The coefficient of determination can have values between -1 and +1.
6. True or False: Whenever r is calculated on the basis of a sample, the value which we obtain for r is only an estimate of the true correlation coefficient which we would obtain if we calculated it for the entire population.
7. Under a "scatter diagram" there is a notation that the coefficient of correlation is .10. What does this mean?

a. plus and minus 10% from the means includes about 68% of the cases
b. one-tenth of the variance of one variable is shared with the other variable
c. one-tenth of one variable is caused by the other variable
d. on a scale from -1 to +1, the degree of linear relationship between the two variables is +.10

8. The correlation coefficient for X and Y is known to be zero. We then can conclude that:

a. X and Y have standard distributions
b. the variances of X and Y are equal
c. there exists no relationship between X and Y
d. there exists no linear relationship between X and Y
e. none of these

9. What would you guess the value of the correlation coefficient to be for the pair of variables: "number of man-hours worked" and "number of units of work completed"?

a. Approximately 0.9
b. Approximately 0.4
c. Approximately 0.0
d. Approximately -0.4
e. Approximately -0.9

10. In a given group, the correlation between height measured in feet and weight measured in pounds is +.68. Which of the following would alter the value of r?

a. height is expressed centimeters.
b. weight is expressed in Kilograms.
c. both of the above will affect r.
d. neither of the above changes will affect r.

13.3 Testing the Significance of the Correlation Coefficient

11. Define a t Test of a Regression Coefficient, and give a unique example of its use.
12. The correlation between scores on a neuroticism test and scores on an anxiety test is high and positive; therefore

a. anxiety causes neuroticism
b. those who score low on one test tend to score high on the other.
c. those who score low on one test tend to score low on the other.
d. no prediction from one test to the other can be meaningfully made.

13.4 Linear Equations

13. True or False? If False, correct it: Suppose a 95% confidence interval for the slope β of the straight line regression of Y on X is given by -3.5 < β < -0.5. Then a two-sided test of the hypothesis H0:β=−1 would result in rejection of H0 at the 1% level of significance.
14. True or False: It is safer to interpret correlation coefficients as measures of association rather than causation because of the possibility of spurious correlation.
15. We are interested in finding the linear relation between the number of widgets purchased at one time and the cost per widget. The following data has been obtained:

X: Number of widgets purchased – 1, 3, 6, 10, 15

Y: Cost per widget(in dollars) – 55, 52, 46, 32, 25

Suppose the regression line is $\hat{y} = -2.5x + 60$. We compute the average price per widget if 30 are purchased and observe which of the following?

a. $\hat{y} = 15$ dollars; obviously, we are mistaken; the prediction $\hat{y}$ is actually +15 dollars.
b. $\hat{y} = 15$ dollars, which seems reasonable judging by the data.
c. $\hat{y} = -15$ dollars, which is obvious nonsense. The regression line must be incorrect.
d. $\hat{y} = -15$ dollars, which is obvious nonsense. This reminds us that predicting Y outside the range of X values in our data is a very poor practice.

16. Discuss briefly the distinction between correlation and causality.
17. True or False: If r is close to + or -1, we shall say there is a strong correlation, with the tacit understanding that we are referring to a linear relationship and nothing else.

13.5 The Regression Equation

18. Suppose that you have at your disposal the information below for each of 30 drivers. Propose a model (including a very brief indication of symbols used to represent independent variables) to explain how miles per gallon vary from driver to driver on the basis of the factors measured.

Information:

1. miles driven per day
2. weight of car
3. number of cylinders in car
4. average speed
5. miles per gallon
6. number of passengers

19. Consider a sample least squares regression analysis between a dependent variable (Y) and an independent variable (X). A sample correlation coefficient of −1 (minus one) tells us that

a. there is no relationship between Y and X in the sample
b. there is no relationship between Y and X in the population
c. there is a perfect negative relationship between Y and X in the population
d. there is a perfect negative relationship between Y and X in the sample.

20. In correlational analysis, when the points scatter widely about the regression line, this means that the correlation is

a. negative.
b. low.
c. heterogeneous.
d. between two measures that are unreliable.

13.6 Interpretation of Regression Coefficients: Elasticity and Logarithmic Transformation

21. In a linear regression, why do we need to be concerned with the range of the independent (X) variable?
22. Suppose one collected the following information where X is diameter of tree trunk and Y is tree height.

X	Y
4	8
2	4
8	18
6	22
10	30
6	8

Regression equation: y^i=−3.6+3.1·Xi $y\hat{}i=-3.6+3.1\cdot Xi$

What is your estimate of the average height of all trees having a trunk diameter of 7 inches?

23. The manufacturers of a chemical used in flea collars claim that under standard test conditions each additional unit of the chemical will bring about a reduction of 5 fleas (i.e. where Xj=amount of chemical Xj=amount of chemical and $X_j =$ amount of chemical and $Y_J = B_0 + B_1 \cdot X_J + E_J, H_0 : B_1 = -5$

Suppose that a test has been conducted and results from a computer include:

Intercept = 60

Slope = −4

Standard error of the regression coefficient = 1.0

Degrees of Freedom for Error = 2000

95% Confidence Interval for the slope −2.04, −5.96

Is this evidence consistent with the claim that the number of fleas is reduced at a rate of 5 fleas per unit chemical?

13.7 Predicting with a Regression Equation

24. True or False? If False, correct it: Suppose you are performing a simple linear regression of Y on X and you test the hypothesis that the slope β is zero against a two-sided alternative. You have n=25 observations and your computed test (t) statistic is 2.6. Then your *P*-value is given by .01 < *P* < .02, which gives borderline significance (i.e. you would reject H0 at α=.02 but fail to reject H00 at α=.011).
25. An economist is interested in the possible influence of "Miracle Wheat" on the average yield of wheat in a district. To do so he fits a linear regression of average yield per year against year after introduction of "Miracle Wheat" for a ten year period.

The fitted trend line is

$$\hat{y}_j = 80 + 1.5 \cdot X_j$$

(Y_j : Average yield in j year after introduction)
(X_j : j year after introduction).

a. What is the estimated average yield for the fourth year after introduction?
b. Do you want to use this trend line to estimate yield for, say, 20 years after introduction? Why? What would your estimate be?

26. An interpretation of r=0.5 is that the following part of the Y-variation is associated with which variation in X:

a. most
b. half
c. very little
d. one quarter
e. none of these

27. Which of the following values of *r* indicates the most accurate prediction of one variable from another?

a. r=1.18
b. r=−.77
c. r=.68

13.8 How to Use Microsoft Excel® for Regression Analysis

28. A computer program for multiple regression has been used to fit $y_j=b_0+b_1 \cdot X_{1j}+b2 \cdot X_{2j}+b_3 \cdot X_{3j}$

Part of the computer output includes:

i	bi	S_{bi}
0	8	1.6
1	2.2	.24
2	-.72	.32
3	0.005	0.002

a. Calculation of confidence interval for b_2 consists of _____± (a student's t value) (_____)
b. The confidence level for this interval is reflected in the value used for _____.
c. The degrees of freedom available for estimating the variance are directly concerned with the value used for _____

29. An investigator has used a multiple regression program on 20 data points to obtain a regression equation with 3 variables. Part of the computer output is:

Variable	**Coefficient**	**Standard Error of bi*bi***
1	0.45	0.21
2	0.80	0.10
3	3.10	0.86

a. 0.80 is an estimate of _____.
b. 0.10 is an estimate of _____.
c. Assuming the responses satisfy the normality assumption, we can be 95% confident that the value of β_2 is in the interval,_____ ± [$t_{.025}$ · _____], where $t_{.025}$ is the critical value of the student's *t*-distribution with ____ degrees of freedom.