CHAPTER 13

CORRELATION AND REGRESSION

OBJECTIVES

After completing this chapter, you should

- understand the meaning of correlation and be able to compute the Pearson r and Spearman r_S correlation coefficients.
- understand the linear regression equation and be able to compute it and use it to predict a value of the Y variable when you know a value on the X variable.

CHAPTER REVIEW

Correlation is defined as the degree of relationship between two or more variables. Although there are many kinds of correlation, the chapter focuses on *linear* correlation, or the degree to which a straight line best describes the relationship between two variables.

The degree of linear relationship between two variables may assume an infinite range of values, but it is customary to speak of three different classes of correlation. *Zero* correlation is defined as no relationship between variables. *Positive* correlation means there is a direct relationship between the variables, such that as one variable increases, so does the other. An inverse relationship in which low values of one variable are associated with high values of the other is called *negative* correlation.

A *scatterplot* is often used to show the relationship between two variables. Scatterplots are graphs in which pairs of scores are plotted, with the scores on one variable plotted on the X axis and scores on the other variable plotted on the Y axis. On the scatterplot, a pattern of points describing a line sloping upward to the right indicates positive correlation, and points indicating a line sloping downward to the right reveal negative correlation. Zero correlation is shown by a random pattern of points on the scatterplot. High correlation between two variables doesn't necessarily mean that one variable *caused* the other.

When the data are at least interval scale, the *Pearson product-moment correlation coefficient,* or *Pearson r,* is used to compute the degree of relationship between two variables. The Pearson r may be defined as the mean of the z-score products for X and Y pairs, where X stands for one variable and Y stands for the other. One approach to understanding the Pearson correlation is based on a close relative of variance, the *covariance,* which is the extent to which two variables vary together. Covariance can be used to derive a simple formula for the Pearson correlation, and we can think of the Pearson r as a standardized covariance between X and Y.

The range of r is from –1 to +1. Restricting the range of either the X or the Y variable lowers the value of r. The *coefficient of determination,* r^2, tells the amount of variability in one variable explained by variability in the other variable.

After computing the Pearson r, we can test it for significance. First, we assume that our sample was taken from a population in which there is no relationship between the two variables; this is just another version of the null hypothesis. Then, we consult Table E, which contains values of r for different degrees of freedom ($N - 2$) with probabilities of either .05 or .01. If our computed coefficient, in absolute value, is equal to or greater than the critical value at the 5% level, we reject the null hypothesis and conclude that our sample probably came from a population in which there is a relationship between the variables.

From the definition of correlation as the degree of linear relationship between two variables, we can use the correlation coefficient to compute the equations for the straight lines best describing the relationship between the variables. The equations (one to predict X and one to predict Y) are called *regression equations,* and we can use them to predict a score on one variable if we know a score on the other. The general form of the equation is $Y = bX + a$, where b is the slope of the line and a is where the line intercepts the Y axis. The regression line is also called the *least squares line.*

The *Spearman rank order correlation coefficient,* r_S, is a computationally simple alternative to r that is useful when the measurement level of one or both variables is ordinal scale. Like the Pearson r, the Spearman coefficient can be tested for significance. To test r_S for significance, we compare its value with critical values in Table F for the appropriate sample size; if our computed value is larger in absolute value than the table value at the 5% level, we reject the null hypothesis and conclude that the two variables are related.

Other correlation coefficients briefly considered in the chapter are the point biserial correlation (r_{pbis}) and the phi coefficient (ϕ). The former is useful when one variable is dichotomous (has only two values) and the other variable is continuous or interval level, whereas the latter is used when both variables are dichotomous. All of the inferential statistical methods covered in the text through this chapter can be tied together under the general linear model, which is a general, relationship-oriented multiple predictor approach to inference.

SYMBOLS

Symbol	*Stands For*
r	Pearson r, Pearson product-moment correlation coefficient
z_X, z_Y	z scores for the X and Y variables, respectively
cov_{XY}	covariance of X and Y
ρ	population correlation coefficient, read “rho”
r_{comp}, r_{crit}	computed value of r and the critical value of r from Table E, respectively
$\hat{Y}$	Y-caret, predicted values for Y based on the regression equation
b	regression coefficient, slope of the regression line
a	Y intercept, value of Y where the regression line crosses the Y axis
s_Y, s_X	standard deviation of the Y variable and the X variable, respectively
r^2	coefficient of determination
r_S	Spearman rank order correlation coefficient
d	difference between the ranks

FORMULAS

Formula 13-2. *Computational formula for the Pearson r*

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

The values needed to compute the equation are: ΣX, ΣY, ΣX^2, ΣY^2, ΣXY, and N. ΣXY is found by multiplying each X by each Y and summing the result.

Formula 13-3. *Regression equation for predicting Y from X*

$$\hat{Y} = \left(\frac{rs_Y}{s_X}\right)X + \left[\overline{Y} - \left(\frac{rs_Y}{s_X}\right)\overline{X}\right]$$

Formula 13-4. *Equation for determining the proportion of variability in data explained by correlation*

$$\text{coefficient of determination} = \frac{\text{explained variation}}{\text{total variation}} = r^2$$

Formula 13-5. *Equation for the Spearman rank order correlation coefficient*

$$r_S = 1 - \frac{6\Sigma d^2}{N\left(N^2 - 1\right)}$$

d is the difference between the *ranks* of individuals on the two variables, and N is the number of pairs of observations.

TERMS TO DEFINE AND/OR IDENTIFY

correlation

linear relationship

positive correlation

scatterplot

negative correlation

zero correlation

Pearson product-moment correlation coefficient

Pearson *r*

covariance

regression equation

least squares line

regression coefficient

multiple regression

coefficient of determination

Spearman rank order correlation coefficient

point biserial correlation coefficient

phi coefficient

general linear model

FILL-IN-THE-BLANK ITEMS

Linear Correlation

The degree of relationship between two or more variables is called (1) _________________. If the relationship is best described by means of a straight line, we call this (2) _________________ _________________.

Classes of correlation

A direct relationship between two variables, in which a high score is associated with a (3) _________________ score and a low score with a (4) _________________ score, is called (5) _________________ correlation. One way to study the relationship between the variables is with a (6) _________________ or graph on which scores for one variable are plotted on the *X* axis and scores for the other variable are plotted on the *Y* axis. An inverse relationship between the variables is called (7) _________________ correlation and is shown by a line sloping (8) _________________ to the right on a scatterplot. If the relationship between the variables is very small or nonexistent, the "class" of correlation is called (9) _________________ correlation. The strength of a relationship between two variables is given by the (10) _________________ _________________ of the correlation coefficient.

Correlation and causation

A high correlation between two variables doesn't automatically mean that one variable (11) _________________ the other. Correlation is necessary but not (12) _________________ to determine causality.

The Pearson Product-Moment Correlation Coefficient

The Pearson *r* is defined as the (13) _________________ of the *z*-score products for *X*-*Y* pairs of scores. The range of *r* is from (14) ______________ to _____________. A (15) ____________ value of *r* indicates a direct relationship between the variables, and a negative value indicates an (16) _________________ relationship. Values of *r* close to (17) _________________ indicate little or no relationship between the variables.

Correlation, variance, and covariance

We can define the (18) _________________ as the extent to which two variables vary together. The variance, then, is a special case of the (19) _________________ of *X* and *X*—of a variable with itself. Standardizing the covariance gives us a simple formula for the (20) _________________ _________________.

The effect of range on correlation

Restricting the range of either the *X* or the *Y* variable (21) _________________ the correlation.

Testing r for significance

To test *r* for significance, we first assume there is (22) _________________ _________________ in the population between the variables; that is, we assume that the underlying population correlation coefficient, (23) _________________, is (24) _________________. Then we look in Table (25) _________________ for values of *r* known to occur 5% or 1% of the time in samples of a given size, converted to (26) _________________, from a population with a (27) _________________ coefficient. If the absolute value of our sample coefficient exceeds the critical table value, then we (28) _________________ the null hypothesis, indicating that there is a significant (29) _________________ between the variables in the population sampled.

The linear regression equation

Correlation is defined as the degree of (30) _________________ relationship between the variables. Based on this definition, we can use correlation for prediction by first computing the equation for the (31) _________________ line that best describes the relationship between the variables. The general equation for the regression equation is (32) _________________, where *b* is the (33) _________________ of the line and *a* is where the line intercepts the (34) _________________ _________________. The regression line is the line that makes the squared (35) _________________ around it as small as possible. Unless *r* is (36) _________________, we must compute separate equations to predict *Y* given *X* and *X* given *Y.* The regression formula can be extended to include more than one predictor; this extension is called (37) _________________ _________________.

The coefficient of determination

The (38) _________________ _________________ _____________________, symbolized by (39) _________________, tells the amount of variability in one variable explained by variability in the other variable. This gives us a method to assess how (40) _________________ the relationship is between *X* and *Y* and is more important than the

(41) _________________ level.

The Spearman Rank Order Correlation Coefficient

The Spearman coefficient is useful as an alternative to *r* because it is easier to (42) _________________. Also, we can use it when the level of measurement on one or both of our variables is

(43) _________________ scale rather than interval scale as required by the Pearson *r*. With

(44) _________________ scale data, the exact length of the intervals between scores cannot be specified.

To compute the Spearman r_S, we first (45) _________________ the scores on each of the variables from highest to lowest and then find the difference between the (46) _________________. If two or more subjects are tied for a particular rank, each subject is given the (47) _________________ of the tied ranks.

Other correlation coefficients

The (48) _________________ _________________ correlation is used when one variable is dichotomous—has only (49) _________________ values—and the other variable is continuous or interval level measurement. When both variables are dichotomous, the (50) _________________ _________________ is used.

A Broader View of Inferential Techniques—The *General Linear Model*

The (51) ______________ ______________technique is the most general of all the techniques we've studied. As such, it is called the (52) _________________ ___________. Basically, what we are saying is that the most general way of looking at data has to do with (53) _________________ between measures. Thus, regression and correlation give us direct information about the statistical significance of a relationship and also about the (54) _________________ of the relationship. Tests such as the *t* test and ANOVA investigate (55) _________________ differences, which is the *other* way to study relationships.

Troubleshooting Your Computations

Any r or r_S computed must fall within the range of values from (56) _________________ to _________________. A common error in computing r_S is forgetting to (57) _________________ the scores on the two variables. Remember that the fractional part of the r_S formula is subtracted from (58) _________________. In computing the regression equation, be particularly careful in handling the last two terms in the equation, (59) _________________. The two numbers are added (60) ________________.

PROBLEMS

1. On the basis of your experience, decide whether the following pairs of variables are positively, negatively, or not correlated.

a. amount of alcohol consumed and speed of reaction to a suddenly appearing stimulus.

b. ratings of physical attractiveness of husband–wife pairs

c. IQ scores and the number of trials to learn a list of nonsense syllables for students in a general psychology class

d. sales of McDonald's hamburgers per day in a city and the number of people committed per day to a state mental institution

e. amount of time spent in practice and the average golf score

f. number of siblings and the likelihood of developing lung cancer

g. length of depression and the probability of suicide

2. The scores of 10 people on standardized scales of introversion and shyness are shown here (high scores on each scale indicate high introversion and shyness).

Person	Introversion	Shyness
1	17	22
2	6	4
3	12	10
4	13	8
5	19	11
6	20	18
7	9	10
8	4	3
9	8	10
10	21	16

Make a scatterplot of the data. Which class of correlation is revealed in the graph?

Compute *r,* and test it for significance. How much of the variability is accounted for by *r*?

3. Using the information in Problem 2, compute the regression equation for $\hat{Y}$.

Use the equation to predict the shyness score of a person with an introversion score of 15.

4. In a physiological psychology class, the first and last exam scores were as follows:

Student	First Score	Last Score
A	60	75
B	88	84
C	99	98
D	62	73
E	86	91
F	92	91
G	99	97
H	78	90
I	92	93
J	62	78
K	61	64
L	75	82
M	92	92
N	86	76
O	58	79
P	32	46
Q	54	67

Compute *r*, and test it for significance.

Find the regression equation for $\hat{Y}$. Use the equation to predict a last exam score for a student who made a 95 on the first test. Do the same for a student who made a 55 on the first test.

5. Ten female monkeys with male offspring are assigned ratings on a dominance test. The ratings of their male offspring are also determined. Assume the ratings are ordinal level measurement at best. Is there a relationship between the ratings? Test your correlation coefficient for significance.

Female Number	Dominance Ratings	Ratings of Offspring
1	10	8
2	9	10
3	9	8
4	7	6
5	6	6
6	5	6
7	5	4
8	5	5
9	3	4
10	2	1

6. A group of students was asked to estimate the amount of time each spends per day reading the newspaper. Then each student was given a 20-item recognition test of current events. The paired scores are:

Student	Time in Minutes	Score
A	25	12
B	40	13
C	55	18
D	10	8
E	5	5
F	5	3
G	30	10
H	45	15

What is the degree of relationship between the variables? Is it significant? How much of the variability in the data is accounted for by *r*?

7. Using figures from *Consumer Reports,* a consumer wants to see whether there is any relationship between the weight of a car and the gas mileage it gets in city driving. The following pairs of scores are taken from the annual car buying guide:

Car	Weight in Thousands of Pounds	mpg in Town
1	2.6	17.4
2	2.1	20.8
3	2.2	19.2
4	2.0	19.8
5	3.2	13.9
6	2.7	13.4
7	3.4	11.8
8	3.8	10.3
9	3.9	9.5

Compute the correlation coefficient, and test it for significance.

8. Using the data from Problem 7, find the regression equation for $\hat{Y}$. What gas mileage can the consumer expect from a car weighing 4,300 pounds?

9. In a study of the parents of schizophrenic children, letters have been independently rated by two psychiatrists for the presence of contradictory ideas and feelings. The rating scale assigns numbers from 0 to 7, with higher numbers indicating more contradictions. Assume the ratings are ordinal scale measurement at best. Compute a correlation coefficient, and test it for significance.

Letter	Rater A	Rater B
A	3	7
B	2	4
C	5	3
D	7	6
E	0	5
F	1	4
G	2	4
H	4	2

10. Two experimenters have independently rated the handling characteristics of 12 rats. Compare the correlation between the ratings of Experimenters A and B, assuming that the ratings are ordinal scale measurement at best. Test your coefficient for significance.

Rat	Experimenter A	Experimenter B
1	15	13
2	13	14
3	10	11
4	5	5
5	8	10
6	6	8
7	3	4
8	7	6
9	7	5
10	2	1
11	2	2
12	2	3

11. The following pairs of scores are heart rate values measured by a physiograph for subjects looking at different stimuli. Compute the most appropriate correlation coefficient, and test it for significance.

Subject	Stimulus A	Stimulus B
1	65.3	71.8
2	75.7	73.5
3	85.6	99.3
4	73.7	81.7
5	69.5	75.7
6	68.2	73.5
7	70.1	79.8
8	72.5	70.3
9	71.0	85.3
10	83.5	107.1

USING SPSS—EXAMPLES AND EXERCISES

SPSS provides procedures for both the Pearson and the Spearman correlation coefficients, as well as procedures for simple and multiple regression. We will also show you how to obtain a scatterplot with regression line to assist in visualizing the relationship between two measures and to check the assumption of a *linear* correlation/regression relationship.

Example—Pearson and Spearman Correlation Coefficients: We will use SPSS to work Problem 5. First, let's assume that the dominance ratings are interval level measurement and compute the Pearson correlation. After we complete that process, we will assume that the ratings are pure ranks (ordinal level measurement) and calculate the Spearman correlation. Finally, we will obtain a scatterplot of the data. The steps are as follows:

1. Start SPSS and enter the data into variable columns named **mother** and **son.**
2. Select *Analyze>Correlate>Bivariate*.
3. Highlight and move both variables into the Variables box and select both the Pearson and Spearman boxes under Correlation Coefficients.
4. Click *Options>Means and Standard deviations>Continue>OK*, and the results should appear in the output Viewer window.
5. To obtain a scatterplot, select *Graph>Scatter>Simple>Define*.
6. Highlight and move **mother** to the *X*-axis box and **son** to the *Y*-axis box; click *OK,* and the plot should appear in the output Viewer window.

Notes on Reading the Output

1. The Correlations box gives the results for the Pearson correlation as a matrix. The first number in the box is the correlation coefficient, $r = .971$; the next value is the significance of the correlation, $p = .000$ ($p < .001$); and the last value is the sample size, $N = 10$. Thus, the correlation is statistically significant. The Correlations box under the section labeled Nonparametric Correlations gives the Spearman correlation ($r_S = .925$) using the same arrangement.
2. In examining the scatterplot, remember that we never connect the points but look at the plot to confirm the sign and strength of the correlation and to check for a nonlinear pattern, which would be a possible violation of our assumption of a linear relationship. You might want to put an oval around the points to assist you in visualizing the relationship.

```
CORRELATIONS
  /VARIABLES=mother son
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE .
```

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
MOTHER	6.1000	2.6437	10
SON	5.8000	2.5298	10

Correlations

		MOTHER	SON
MOTHER	Pearson Correlation	1.000	.917**
	Sig. (2-tailed)	.	.000
	N	10	10
SON	Pearson Correlation	.917**	1.000
	Sig. (2-tailed)	.000	.
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

```
NONPAR CORR
  /VARIABLES=mother son
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Nonparametric Correlations

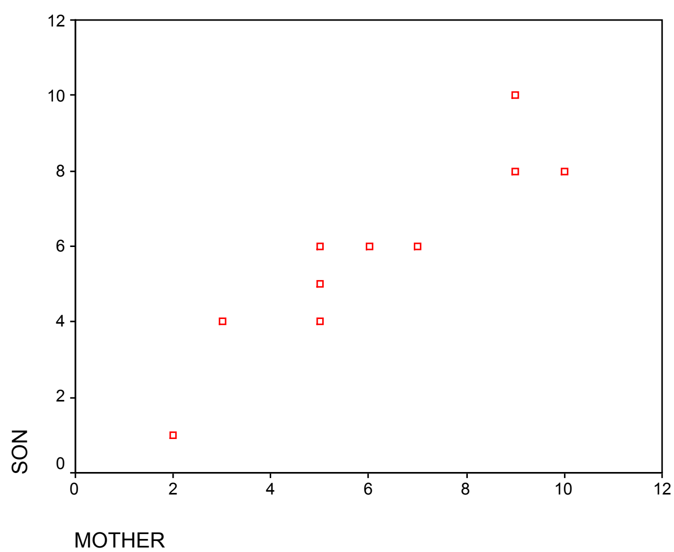
Correlations

			MOTHER	SON
Spearman's rho	MOTHER	Correlation Coefficient	1.000	.925**
		Sig. (2-tailed)	.	.000
		N	10	10
	SON	Correlation Coefficient	.925**	1.000
		Sig. (2-tailed)	.000	.
		N	10	10

**. Correlation is significant at the .01 level (2-tailed).

```
GRAPH
  /SCATTERPLOT(BIVAR)=mother WITH son
  /MISSING=LISTWISE .
```

Graph



Example—Regression: Let's use SPSS to find the regression equation for Problem 7. Also, we will produce a scatterplot with a regression line. The procedure is as follows:

1. Start SPSS and enter the data under the variable names **weight** and **mpg**.
2. Select *Analyze*>*Regression*>*Linear*.
3. Move **mpg** into the Dependent box and **weight** into the Independent(s) box; click *Statistics*>*Descriptives* (*Estimates* and *Model* fit should already be checked by default)>*Continue*>*OK*. The results should appear in the output Viewer window.
4. To get the scatterplot, switch to the Data Editor window and follow the instructions in the previous example on producing it. Note that in a regression problem, we want the dependent variable—the measure we want to predict—to be plotted on the vertical (Y) axis, so **mpg** should appear on the Y axis.
5. Once the scatterplot has appeared in the output Viewer window, double-click on the chart and maximize the chart window. In the Menu Bar, click *Chart*>*Options*. In the Scatterplot Options box, select *Total*>*OK*, and the regression line should appear on the graph. Because we are finished editing the chart, select *File*>*Close* to close the chart window. If necessary, switch to the output Viewer window, in which you should find the scatterplot with regression line.

Notes on Reading the Output

1. As with the previous examples, you should find the results relatively easy to identify and interpret in this output. The Coefficients box gives the intercept and slope terms in the column labeled B under the section labeled Unstandardized Coefficients. The term in the row labeled (Constant) is the intercept term, $a = 31.426$, and the term in the row labeled WEIGHT is the slope, $b = -5.678$. Thus, the regression equation could be written as follows:

$$\hat{Y} = 31.426 - 5.678X \text{ or}$$
$$\text{MPG} = 31.426 - 5.678(\text{WEIGHT}).$$

2. The scatterplot with regression line shows a close fit with no indication of nonlinearity.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mpg
  /METHOD=ENTER weight  .
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
MPG	15.1222	4.2763	9
WEIGHT	2.8778	.7259	9

Correlations

		MPG	WEIGHT
Pearson Correlation	MPG	1.000	-.964
	WEIGHT	-.964	1.000
Sig. (1-tailed)	MPG	.	.000
	WEIGHT	.000	.
N	MPG	9	9
	WEIGHT	9	9

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	WEIGHT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: MPG

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.964[a]	.929	.919	1.2184

a. Predictors: (Constant), WEIGHT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	135.904	1	135.904	91.548	.000[a]
	Residual	10.392	7	1.485		
	Total	146.296	8			

a. Predictors: (Constant), WEIGHT

b. Dependent Variable: MPG

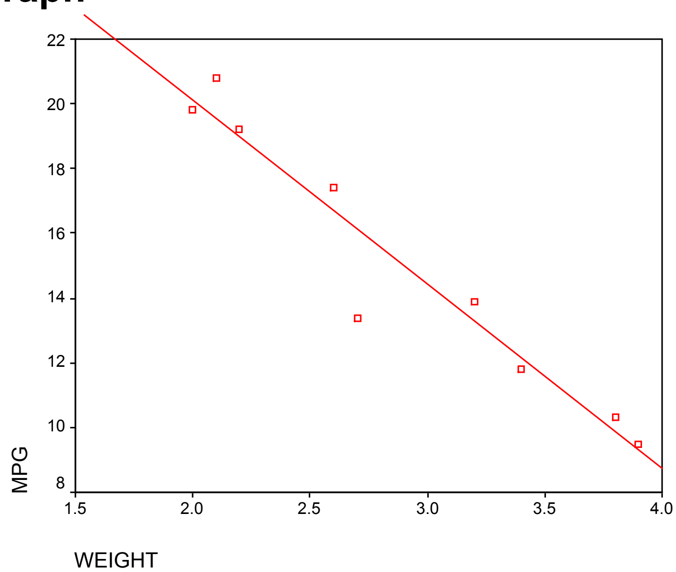
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	31.462	1.755		17.923	.000
	WEIGHT	-5.678	.593	-.964	-9.568	.000

a. Dependent Variable: MPG

```
GRAPH
  /SCATTERPLOT(BIVAR)=weight WITH mpg
  /MISSING=LISTWISE .
```

Graph



Exercises Using SPSS

1. Use SPSS to work Problem 6.
2. Use SPSS and the data from Problem 6 to compute a regression equation predicting the current events score from time spent reading the newspaper. Give the regression equation and obtain a scatterplot with regression line.
3. Use SPSS to work Problem 9, assuming the data are pure ranks. Obtain a scatterplot.

CHECKING YOUR PROGRESS: A SELF-TEST

1. Match the following:

_______ positive correlation	**a.** a straight line describes the relationship between two variables
_______ negative correlation	**b.** coefficient of determination
_______ zero correlation	**c.** *Y* intercept of the regression line
_______ ρ	**d.** no relationship between the variables
_______ scatterplot	**e.** direct relationship between the variables
_______ linear correlation	**f.** inverse relationship between the variables
_______ regression equation	**g.** population correlation coefficient
_______ r^2	**h.** used for prediction
_______ *b*	**i.** graph used to show the relationship between two variables
_______ *a*	**j.** slope of the regression line

2. The ACT math and science scores for eight students are shown here. Compute *r,* and test it for significance.

Student	Math ACT	Science ACT
A	26	24
B	22	24
C	13	10
D	30	31
E	12	17
F	15	15
G	19	21
H	20	16

3. Use the data from Problem 2 to compute a regression equation, and use the equation to predict a science ACT score for a student scoring 33 on the math ACT.

4. Without knowing who is married to whom, an observer has rated the attractiveness of 10 couples on a 10-point scale. Compute the appropriate correlation coefficient, and test it for significance. Assume that the ratings are ordinal scale measurement at best.

Couple	Wife's Rating	Husband's Rating
A	7	6
B	6	8
C	5	4
D	8	9
E	3	5
F	1	2
G	5	2
H	9	9
I	10	7
J	7	5