(a) $r > 0.3$

(b) $r > 0.7$

(c) $|r| > 0.7$

14.15 Complete this statement: H_0: ρ = ____________

14.16 Fill in the blanks: The least squares regression line is given by $\hat{y}$ = ____________ + ____________ x.

14.17 In the above statement, what does $\hat{y}$ represent?

14.18 What symbol is used to denote the intercept of a regression line?

14.19 What represents the predicted change in Y unit X?

14.20 What symbol represents the slope estimate?

14.21 What symbol represents the slope parameter?

14.22 What is wrong with this statement? "A 95% confidence interval for b is -0.91 to -0.42."

14.23 What is the distance of a data point from the regression line called?

14.24 Besides linearity, what conditions are needed to infer population slope β?

14.25 Besides linearity, what conditions are needed to infer population correlation coefficient ρ?

14.26 Complete this statement: (residual i) = (observed Y value of i) − ____________.

14.27 In symbols, residual$_i$ = y_i − ____________.

14.28 What is wrong with this statement? "H_0: $r = 0$."

14.29 Fix this statement: "H_a: $\beta = 0$". (Four fixes are possible.)

Exercises

chapter 14

14.3 ***Bicycle helmet use, n = 12.*** Exercise 14.1 introduced data for a cross-sectional survey of bicycle helmet use in Northern California. Table 14.5 lists the data. Exercise 14.1(a) revealed that observation 13 (Los Arboles) was an outlier.

(a) After eliminating the outlier from the data set (n is now equal to 12), calculate least squares regression coefficients a and b. Interpret these statistics.

(b) Calculate the 95% confidence interval for slope parameter β.

(c) Use the 95% confidence interval to predict whether the slope is significant at $\alpha = 0.05$.

(d) *Optional:* Determine the residuals for each of the 12 data points in the analysis. Plot these residuals as a stemplot. Check for major departures from Normality.

14.4 ***Mental health care.*** Exercise 14.2 introduced historical data in which the explanatory variable was the reciprocal of the distance to the nearest health facility (miles^{-1}, variable name REC_DIST) and the response variable was the percent of patients cared for at home (variable name PHOME2). Table 14.6 lists the data. Eliminate observation 13 (Nantucket) and then determine the least squares regression coefficients for the data.

14.5 ***Anscombe's quartet.*** "Graphs are essential to good statistical analysis," so starts a 1973 article by Anscombe.[9] This article demonstrates why it is important to *look* at the data before analyzing it numerically. Table 14.10 contains four different data sets. Each of the data sets produces these identical numerical results:

$$n = 11 \quad \bar{x} = 9.0 \quad \bar{y} = 7.5 \quad r = 0.82 \quad \hat{y} = 3 + 0.5X \quad P = 0.0022$$

Figure 14.17 shows scatterplots for each of the data sets. Consider the relevance of the numerical statistics in light of the scatterplots. Would you use correlation or regression to analyze any of these data sets? Explain your reasoning in each instance.

14.6 ***Domestic water and dental cavities.*** Table 14.11 contains data from a historically important study of water fluoridation and dental caries in 21 North American cities.

(a) Construct a scatterplot of FLUORIDE and CARIES. Discuss the plot. Are there any outliers? Is the relationship linear? If the relationship is not linear, what type of relation *is* evident? Would linear regression be warranted under these circumstances?

(b) Although a single straight line does not fit these data, we may build a valid linear model by reexpressing the data through a mathematical transformation. Apply logarithmic transforms (base *e*) to both FLUORIDE and CARIES. Create a new plot with the transformed data. Discuss the results.

(c) Calculate the coefficients for a least square regression line for the ln–ln transformed data. Interpret the slope estimate.

(d) Calculate r^2 for ln–ln transformed data. Interpret the result.

[9] Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27,* 17–2... Data are stored in ANSCOMB.*. Anscombe (1973) states "Most textbooks on statistical methods pay little attention to graphs. Few of us escape being indoctrinated with these *notions:* (1) Numerical calculations are exact, but graphs are rough; (2) For any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis; (3) Performing intricate calculations is virtuous, whereas actually looking at data is cheating." This text attempts to avoid these false notions.

TABLE 14.10 Anscombe's quartet, Exercise 14.5.

Data set I		Data set II		Data set III		Data set IV	
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
[illegible]	8.04	10.0	9.14	10.0	7.46	8.0	6.58
[illegible]	6.95	8.0	8.14	8.0	6.77	8.0	5.76
[illegible]	7.58	13.0	8.74	13.0	12.74	8.0	7.71
[illegible]	8.81	9.0	8.77	9.0	7.11	8.0	8.84
[illegible]	8.33	11.0	9.26	11.0	7.81	8.0	8.47
[illegible]	9.96	14.0	8.10	14.0	8.84	8.0	7.04
[illegible]	7.24	6.0	6.13	6.0	6.08	8.0	5.25
[illegible]	4.26	4.0	3.10	4.0	5.39	19.0	12.50
[illegible]	10.84	12.0	9.13	12.0	8.15	8.0	5.56
[illegible]	4.82	7.0	7.26	7.0	6.42	8.0	7.91
[illegible]	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Data from Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, 17–21.

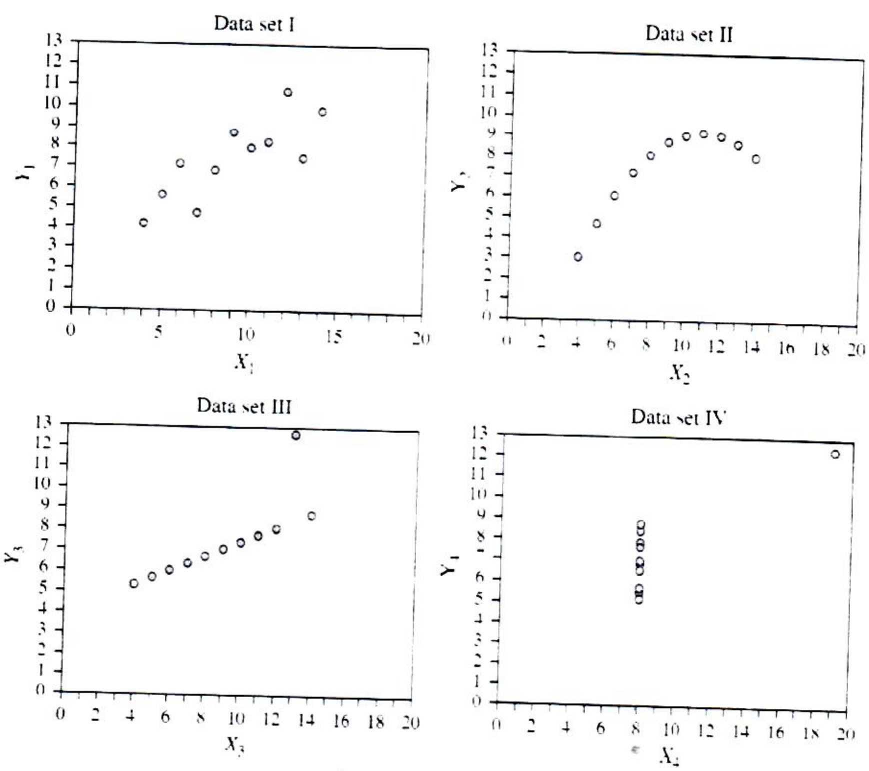

FIGURE 14.17 Anscombe's quartet, Exercise 14.5.

14.7 ***Domestic water and dental cavities, range restriction.*** Another [illegible] the data presented in Exercise 14.6 is to restrict the analysis to a [illegible] be described with a straight line. This is called a range restriction

(a) Use the scatterplot you drew in Exercise 14.6(a) to determine [illegible] is a range of FLUORIDE in which the relationship between [illegible] CARIES is fairly straight. Restrict the data to this range. Then [illegible] least squares line for this segment of the data. Interpret b.

TABLE 14.11 Fluoride in public water (parts per million, variable name FLUORIDE) and dental caries experience in permanent teeth per 100 children (variable CARIES).

CITYID	FLUORIDE	CARIES
1	1.9	236
2	2.6	246
3	1.8	252
4	1.2	258
5	1.2	281
6	1.2	303
7	1.3	323
8	0.9	343
9	0.6	412
10	0.5	444
11	0.4	556
12	0.3	652
13	0.0	673
14	0.2	703
15	0.1	706
16	0.0	722
17	0.2	733
18	0.1	772
19	0.0	810
20	0.1	823
21	0.1	1037

Data from Dean, H. T., Arnold, F. A., Jr., & Elvove, E. (1942). Domestic water and dental caries. *Public Health Reports, 57*, 1155–117[illegible] Data are stored online in the file DEAN1942[illegible]

(b) Calculate r^2 for this range-restricted model. Which [illegible] this model or the ln–ln transformed model from the [illegible]

(c) Which model do you prefer, this model or the one [illegible] Explain your reasoning.

14-8

14.8 ***Correlation matrix.*** Statistical packages can easily calculate correlation coefficients for multiple pairings of variables. Results are often reported in the form of a **correlation matrix**. Figure 14.18 displays the correlation matrix for data from a study of geographic variation in cancer rates.[r] The variables are:

CIG	cigarettes sold per capita
BLAD	bladder cancer deaths per 100,000
LUNG	lung cancer deaths per 100,000
KID	kidney cancer deaths per 100,000
LEUK	leukemia cancer deaths per 100,000

Notice that the values for each correlation coefficient appear twice in the matrix, each time the variables intersect in either row or column order. For example, the value $r = 0.704$ occurs for CIG and BLAD and for BLAD and CIG. The correlation of 1 across the diagonal reflects the trivial fact that each variable is perfectly correlated with itself. Review this correlation matrix and interpret its results.

14-9

14.9 ***True or false?*** Identify which of these statements are true and which are false.

(a) Correlation coefficient r quantifies the relationship between quantitative variables X and Y.

(b) Correlation coefficient r quantifies the linear relation between quantitative variables X and Y.

(c) The closer r is to 1, the stronger is the linear relation between X and Y.

(d) The closer r is to -1 or 1, the stronger is the linear relation between X and Y.

(e) If r is close to zero, X and Y are unrelated.

(f) If r is close to zero, X and Y are not related in a linear way.

(g) The value of r changes when the units of measure are changed.

(h) The value of b changes when the units of measure are changed.

14.10 ***Memory of food intake.*** Retrospective studies of diet and health often rely on recall of distant dietary histories. The validity and reliability of such information is often suspected. An epidemiologic study asked middle-aged adults (median age 50) to recall food intake at ages 6, 18, and 30 years. Recall was validated by comparing recalled results to historical information collected during earlier time periods. Correlations rarely exceeded $r = 0.3$.[s] What do you conclude from this result?

---

[r] Fraumeni, J. F., Jr. (1968). Cigarette smoking and cancers of the urinary tract: Geographic variation in the United States. *Journal of the National Cancer Institute, 41*(5), 1205–1211.

[s] Dwyer, J. T., Gardner, J., Halvorsen, K., Krall, E. A., Cohen, A., & Valadian, I. (1989). Memory of food intake in the distant past. *American Journal of Epidemiology, 130*(5), 1033–1046.

Correlations

		CIG	BLAD	LUNG	KID
CIG	Pearson correlation	1	0.704**	0.697**	[illegible]
	Sig. (two-tailed)	.	0.000	0.000	[illegible]
	N	44	44	44	[illegible]
BLAD	Pearson correlation	0.704**	1	0.659**	[illegible]
	Sig. (two-tailed)	0.000	.	0.000	[illegible]
	N	44	44	44	[illegible]
LUNG	Pearson correlation	0.697**	0.659**	1	[illegible]
	Sig. (two-tailed)	0.000	0.000	.	[illegible]
	N	44	44	44	[illegible]
KID	Pearson correlation	0.487**	0.359*	0.283	[illegible]
	Sig. (two-tailed)	0.001	0.017	0.063	[illegible]
	N	44	44	44	[illegible]
LEUK	Pearson correlation	−0.068	0.162	−0.152	[illegible]
	Sig. (two-tailed)	0.659	0.293	0.326	[illegible]
	N	44	44	44	[illegible]

** Correlation is significant at the 0.01 level (two-tailed).
* Correlation is significant at the 0.05 level (two-tailed).

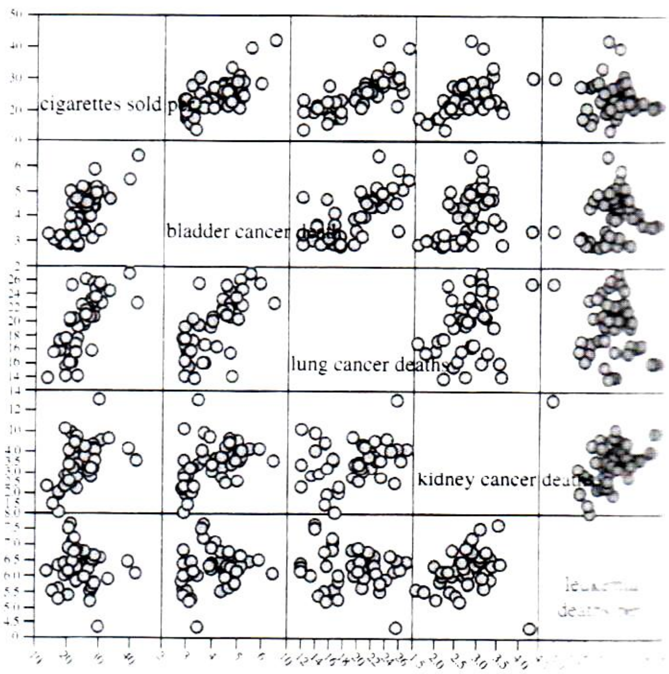

FIGURE 14.18 Correlation matrix and scatterplot matrix for [illegible]