which is the *sample variance* of Y. The standard error of $\hat{\beta}_0$ is then $\hat{\sigma}/\sqrt{n} = s_y/\sqrt{n}$, which is the familiar standard error of the sample mean $\bar{y}$. The t-Test for testing Model 1 against Model 2 is

$$t_1 = \frac{\hat{\beta}_0 - 0}{\text{s.e.}(\hat{\beta}_0)} = \frac{\bar{y}}{s_y/\sqrt{n}}\,, \tag{2.60}$$

which is the same as the one-sample t-Test in (2.57).

The second example occurs in connection with the *paired two-sample t-Test*. For example, to test whether a given diet is effective in weight reduction, a random sample of n people is chosen and each person in the sample follows the diet for a specified period of time. Each person's weight is measured at the beginning of the diet and at the end of the period. Let Y_1 and Y_2 denote the weight at the beginning and at the end of diet period, respectively. Let $Y = Y_1 - Y_2$ be the difference between the two weights. Then Y is a random variable with mean μ and variance σ^2. Consequently, testing whether or not the diet is effective is the same as testing $\text{H}_0 : \mu = 0$ against $\text{H}_1 : \mu > 0$. With the definition of Y and assuming that Y is normally distributed, the well-known paired two-sample t-Test is the same as the test in (2.57). This situation can be modeled as in (2.58) and the test in (2.60) can be used to test whether the diet is effective in weight reduction.

The above two examples show that the one-sample and the paired two-sample tests can be obtained as special cases using regression analysis.

2.12 BIBLIOGRAPHIC NOTES

The standard theory of regression analysis is developed in a number of good text books, some of which have been written to serve specific disciplines. Each provides a complete treatment of the standard results. The books by Snedecor and Cochran (1980), Fox (1984), and Kmenta (1986) develop the results using simple algebra and summation notation. The development in Searle (1971), Rao (1973), Seber (1977), Myers (1990), Sen and Srivastava (1990), Green (1993), Graybill and Iyer (1994), and Draper and Smith (1998) lean more heavily on matrix algebra.

EXERCISES

2.1 Using the data in Table 2.6: (table on lecture slides)

(a) Compute Var(Y) and Var(X).

(b) Prove or verify that $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$.

(c) Prove or verify that any standardized variable has a mean of 0 and a standard deviation of 1.

(d) Prove or verify that the three formulas for Cor(Y, X) in (2.5), (2.6), and (2.7) are identical. (equations on lecture slides)

(e) Prove or verify that the three formulas for $\hat{\beta}_1$ in (2.14) and (2.20) are identical.

2.2 Explain why you would or wouldn't agree with each of the following statements:

(a) $\text{Cov}(Y, X)$ and $\text{Cor}(Y, X)$ can take values between $-\infty$ and $+\infty$.

(b) If $\text{Cov}(Y, X) = 0$ or $\text{Cor}(Y, X) = 0$, one can conclude that there is no relationship between Y and X.

(c) The least squares line fitted to the points in the scatter plot of Y versus $\hat{Y}$ has a zero intercept and a unit slope.

2.3 Using the regression output in Table 2.9, test the following hypotheses using $\alpha = 0.1$: (table on lecture slides)

(a) $H_0 : \beta_1 = 15$ versus $H_1 : \beta_1 \neq 15$

(b) $H_0 : \beta_1 = 15$ versus $H_1 : \beta_1 > 15$

(c) $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$

(d) $H_0 : \beta_0 = 5$ versus $H_1 : \beta_0 \neq 5$

2.4 Using the regression output in Table 2.9, construct the 99% confidence interval for β_0. (table on lecture slides)

2.5 When fitting the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ to a set of data using the least squares method, each of the following statements can be proven to be true. Prove each statement mathematically or demonstrate its correctness numerically (using the data in Table 2.5):

(a) The sum of the ordinary least squares residuals is zero.

(b) The two tests in (2.26) and (2.32) are equivalent. (equations on lecture slides)

(c) The scatter plot of Y versus X and the scatter plot of Y versus $\hat{Y}$ have identical patterns.

(d) The correlation coefficient between Y and $\hat{Y}$ must be nonnegative.

2.6 Using the data in Table 2.5, and the fitted values and the residuals in Table 2.7, verify that: (table on lecture slides)

(a) $\text{Cor}(Y, X) = \text{Cor}(Y, \hat{Y}) = 0.994$

(b) SST = 27768.348

(c) SSE = 348.848

2.7 Verify that the four data sets in Table 2.4 give identical results for the following quantities: (table on lecture slides)

(a) $\hat{\beta}_0$ and $\hat{\beta}_1$ (b) $\text{Cor}(Y, X)$

(c) R^2 (d) The t-Test

2.8 When fitting a simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ to a set of data using the least squares method, suppose that $H_0 : \beta_1 = 0$ was not

Table 2.10 Regression Output When Y is Regressed on X for Labor Force Participation Rate of Women

Variable	Coefficient	s.e.	t-Test	p-value
Constant	0.203311	0.0976	2.08	0.0526
X	0.656040	0.1961	3.35	< 0.0038
$n = 19$	$R^2 = 0.397$	$R_a^2 = 0.362$	$\hat{\sigma} = 0.0566$	df = 17

rejected. This implies that the model can be written simply as: $Y = \beta_0 + \varepsilon$. The least squares estimate of β_0 is $\hat{\beta}_0 = \bar{y}$. (Can you prove that?)

(a) What are the ordinary least squares residuals in this case?

(b) Show that the ordinary least squares residuals sum up to zero.

2.9 Let Y and X denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the United States. The regression output for this data set is shown in Table 2.10. It was also found that SSR = 0.0358 and SSE = 0.0544. Suppose that the model $Y = \beta_0 + \beta_1 X + \varepsilon$ satisfies the usual regression assumptions.

(a) Compute $\text{Var}(Y)$ and $\text{Cor}(Y, X)$.

(b) Suppose that the participation rate of women in 1968 in a given city is 45%. What is the estimated participation rate of women in 1972 for the same city?

(c) Suppose further that the mean and variance of the participation rate of women in 1968 are 0.5 and 0.005, respectively. Construct the 95% confidence interval for the estimate in (b).

(d) Construct the 95% confidence interval for the slope of the true regression line, β_1.

(e) Test the hypothesis: $\text{H}_0 : \beta_1 = 1$ versus $\text{H}_1 : \beta_1 > 1$ at the 5% significance level.

(f) If Y and X were reversed in the above regression, what would you expect R^2 to be?

2.10 One may wonder if people of similar heights tend to marry each other. For this purpose, a sample of newly married couples was selected. Let X be the height of the husband and Y be the height of the wife. The heights (in centimeters) of husbands and wives are found in Table 2.11. The data can also be found at the book's Website.

(a) Compute the covariance between the heights of the husbands and wives.

(b) What would the covariance be if heights were measured in inches rather than in centimeters?

Table 2.11 Heights of Husband (H) and Wife (W) in (Centimeters)

Row	H	W	Row	H	W	Row	H	W
1	186	175	33	180	166	65	181	175
2	180	168	34	188	181	66	170	169
3	160	154	35	153	148	67	161	149
4	186	166	36	179	169	68	188	176
5	163	162	37	175	170	69	181	165
6	172	152	38	165	157	70	156	143
7	192	179	39	156	162	71	161	158
8	170	163	40	185	174	72	152	141
9	174	172	41	172	168	73	179	160
10	191	170	42	166	162	74	170	149
11	182	170	43	179	159	75	170	160
12	178	147	44	181	155	76	165	148
13	181	165	45	176	171	77	165	154
14	168	162	46	170	159	78	169	171
15	162	154	47	165	164	79	171	165
16	188	166	48	183	175	80	192	175
17	168	167	49	162	156	81	176	161
18	183	174	50	192	180	82	168	162
19	188	173	51	185	167	83	169	162
20	166	164	52	163	157	84	184	176
21	180	163	53	185	167	85	171	160
22	176	163	54	170	157	86	161	158
23	185	171	55	176	168	87	185	175
24	169	161	56	176	167	88	184	174
25	182	167	57	160	145	89	179	168
26	162	160	58	167	156	90	184	177
27	169	165	59	157	153	91	175	158
28	176	167	60	180	162	92	173	161
29	180	175	61	172	156	93	164	146
30	157	157	62	184	174	94	181	168
31	170	172	63	185	160	95	187	178
32	186	181	64	165	152	96	181	170

(c) Compute the correlation coefficient between the heights of the husband and wife.

(d) What would the correlation be if heights were measured in inches rather than in centimeters?

(e) What would the correlation be if every man married a woman exactly 5 centimeters shorter than him?

(f) We wish to fit a regression model relating the heights of husbands and wives. Which one of the two variables would you choose as the response variable? Justify your answer.

(g) Using your choice of the response variable in Exercise 2.10(f), test the null hypothesis that the slope is zero.

(h) Using your choice of the response variable in 2.10(f), test the null hypothesis that the intercept is zero.

2.11 Consider fitting a simple linear regression model through the origin, $Y = \beta_1 X + \varepsilon$, to a set of data using the least squares method. (equations on lecture slides)

(a) Give an example of a situation where fitting the model (2.49) is justified by theoretical or other physical and material considerations.

(b) Show that least squares estimate of β_1 is as given in (2.50).

(c) Show that the residuals $e_1, e_2, \cdots, e_n$ will not necessarily add up to zero.

(d) Give an example of a data set Y and X in which R^2 in (2.46) but computed from fitting (2.49) to the data is negative.

(e) Which goodness of fit measures would you use to compare model (2.49) with model (2.48)?

2.12 In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands) (*Source*: *Gale Directory of Publications*, 1994). The data are given in Table 2.12 and can be found at the book's Website.

(a) Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggest a linear relationship between daily and Sunday circulation? Do you think this is a plausible relationship?

(b) Fit a regression line predicting Sunday circulation from daily circulation.

(c) Obtain the 95% confidence intervals for β_0 and β_1.

(d) Is there a significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test. Indicate what hypothesis you are testing and your conclusion.

(e) What proportion of the variability in Sunday circulation is accounted for by daily circulation?

(f) Provide an interval estimate (based on 95% level) for the average Sunday circulation of newspapers with daily circulation of 500,000.

Table 2.12 Newspapers Data: Daily and Sunday Circulations (in Thousands)

Newspaper	Daily	Sunday
Baltimore Sun	391.952	488.506
Boston Globe	516.981	798.298
Boston Herald	355.628	235.084
Charlotte Observer	238.555	299.451
Chicago Sun Times	537.780	559.093
Chicago Tribune	733.775	1133.249
Cincinnati Enquirer	198.832	348.744
Denver Post	252.624	417.779
Des Moines Register	206.204	344.522
Hartford Courant	231.177	323.084
Houston Chronicle	449.755	620.752
Kansas City Star	288.571	423.305
Los Angeles Daily News	185.736	202.614
Los Angeles Times	1164.388	1531.527
Miami Herald	444.581	553.479
Minneapolis Star Tribune	412.871	685.975
New Orleans Times-Picayune	272.280	324.241
New York Daily News	781.796	983.240
New York Times	1209.225	1762.015
Newsday	825.512	960.308
Omaha World Herald	223.748	284.611
Orange County Register	354.843	407.760
Philadelphia Inquirer	515.523	982.663
Pittsburgh Press	220.465	557.000
Portland Oregonian	337.672	440.923
Providence Journal-Bulletin	197.120	268.060
Rochester Democrat & Chronicle	133.239	262.048
Rocky Mountain News	374.009	432.502
Sacramento Bee	273.844	338.355
San Francisco Chronicle	570.364	704.322
St. Louis Post-Dispatch	391.286	585.681
St. Paul Pioneer Press	201.860	267.781
Tampa Tribune	321.626	408.343
Washington Post	838.902	1165.567

(g) The particular newspaper that is considering a Sunday edition has a daily circulation of 500,000. Provide an interval estimate (based on 95% level) for the predicted Sunday circulation of this paper. How does this interval differ from that given in (f)?

(h) Another newspaper being considered as a candidate for a Sunday edition has a daily circulation of 2,000,000. Provide an interval estimate for the predicted Sunday circulation for this paper? How does this interval compare with the one given in (g)? Do you think it is likely to be accurate?

2.13 Let $y_1, y_2, \cdots, y_n$ be a sample drawn from a normal population with unknown mean μ and unknown variance σ^2. One way to estimate μ is to fit the linear model

$$y_i = \mu + \varepsilon; \quad i = 1, 2, \cdots, n, \tag{2.61}$$

and use the least squares (LS), that is, to minimize the sum of squares, $\sum_{i=1}^{n}(y_i - \mu)^2$. Another way is to use the least absolute value (LAV), that is, to minimize the sum of absolute value of the vertical distances, $\sum_{i=1}^{n}|y_i - \mu|$.

(a) Show that the least squares estimate of μ is the sample mean $\bar{y}$.

(b) Show that the LAV estimate of μ is the sample median.

(c) State one advantage and one disadvantage of the sample mean.

(d) State one advantage and one disadvantage of the sample median.

(e) Which of the above two estimates of μ would you choose? Why?

2.14 An alternative to the least squares method is the orthogonal regression method. According to the orthogonal regression method, the estimated regression coefficients in the simple regression model are obtained by minimizing the sum of squares of the perpendicular distances from each point to the regression line. Show that the intercept and the slope of the line that minimizes the sum of the squared orthogonal distances are obtained by finding β_0 and β_1 that minimize the function

$$g(\beta_0, \beta_1) = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{1 + \beta_1^2}. \tag{2.62}$$

Unlike the least squares criterion, there is no closed-form solution to the minimization problem in (2.62). A solution, however, can be obtained using iterative methods. This is one reason for the popularity of the least squares method.