These examples demonstrate the need for care in drawing conclusions about causal relations from regression analysis. Regression analysis by itself provides no information about causal patterns and must be supplemented by additional analyses to obtain insights about causal relations.

Use of Computers

Because regression analysis often entails lengthy and tedious calculations, computers are usually utilized to perform the necessary calculations. Almost every statistics package for computers contains a regression component. While packages differ in many details, their basic regression output tends to be quite similar.

After an initial explanation of required regression calculations, we shall rely on computer calculations for all subsequent examples. We illustrate computer output by presenting output and graphics from BMDP (Ref. 1.1), MINITAB (Ref. 1.2), SAS (Ref. 1.3), SPSS (Ref. 1.4), SYSTAT (Ref. 1.5), JMP (Ref. 1.6), S-Plus (Ref. 1.7), and MATLAB (Ref. 1.8).

1.3 Simple Linear Regression Model with Distribution of Error Terms Unspecified

Formal Statement of Model

In Part I we consider a basic regression model where there is only one predictor variable and the regression function is linear. The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1.1}$$

where:

Y_i is the value of the response variable in the ith trial

β_0 and β_1 are parameters

X_i is a known constant, namely, the value of the predictor variable in the ith trial

ε_i is a random error term with mean $E\{\varepsilon_i\} = 0$ and variance $\sigma^2\{\varepsilon_i\} = \sigma^2$; ε_i and ε_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all $i, j; i \neq j$)

$i = 1, \ldots, n$

Regression model (1.1) is said to be *simple, linear in the parameters,* and *linear in the predictor variable.* It is "simple" in that there is only one predictor variable, "linear in the parameters," because no parameter appears as an exponent or is multiplied or divided by another parameter, and "linear in the predictor variable," because this variable appears only in the first power. A model that is linear in the parameters and in the predictor variable is also called a *first-order model.*

Important Features of Model

1. The response Y_i in the ith trial is the sum of two components: (1) the constant term $\beta_0 + \beta_1 X_i$ and (2) the random term ε_i. Hence, Y_i is a random variable.
2. Since $E\{\varepsilon_i\} = 0$, it follows from (A.13c) in Appendix A that:

$$E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \beta_0 + \beta_1 X_i + E\{\varepsilon_i\} = \beta_0 + \beta_1 X_i$$

Note that $\beta_0 + \beta_1 X_i$ plays the role of the constant a in (A.13c).

b. In this study, could the error term variance σ^2 be estimated without fitting a regression line? Explain.

1.40. In fitting regression model (1.1), it was found that observation Y_i fell directly on the fitted regression line (i.e., $Y_i = \hat{Y}_i$). If this case were deleted, would the least squares regression line fitted to the remaining $n - 1$ cases be changed? [*Hint*: What is the contribution of case i to the least squares criterion Q in (1.8)?]

1.41. (Calculus needed.) Refer to the regression model $Y_i = \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$, in Exercise 1.29.

a. Find the least squares estimator of β_1.

b. Assume that the error terms ε_i are independent $N(0, \sigma^2)$ and that σ^2 is known. State the likelihood function for the n sample observations on Y and obtain the maximum likelihood estimator of β_1. Is it the same as the least squares estimator?

c. Show that the maximum likelihood estimator of β_1 is unbiased.

1.42. **Typographical errors.** Shown below are the number of galleys for a manuscript (X) and the dollar cost of correcting typographical errors (Y) in a random sample of recent orders handled by a firm specializing in technical manuscripts. Assume that the regression model $Y_i = \beta_1 X_i + \varepsilon_i$ is appropriate, with normally distributed independent error terms whose variance is $\sigma^2 = 16$.

i:	1	2	3	4	5	6
X_i:	7	12	4	14	25	30
Y_i:	128	213	75	250	446	540

a. State the likelihood function for the six Y observations, for $\sigma^2 = 16$.

b. Evaluate the likelihood function for $\beta_1 = 17$, 18, and 19. For which of these β_1 values is the likelihood function largest?

c. The maximum likelihood estimator is $b_1 = \sum X_i Y_i / \sum X_i^2$. Find the maximum likelihood estimate. Are your results in part (b) consistent with this estimate?

d. Using a computer graphics or statistics package, evaluate the likelihood function for values of β_1 between $\beta_1 = 17$ and $\beta_1 = 19$ and plot the function. Does the point at which the likelihood function is maximized correspond to the maximum likelihood estimate found in part (c)?

Projects

1.43. Refer to the **CDI** data set in Appendix C.2. The number of active physicians in a CDI (Y) is expected to be related to total population, number of hospital beds, and total personal income. Assume that first-order regression model (1.1) is appropriate for each of the three predictor variables.

a. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions.

b. Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?

c. Calculate *MSE* for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

1.44. Refer to the **CDI** data set in Appendix C.2.

a. For each geographic region, regress per capita income in a CDI (Y) against the percentage of individuals in a county having at least a bachelor's degree (X). Assume that

first-order regression model (1.1) is appropriate for each region. State the estimated regression functions.

b. Are the estimated regression functions similar for the four regions? Discuss.

c. Calculate *MSE* for each region. Is the variability around the fitted regression line approximately the same for the four regions? Discuss.

1.45. Refer to the **SENIC** data set in Appendix C.1. The average length of stay in a hospital (Y) is anticipated to be related to infection risk, available facilities and services, and routine chest X-ray ratio. Assume that first-order regression model (1.1) is appropriate for each of the three predictor variables.

a. Regress average length of stay on each of the three predictor variables. State the estimated regression functions.

b. Plot the three estimated regression functions and data on separate graphs. Does a linear relation appear to provide a good fit for each of the three predictor variables?

c. Calculate *MSE* for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

1.46. Refer to the **SENIC** data set in Appendix C.1.

a. For each geographic region, regress average length of stay in hospital (Y) against infection risk (X). Assume that first-order regression model (1.1) is appropriate for each region. State the estimated regression functions.

b. Are the estimated regression functions similar for the four regions? Discuss.

c. Calculate *MSE* for each region. Is the variability around the fitted regression line approximately the same for the four regions? Discuss.

1.47. Refer to **Typographical errors** Problem 1.42. Assume that first-order regression model (1.1) is appropriate, with normally distributed independent error terms whose variance is $\sigma^2 = 16$.

a. State the likelihood function for the six observations, for $\sigma^2 = 16$.

b. Obtain the maximum likelihood estimates of β_0 and β_1, using (1.27).

c. Using a computer graphics or statistics package, obtain a three-dimensional plot of the likelihood function for values of β_0 between $\beta_0 = -10$ and $\beta_0 = 10$ and for values of β_1 between $\beta_1 = 17$ and $\beta_1 = 19$. Does the likelihood appear to be maximized by the maximum likelihood estimates found in part (b)?

2.57. The normal error regression model (2.1) is assumed to be applicable.

a. When testing H_0: $\beta_1 = 5$ versus H_a: $\beta_1 \neq 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

b. When testing H_0: $\beta_0 = 2$, $\beta_1 = 5$ versus H_a: not both $\beta_0 = 2$ and $\beta_1 = 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

2.58. The random variables Y_1 and Y_2 follow the bivariate normal distribution in (2.74). Show that if $\rho_{12} = 0$, Y_1 and Y_2 are independent random variables.

2.59. (Calculus needed.)

a. Obtain the maximum likelihood estimators of the parameters of the bivariate normal distribution in (2.74).

b. Using the results in part (a), obtain the maximum likelihood estimators of the parameters of the conditional probability distribution of Y_1 for any value of Y_2 in (2.80).

c. Show that the maximum likelihood estimators of $\alpha_{1|2}$ and β_{12} obtained in part (b) are the same as the least squares estimators (1.10) for the regression coefficients in the simple linear regression model.

2.60. Show that test statistics (2.17) and (2.87) are equivalent.

2.61. Show that the ratio *SSR/SSTO* is the same whether Y_1 is regressed on Y_2 or Y_2 is regressed on Y_1. [*Hint*: Use (1.10a) and (2.51).]

Projects

2.62. Refer to the **CDI** data set in Appendix C.2 and Project 1.43. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians?

2.63. Refer to the **CDI** data set in Appendix C.2 and Project 1.44. Obtain a separate interval estimate of β_1 for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

2.64. Refer to the **SENIC** data set in Appendix C.1 and Project 1.45. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability of the average length of stay?

2.65. Refer to the **SENIC** data set in Appendix C.1 and Project 1.46. Obtain a separate interval estimate of β_1 for each region. Use a 95 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

2.66. Five observations on Y are to be taken when $X = 4$, 8, 12, 16, and 20, respectively. The true regression function is $E\{Y\} = 20 + 4X$, and the ε_i are independent $N(0, 25)$.

a. Generate five normal random numbers, with mean 0 and variance 25. Consider these random numbers as the error terms for the five Y observations at $X = 4, 8, 12, 16$, and 20 and calculate Y_1, Y_2, Y_3, Y_4, and Y_5. Obtain the least squares estimates b_0 and b_1 when fitting a straight line to the five cases. Also calculate $\hat{Y}_h$ when $X_h = 10$ and obtain a 95 percent confidence interval for $E\{Y_h\}$ when $X_h = 10$.

b. Repeat part (a) 200 times, generating new random numbers each time.

c. Make a frequency distribution of the 200 estimates b_1. Calculate the mean and standard deviation of the 200 estimates b_1. Are the results consistent with theoretical expectations?

d. What proportion of the 200 confidence intervals for $E\{Y_h\}$ when $X_h = 10$ include $E\{Y_h\}$? Is this result consistent with theoretical expectations?

3.23. A linear regression model with intercept $\beta_0 = 0$ is under consideration. Data have been obtained that contain replications. State the full and reduced models for testing the appropriateness of the regression function under consideration. What are the degrees of freedom associated with the full and reduced models if $n = 20$ and $c = 10$?

Projects

3.24. **Blood pressure.** The following data were obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old.

i:	1	2	3	4	5	6	7	8
X_i:	5	8	11	7	13	12	12	6
Y_i:	63	67	74	64	75	69	90	60

a. Assuming normal error regression model (2.1) is appropriate, obtain the estimated regression function and plot the residuals e_i against X_i. What does your residual plot show?

b. Omit case 7 from the data and obtain the estimated regression function based on the remaining seven cases. Compare this estimated regression function to that obtained in part (a). What can you conclude about the effect of case 7?

c. Using your fitted regression function in part (b), obtain a 99 percent prediction interval for a new Y observation at $X = 12$. Does observation Y_7 fall outside this prediction interval? What is the significance of this?

3.25. Refer to the **CDI** data set in Appendix C.2 and Project 1.43. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?

3.26. Refer to the **CDI** data set in Appendix C.2 and Project 1.44. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variances? What other conclusions do you draw from your plots?

3.27. Refer to the **SENIC** data set in Appendix C.1 and Project 1.45.

a. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more apt in one case than in the others?

b. Obtain the fitted regression function for the relation between length of stay and infection risk after deleting cases 47 ($X_{47} = 6.5$, $Y_{47} = 19.56$) and 112 ($X_{112} = 5.9$, $Y_{112} = 17.94$). From this fitted regression function obtain separate 95 percent prediction intervals for new Y observations at $X = 6.5$ and $X = 5.9$, respectively. Do observations Y_{47} and Y_{112} fall outside these prediction intervals? Discuss the significance of this.

3.28. Refer to the **SENIC** data set in Appendix C.1 and Project 1.46. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variances? What other conclusions do you draw from your plots?

3.29. Refer to **Copier maintenance** Problem 1.20.

a. Divide the data into four bands according to the number of copiers serviced (X). Band 1 ranges from $X = .5$ to $X = 2.5$; band 2 ranges from $X = 2.5$ to $X = 4.5$; and so forth. Determine the median value of X and the median value of Y in each of the bands and develop

1	2	3	4	5	6	7	8	9	10	11	12
1	7.13	55.7	4.1	9.0	39.6	279	2	4	207	241	60.0
2	8.82	58.2	1.6	3.8	51.7	80	2	2	51	52	40.0
3	8.34	56.9	2.7	8.1	74.0	107	2	3	82	54	20.0
...	...	...	...	...	...	...	...	...	...	...	...
111	7.70	56.9	4.4	12.2	67.9	129	2	4	85	136	62.9
112	17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9
113	9.41	59.5	3.1	20.6	91.7	29	2	3	20	22	22.9

Data Set C.2 CDI

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The 17 variables are:

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W

Source: Geospatial and Statistical Data Center, University of Virginia.

1	2	3	4	5	6	7	8	9	10
1	Los_Angeles	CA	4060	8863164	32.1	9.7	23677	27700	688936
2	Cook	IL	946	5105067	29.2	12.4	15153	21550	436936
3	Harris	TX	1729	2818199	31.3	7.1	7553	12449	253526
...	...	...	...	...	...	...	...	...	...
438	Montgomery	TN	539	100498	35.7	7.9	87	188	6537
439	Maui	HI	1159	100374	26.2	11.3	192	182	7130
440	Morgan	AL	582	100043	26.3	11.7	122	464	4693

11	12	13	14	15	16	17
70.0	22.3	11.6	8.0	20786	184230	4
73.4	22.8	11.1	7.2	21729	110928	2
74.9	25.4	12.5	5.7	19517	55003	3
...	...	...	...	...	...	...
77.9	16.5	10.8	8.0	13169	1323	3
77.0	17.8	5.7	3.2	18504	1857	4
69.4	15.5	9.4	7.1	16458	1647	3

Data Set C.3 Market Share

Company executives from a large packaged foods manufacturer wished to determine which factors influence the market share of one of its products. Data were collected from a national database (Nielsen) for 36 consecutive months. Each line of the data set has an identification number and provides information on 6 other variables for each month. The data presented here are for September, 1999, through August, 2002. The variables are:

Variable Number	Variable Name	Description
1	Identification number	1–36
2	Market share	Average monthly market share for product (percent)
3	Price	Average monthly price of product (dollars)
4	Gross Nielsen rating points	An index of the amount of advertising exposure that the product received
5	Discount price	Presence or absence of discount price during period: 1 if discount, 0 otherwise
6	Package promotion	Presence or absence of package promotion during period: 1 if promotion present, 0 otherwise
7	Month	Month (Jan–Dec)
8	Year	Year (1999–2002)

1	2	3	4	5	6	7	8
1	3.15	2.198	498	1	1	Sep	1999
2	2.52	2.186	510	0	0	Oct	1999
3	2.64	2.293	422	1	1	Nov	1999
...	...	...	...	...	...	...	...
34	2.80	2.518	270	1	0	Jun	2002
35	2.48	2.497	322	0	1	Jul	2002
36	2.85	2.781	317	1	1	Aug	2002