CHAPTER 4

Drawing Conclusions

The computations that are done in multiple linear regression, including drawing graphs, creation of terms, fitting models, and performing tests, will be similar in most problems. Interpreting the results, however, may differ by problem, even if the outline of the analysis is the same. Many issues play into drawing conclusions, and some of them are discussed in this chapter.

4.1 UNDERSTANDING PARAMETER ESTIMATES

Parameters in mean functions have *units* attached to them. For example, the fitted mean function for the fuel consumption data is

$$\mathrm{E}(Fuel|X) = 154.19 - 4.23\,Tax + 0.47\,Dlic - 6.14\,Income + 18.54\log(Miles)$$

Fuel is measured in gallons, and so all the quantities on the right of this equation must also be in gallons. The intercept is 154.19 gallons. Since *Income* is measured in thousands of dollars, the coefficient for *Income* must be in gallons per thousand dollars of income. Similarly, the units for the coefficient for *Tax* is gallons per cent of tax.

4.1.1 Rate of Change

The usual interpretation of an estimated coefficient is as a rate of change: increasing *Tax* rate by one cent should decrease consumption, all other factors being held constant, by about 4.23 gallons per person. This assumes that a predictor can in fact be changed without affecting the other terms in the mean function and that the available data will apply when the predictor is so changed. The fuel data are *observational* since the assignment of values for the predictors was not under the control of the analyst, so whether increasing taxes would *cause*

Applied Linear Regression, Third Edition, by Sanford Weisberg
ISBN 0-471-66379-4

a decrease in fuel consumption cannot be assessed from these data. From these data, we can observe *association* but not cause: states with higher tax rates are *observed* to have lower fuel consumption. To draw conclusions concerning the effects of changing tax rates, the rates must in fact be changed and the results observed.

The coefficient estimate of log(*Miles*) is 18.55, meaning that a change of one unit in log(*Miles*) is associated with an 18.55 gallon per person increase in consumption. States with more roads have higher per capita fuel consumption. Since we used base-two logarithms in this problem, increasing log(*Miles*) by one unit means that the value of *Miles doubles*. If we double the amount of road in a state, we expect to increase fuel consumption by about 18.55 gallons per person. If we had used base-ten logarithms, then the fitted mean function would be

$$\mathrm{E}(Fuel|X) = 154.19 - 4.23\,Tax + 0.47\,Dlic - 6.14\,Income + 61.61\log_{10}(Miles)$$

The only change in the fitted model is for the coefficient for the log of Miles, which is now interpreted as a change in expected *Fuel* consumption when $\log_{10}(Miles)$ increases by one unit, or when *Miles* is multiplied by 10.

4.1.2 Signs of Estimates

The sign of a parameter estimate indicates the direction of the relationship between the term and the response. In multiple regression, if the terms are correlated, the sign of a coefficient may change depending on the other terms in the model. While this is mathematically possible and, occasionally, scientifically reasonable, it certainly makes interpretation more difficult. Sometimes this problem can be removed by redefining the terms into new linear combinations that are easier to interpret.

4.1.3 Interpretation Depends on Other Terms in the Mean Function

The value of a parameter estimate not only depends on the other terms in a mean function but it can also change if the other terms are replaced by linear combinations of the terms.

Berkeley Guidance Study

Data from the Berkeley Guidance Study on the growth of boys and girls are given in Problem 3.1. As in Problem 3.1, we will view *Soma* as the response, but consider the three predictors *WT2*, *WT9*, *WT18* for the $n = 70$ girls in the study. The scatterplot matrix for these four variables is given in Figure 4.1. First look at the last row of this figure, giving the marginal response plots of *Soma* versus each of the three potential predictors. For each of these plots, we see that *Soma* is increasing with the potential predictor on the average, although the relationship is strongest at the oldest age and weakest at the youngest age. The two-dimensional plots of each pair of predictors suggest that the predictors are correlated among themselves. Taken together, we have evidence that the regression on all three predictors cannot

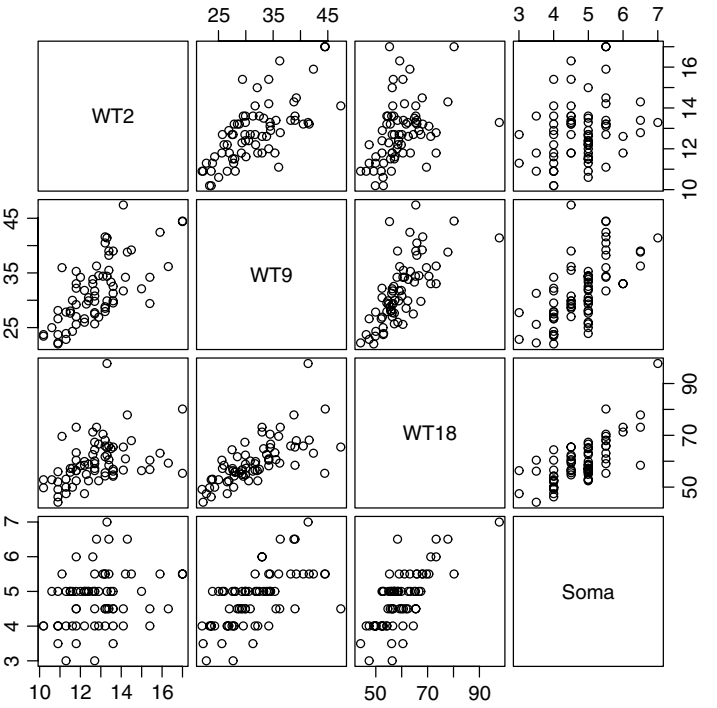

FIG. 4.1 Scatterplot matrix for the girls in the Berkeley Guidance Study.

be viewed as just the sum of the three separate simple regressions because we must account for the correlations between the terms.

We will proceed with this example using the three original predictors as terms and *Soma* as the response. We are encouraged to do this because of the appearance of the scatterplot matrix. Since each of the two-dimensional plots appear to be well summarized by a straight-line mean function, we will see later that this suggests that the regression of the response on the original predictors without transformation is likely to be appropriate.

The parameter estimates for the regression of *Soma* on *WT2*, *WT9*, and *WT18* given in the column marked "Model 1" in Table 4.1 leads to the unexpected conclusion that heavier girls at age two may tend to be thinner, have lower expected somatotype, at age 18. We reach this conclusion because the t-statistic for testing the coefficient equal to zero, which is not shown in the table, has a significance level of about 0.06. The sign, and the weak significance, may be due to the correlations between the terms. In place of the preceding variables, consider the following:

$$WT2 = \text{ Weight at age 2}$$

$$DW9 = WT9 - WT2 = \text{ Weight gain from age 2 to 9}$$

$$DW18 = WT18 - WT9 = \text{ Weight gain from age 9 to 18}$$

TABLE 4.1 Regression of *Soma* on Different Combinations of Three Weight Variables for the $n = 70$ Girls in the Berkeley Guidance Study

Term	Model 1	Model 2	Model 3
(Intercept)	1.5921	1.5921	1.5921
WT2	−0.1156	−0.0111	−0.1156
WT9	0.0562		0.0562
WT18	0.0483		0.0483
DW9		0.1046	NA
DW18		0.0483	NA

Since all three original terms measure weight, combining them in this way is reasonable. If the variables measured different quantities, then combining them could lead to conclusions that are even less useful than those originally obtained. The parameter estimates for *Soma* on *WT2*, *DW9*, and *DW18* are given in the column marked "Model 2" in Table 4.1. Although not shown in the table, summary statistics for the regression like R^2 and $\hat{\sigma}^2$ are identical for all the mean functions in Table 4.1 but coefficient estimates and t-tests are not the same. For example, the slope estimate for *WT2* is about -0.12, with $t = -1.87$ in the column "Model 1," while in Model 2, the estimate is about one-tenth the size, and the t-value is -0.21. In the former case, the effect of *WT2* appears plausible, while in the latter it does not. Although the estimate is negative in each, we would be led in the latter case to conclude that the effect of *WT2* is negligible. Thus, interpretation of the effect of a variable depends not only on the other variables in a model but also upon which linear transformation of those variables is used.

Another interesting feature of Table 4.1 is that the estimate for *WT18* in Model 1 is identical to the estimate for *DW18* in Model 2. In Model 1, the estimate for *WT18* is the effect on *Soma* of changing *WT18* by one unit, with all other terms held fixed. In Model 2, the estimate for *DW18* is the change in *Soma* when *DW18* changes by one unit, when all other terms are held fixed. *But the only way* DW18 = WT18 − WT9 *can be changed by one unit with the other variables including* WT9 = DW9 − WT2 *held fixed is by changing* WT18 *by one unit.* Consequently, the terms *WT18* in Model 1 and *DW18* in Model 2 play identical roles and therefore we get the same estimates.

The linear transformation of the three weight variables we have used so far could be replaced by other linear combinations, and, depending on the context, others might be preferred. For example, another set might be

$$AVE = (WT2 + WT9 + WT18)/3$$

$$LIN = WT18 - WT2$$

$$QUAD = WT2 - 2WT9 + WT18$$

This transformation focuses on the fact that *WT2*, *WT9* and *WT18* are ordered in time and are more or less equally spaced. Pretending that the weight measurements

are equally spaced, *AVE*, *LIN* and *QUAD* are, respectively, the average, linear, and quadratic time trends in weight gain.

4.1.4 Rank Deficient and Over-Parameterized Mean Functions

In the last example, several combinations of the basic predictors *WT2*, *WT9*, and *WT18* were studied. One might naturally ask what would happen if more than three combinations of these predictors were used in the same regression model. As long as we use linear combinations of the predictors, as opposed to nonlinear combinations or transformations of them, we cannot use more than three, the number of linearly independent quantities.

To see why this is true, consider adding *DW9* to the mean function including *WT2*, *WT9* and *WT18*. As in Chapter 3, we can learn about adding *DW9* using an added-variable plot of the residuals from the regression of *Soma* on *WT2*, *WT9* and *WT18* versus the residuals from the regression of *DW9* on *WT2*, *WT9* and *WT18*. Since *DW9* can be written as an exact linear combination of the other predictors, $DW9 = WT9 - WT2$, the residuals from this second regression are all exactly zero. A slope coefficient for *DW9* is thus not defined after adjusting for the other three terms. We would say that the four terms *WT2*, *WT9*, *WT18*, and *DW9* are *linearly dependent*, since one can be determined exactly from the others. The three variables *WT2*, *WT9* and *WT18* are *linearly independent* because one of them cannot be determined exactly by a linear combination of the others. The maximum number of linearly independent terms that could be included in a mean function is called the *rank* of the data matrix $\mathbf{X}$.

Model 3 in Table 4.1 gives the estimates produced in a computer package when we tried to fit using an intercept and the five terms *WT2*, *WT9*, *WT18*, *DW9*, and *DW18*. Most computer programs, including this one, will select the first three, and the estimated coefficients for them. For the remaining terms, this program sets the estimates to "NA," a code for a missing value; the word *aliased* is sometimes used to indicate a term that is a linear combination of terms already in the mean function, and so a coefficient for it is not estimable.

Mean functions that are over-parameterized occur most often in designed experiments. The simplest example is the one-way design. Suppose that a unit is assigned to one of three treatment groups, and let $X_1 = 1$ if the unit is in group one and zero otherwise, $X_2 = 1$ if the unit is in group two and zero otherwise, and $X_3 = 1$ if the unit is in group three and zero otherwise. For each unit, we must have $X_1 + X_2 + X_3 = 1$ since each unit is in only one of the three groups. We therefore cannot fit the model

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

because the sum of the X_j is equal to the column of ones, and so, for example, $X_3 = 1 - X_1 - X_2$. To fit a model, we must do something else. The options are: (1) place a constraint like $\beta_1 + \beta_2 + \beta_3 = 0$ on the parameters; (2) exclude one of the X_j from the model, or (3) leave out an explicit intercept. All of these options will in some sense be equivalent, since the same R^2, σ^2 and overall F-test and

predictions will result. Of course, some care must be taken in using parameter estimates, since these will surely depend on the parameterization used to get a full rank model. For further reading on matrices and models of less than full rank, see, for example, Searle (1971, 1982).

4.1.5 Tests

Even if the fitted model were correct and errors were normally distributed, tests and confidence statements for parameters are difficult to interpret because correlations among the terms lead to a multiplicity of possible tests. Sometimes, tests of effects adjusted for other variables are clearly desirable, such as in assessing a treatment effect after adjusting for other variables to reduce variability. At other times, the order of fitting is not clear, and the analyst must expect ambiguous results. In most situations, the only true test of significance is repeated experimentation.

4.1.6 Dropping Terms

Suppose we have a sample of n rectangles from which we want to model $\log(Area)$ as a function of $\log(Length)$, perhaps through the simple regression mean function

$$\mathrm{E}(\log(Area)|\log(Length)) = \eta_0 + \eta_1 \log(Length) \tag{4.1}$$

From elementary geometry, we know that $Area = Length \times Width$, and so the "true" mean function for $\log(Area)$ is

$$\mathrm{E}(\log(Area)|\log(Length), \log(Width)) = \beta_0 + \beta_1 \log(Length) + \beta_2 \log(Width) \tag{4.2}$$

with $\beta_0 = 0$, and $\beta_1 = \beta_2 = 1$. The questions of interest are: (1) can the incorrect mean function specified by (4.1) provide a useful approximation to the true mean function (4.2), and if so, (2) what are the relationships between ηs, in (4.1) and the βs in (4.2)?

The answers to these questions comes from Appendix A.2.4. Suppose that the true mean function were

$$\mathrm{E}(Y|X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2) = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_1 + \boldsymbol{\beta}_2'\mathbf{x}_2 \tag{4.3}$$

but we want to fit a mean function with X_1 only. The mean function for $Y|X_1$ is obtained by averaging (4.3) over X_2,

$$\begin{aligned} \mathrm{E}(Y|X_1 = \mathbf{x}_1) &= \mathrm{E}\left[\mathrm{E}(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1\right] \\ &= \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_1 + \boldsymbol{\beta}_2'\mathrm{E}(X_2|X_1 = \mathbf{x}_1) \end{aligned} \tag{4.4}$$

We cannot, in general, simply drop a set of terms from a correct mean function, but we need to substitute the conditional expectation of the terms dropped given the terms that remain in the mean function.

In the context of the rectangles example, we get

$$\mathrm{E}(\log(Area)|\log(Length)) = \eta_0 + \eta_1 \log(Length) + \beta_2 \mathrm{E}(\log(Width)|\log(Length)) \tag{4.5}$$

The answers to the questions posed depend on the mean function for the regression of log(*Width*) on log(*Length*). This conditional expectation has little to do with the area of rectangles, but much to do with the way we obtain a sample of rectangles to use in our study. We will consider three cases.

In the first case, imagine that each of the rectangles in the study is formed by sampling a log(*Length*) and a log(*Width*) from independent distributions. If the mean of the log(*Width*) distribution is W, then by independence

$$\mathrm{E}(\log(Width)|\log(Length)) = \mathrm{E}(\log(Width)) = W$$

Substituting into (4.5),

$$\begin{aligned}\mathrm{E}(\log(Area)|\log(Length)) &= \beta_0 + \beta_1\log(Length) + \beta_2 W\\ &= (\beta_0 + \beta_2 W) + \beta_1\log(Length)\\ &= W + \log(Length)\end{aligned}$$

where the last equation follows by substituting $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$. For this case, the mean function (4.1) would be appropriate for the regression of log(*Area*) on log(*Width*). The intercept for the mean function (4.1) would be W, and so it depends on the distribution of the widths in the data. The slope for log(*Length*) is the same for fitting (4.1) or (4.2).

In the second case, suppose that

$$E(\log(Width)|\log(Length)) = \gamma_0 + \gamma_1\log(Length)$$

so the mean function for the regression of log(*Width*) on log(*Length*) is a straight line. This could occur, for example, if the rectangles in our study were obtained by sampling from a family of similar rectangles, so the ratio *Width*/*Length* is the same for all rectangles in the study. Substituting this into (4.5) and simplifying gives

$$\begin{aligned}\mathrm{E}(\log(Area)|\log(Length)) &= \beta_0 + \beta_1\log(Length) + \beta_2(\gamma_0 + \gamma_1\log(Length))\\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)\log(Length)\\ &= \gamma_0 + (1 + \gamma_1)\log(Length)\end{aligned}$$

Once again fitting using (4.1) will be appropriate, but the values of $\eta_0 = \gamma_0$ and $\eta_1 = 1 + \gamma_1$ depend on the parameters of the regression of log(*Width*) on log(*Length*). The γs are a characteristic of the sampling plan, not of rectangles. Two experimenters who sample rectangles of different shapes will end up estimating different parameters.

For a final case, suppose that the mean function

$$E(\log(Width)|\log(Length)) = \gamma_0 + \gamma_1\log(Length) + \gamma_2\log(Length)^2$$

is quadratic. Substituting into (4.5), setting $\beta_0 = 0, \beta_1 = \beta_2 = 1$ and simplifying gives

$$\begin{aligned}\mathrm{E}(\log(Area)|\log(Length)) &= \beta_0 + \beta_1\log(Length) \\ &\quad + \beta_2\left(\gamma_0 + \gamma_1\log(Length) + \gamma_2\log(Length)^2\right) \\ &= \gamma_0 + (1+\gamma_1)\log(Length) + \gamma_2\log(Length)^2\end{aligned}$$

which is a quadratic function of $\log(Length)$. If the mean function is quadratic, or any other function beyond a straight line, then fitting (4.1) is inappropriate.

From the above three cases, we see that both the mean function and the parameters for the response depend on the mean function for the regression of the removed terms on the remaining terms. If the mean function for the regression of the removed terms on the retained terms is not linear, then a linear mean function will not be appropriate for the regression problem with fewer terms.

Variances are also affected when terms are dropped. Returning to the true mean function given by (4.3), the general result for the regression of Y on X_1 alone is, from Appendix A.2.4,

$$\begin{aligned}\mathrm{Var}(Y|X_1 = \mathbf{x}_1) &= \mathrm{E}\left[\mathrm{Var}(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1\right] \\ &\quad + \mathrm{Var}\left[\mathrm{E}(Y|X_1 = \mathbf{x}_1, X_2)|X_1 = \mathbf{x}_1\right] \\ &= \sigma^2 + \boldsymbol{\beta}_2'\mathrm{Var}(X_2|X_1 = \mathbf{x}_1)\boldsymbol{\beta}_2 \qquad (4.6)\end{aligned}$$

In the context of the rectangles example, $\beta_2 = 1$ and we get

$$\mathrm{Var}(\log(Area)|\log(Length)) = \sigma^2 + \mathrm{Var}(\log(Width)|\log(Length))$$

Although fitting (4.1) can be appropriate if $\log(Width)$ and $\log(Length)$ are linearly related, the errors for this mean function can be much larger than those for (4.2) if $\mathrm{Var}(\log(Width)|\log(Length))$ is large. If $\mathrm{Var}(\log(Width)|\log(Length))$ is small enough, then fitting (4.2) can actually give answers that are nearly as accurate as fitting with the true mean function (4.2).

4.1.7 Logarithms

If we start with the simple regression mean function,

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x$$

a useful way to interpret the coefficient β_1 is as the first derivative of the mean function with respect to x,

$$\frac{d\mathrm{E}(Y|X = x)}{dx} = \beta_1$$

We recall from elementary geometry that the first derivative is the rate of change, or the slope of the tangent to a curve, at a point. Since the mean function for

simple regression is a straight line, the slope of the tangent is the same value β_1 for any value of x, and β_1 completely characterizes the change in the mean when the predictor is changed for any value of x.

When the predictor is replaced by $\log(x)$, the mean function as a function of x

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 \log(x)$$

is no longer a straight line, but rather it is a curve. The tangent at the point $x > 0$ is

$$\frac{d\mathrm{E}(Y|X = x)}{dx} = \frac{\beta_1}{x}$$

The slope of the tangent is different for each x and the effect of changing x on $\mathrm{E}(Y|X = x)$ is largest for small values of x and gets smaller as x is increased.

When the response is in log scale, we can get similar approximate results by exponentiating both sides of the equation:

$$\mathrm{E}(\log(Y)|X = x) = \beta_0 + \beta_1 x$$
$$\mathrm{E}(Y|X = x) \approx e^{\beta_0} e^{\beta_1 x}$$

Differentiating this second equation gives

$$\frac{d\mathrm{E}(Y|X = x)}{dx} = \beta_1 \mathrm{E}(Y|X = x)$$

The rate of change at x is thus equal to β_1 times the mean at x. We can also write

$$\frac{d\mathrm{E}(Y|X = x)/dx}{\mathrm{E}(Y|X = x)} = \beta_1$$

is constant, and so β_1 can be interpreted as the constant rate of change in the response per unit of response.

4.2 EXPERIMENTATION VERSUS OBSERVATION

There are fundamentally two types of predictors that are used in a regression analysis, *experimental* and *observational*. Experimental predictors have values that are under the control of the experimenter, while for observational predictors, the values are observed rather than set. Consider, for example, a hypothetical study of factors determining the yield of a certain crop. Experimental variables might include the amount and type of fertilizers used, the spacing of plants, and the amount of irrigation, since each of these can be assigned by the investigator to the units, which are plots of land. Observational predictors might include characteristics of the plots in the study, such as drainage, exposure, soil fertility, and weather variables. All of these are beyond the control of the experimenter, yet may have important effects on the observed yields.

The primary difference between experimental and observational predictors is in the inferences we can make. From experimental data, we can often infer causation.

If we assign the level of fertilizer to plots, usually on the basis of a randomization scheme, and observe differences due to levels of fertilizer, we can infer that the fertilizer is causing the differences. Observational predictors allow weaker inferences. We might say that weather variables are associated with yield, but the causal link is not available for variables that are not under the experimenter's control. Some experimental designs, including those that use randomization, are constructed so that the effects of observational factors can be ignored or used in analysis of covariance (see, e.g., Cox, 1958; Oehlert, 2000).

Purely observational studies that are not under the control of the analyst can only be used to predict or model the events that were observed in the data, as in the fuel consumption example. To apply observational results to predict future values, additional assumptions about the behavior of future values compared to the behavior of the existing data must be made. From a purely observational study, we cannot infer a causal relationship without additional information external to the observational study.

Feedlots

A *feedlot* is a farming operation that includes large number of cattle, swine or poultry in a small area. Feedlots are efficient producers of animal products, and can provide high-paying skilled jobs in rural areas. They can also cause environmental problems, particularly with odors, ground water pollution, and noise.

Taff, Tiffany, and Weisberg (1996) report a study on the effect of feedlots on property values. This study was based on all 292 rural residential property sales in two southern Minnesota counties in 1993–94. Regression analysis was used. The response was sale price. Predictors included house characteristics such as size, number of bedrooms, age of the property, and so on. Additional predictors described the relationship of the property to existing feedlots, such as distance to the nearest feedlot, number of nearby feedlots, and related features of the feedlots such as their size. The "feedlot effect" could be inferred from the coefficients for the feedlot variables.

In the analysis, the coefficient estimates for feedlot effects were generally positive and judged to be nonzero, meaning that close proximity to feedlots was associated with an *increase* in sale prices. While association of the opposite sign was expected, the positive sign is plausible if the positive economic impact of the feedlot outweighs the negative environmental impact. The positive effect is estimated to be small, however, and equal to 5% or less of the sale price of the homes in the study.

These data are purely observational, with no experimental predictors. The data collectors had no control over the houses that actually sold, or siting of feedlots. Consequently, any inference that nearby feedlots *cause* increases in sale price is unwarranted from this study. Given that we are limited to association, rather than causation, we might next turn to whether we can generalize the results. Can we infer the same association to houses that were *not* sold in these counties during this period? We have no way of knowing from the data if the same

relationship would hold for homes that did not sell. For example, some homeowners may have perceived that they could not get a reasonable price and may have decided not to sell. This would create a bias in favor of a positive effect of feedlots.

Can we generalize geographically, to other Minnesota counties or to other places in the Midwest United States? The answer to this question depends on the characteristics of the two counties studied. Both are rural counties with populations of about 17,000. Both have very low property values with median sale price in this period of less than $50,000. Each had different regulations for operators of feedlots, and these regulations could impact pollution problems. Applying the results to a county with different demographics or regulations cannot be justified by these data alone, and additional information and assumptions are required.

Joiner (1981) coined the picturesque phrase *lurking variable* to describe a predictor variable not included in a mean function that is correlated with terms in the mean function. Suppose we have a regression with predictors X that are included in the regression and a lurking variable L not included in the study, and that the true regression mean function is

$$\mathrm{E}(Y|X = \mathbf{x}, L = \ell) = \beta_0 + \sum_{j-1}^{p} \beta_j x_j + \delta\ell \tag{4.7}$$

with $\delta \neq 0$. We assume that X and L are correlated and for simplicity we assume further that $\mathrm{E}(L|X = \mathbf{x}) = \gamma_0 + \sum \gamma_j x_j$. When we fit the incorrect mean function that ignores the lurking variable, we get, from Section 4.1.6,

$$\begin{aligned}\mathrm{E}(Y|X = \mathbf{x}) &= \beta_0 + \sum_{j-1}^{p} \beta_j x_j + \delta\mathrm{E}(L|X = \mathbf{x}) \\ &= (\beta_0 + \delta\gamma_0) + \sum_{j-1}^{p} (\beta_j + \delta\gamma_j)x_j \end{aligned} \tag{4.8}$$

Suppose we are particularly interested in inferences about the coefficient for X_1, and, unknown to us, β_1 in (4.7) is equal to zero. If we were able to fit with the lurking variable included, we would probably conclude that X_1 is not an important predictor. If we fit the incorrect mean function (4.8), the coefficient for X_1 becomes $(\beta_1 + \delta\gamma_1)$, which will be non zero if $\gamma_1 \neq 0$. The lurking variable masquerades as the variable of interest to give an incorrect inference. A lurking variable can also hide the effect of an important variable if, for example, $\beta_1 \neq 0$ but $\beta_1 + \delta\gamma_1 = 0$.

All large observational studies like this feedlot study potentially have lurking variables. For this study, a casino had recently opened near these counties, creating many jobs and a demand for housing that might well have overshadowed any effect of feedlots. In experimental data with random assignment, the potential effects of lurking variables are greatly decreased, since the random assignment guarantees that

the correlation between the terms in the mean function and any lurking variable is small or zero.

The interpretation of results from a regression analysis depend on the details of the data design and collection. The feedlot study has extremely limited scope, and is but one element to be considered in trying to understand the effect of feedlots on property values. Studies like this feedlot study are easily misused. As recently as spring 2004, the study was cited in an application for a permit to build a feedlot in Starke county, Indiana, claiming that the study supports the positive effect of feedlots on property values, confusing association with causation, and inferring generalizability to other locations without any logical foundation for doing so.

4.3 SAMPLING FROM A NORMAL POPULATION

Much of the intuition for the use of least squares estimation is based on the assumption that the observed data are a sample from a multivariate normal population. While the assumption of multivariate normality is almost never tenable in practical regression problems, it is worthwhile to explore the relevant results for normal data, first assuming random sampling and then removing that assumption.

Suppose that all of the observed variables are normal random variables, and the observations on each case are independent of the observations on each other case. In a two-variable problem, for the ith case observe (x_i, y_i), and suppose that

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \tag{4.9}$$

Equation (4.9) says that x_i and y_i are each realizations of normal random variables with means μ_x and μ_y, variances σ_x^2 and σ_y^2 and correlation ρ_{xy}. Now, suppose we consider the conditional distribution of y_i given that we have already observed the value of x_i. It can be shown (see e.g., Lindgren, 1993; Casella and Berger, 1990) that the conditional distribution of y_i given x_i, is normal and,

$$y_i|x_i \sim \mathrm{N}\left(\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), \sigma_y^2(1-\rho_{xy}^2)\right) \tag{4.10}$$

If we define

$$\beta_0 = \mu_y - \beta_1\mu_x \qquad \beta_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x} \qquad \sigma^2 = \sigma_y^2(1-\rho_{xy}^2) \tag{4.11}$$

then the conditional distribution of y_i given x_i is simply

$$y_i|x_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2) \tag{4.12}$$

which is essentially the same as the simple regression model with the added assumption of normality.

Given random sampling, the five parameters in (4.9) are estimated, using the notation of Table 2.1, by

$$\begin{array}{lll} \hat{\mu}_x = \overline{x} & \hat{\sigma}_x^2 = \mathrm{SD}_x^2 & \hat{\rho}_{xy} = r_{xy} \\ \hat{\mu}_y = \overline{y} & \hat{\sigma}_y^2 = \mathrm{SD}_y^2 & \end{array} \tag{4.13}$$

Estimates of β_0 and β_1 are obtained by substituting estimates from (4.13) for parameters in (4.11), so that $\hat{\beta}_1 = r_{xy}\mathrm{SD}_y/\mathrm{SD}_x$, and so on, as derived in Chapter 2. However, $\hat{\sigma}^2 = [(n-1)/(n-2)]\mathrm{SD}_y^2(1-r_{xy}^2)$ to correct for degrees of freedom.

If the observations on the ith case are y_i and a $p \times 1$ vector $\mathbf{x}_i$ not including a constant, multivariate normality is shown symbolically by

$$\begin{pmatrix} \mathbf{x}_i \\ y_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy} & \sigma_y^2 \end{pmatrix} \right) \tag{4.14}$$

where $\boldsymbol{\Sigma}_{xx}$ is a $p \times p$ matrix of variances and covariances between the elements of $\mathbf{x}_i$ and $\boldsymbol{\Sigma}_{xy}$ is a $p \times 1$ vector of covariances between $\mathbf{x}_i$ and y_i. The conditional distribution of y_i given x_i is then

$$y_i|\mathbf{x}_i \sim \mathrm{N}\left((\mu_y - \boldsymbol{\beta}^{*\prime}\boldsymbol{\mu}_x) + \boldsymbol{\beta}^{*\prime}\mathbf{x}_i, \sigma^2\right) \tag{4.15}$$

If $\mathcal{R}^2$ is the population multiple correlation,

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}; \quad \sigma^2 = \sigma_y^2\boldsymbol{\Sigma}_{xy}'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} = \sigma_y^2(1-\mathcal{R}^2)$$

The formulas for $\hat{\boldsymbol{\beta}}^*$ and σ^2 and the formulas for their least squares estimators differ only by the substitution of estimates for parameters, with $(n-1)^{-1}(\mathcal{X}'\mathcal{X})$ estimating $\boldsymbol{\Sigma}_{xx}$, and $(n-1)^{-1}(\mathcal{X}'\mathcal{Y})$ estimating $\boldsymbol{\Sigma}_{xy}$.

4.4 MORE ON R^2

The conditional distribution in (4.10) or (4.15) does not depend on random sampling, but only on normal distributions, so whenever multivariate normality seems reasonable, a linear regression model is suggested for the conditional distribution of one variable, given the others. However, if random sampling is not used, some of the usual summary statistics, including R^2, lose their connection to population parameters.

Figure 4.2a repeats Figure 1.1, the scatterplot of *Dheight* versus *Mheight* for the heights data. These data closely resemble a bivariate normal sample, and so $R^2 = 0.24$ estimates the population $\mathcal{R}^2$ for this problem. Figure 4.2b repeats this last figure, except that all cases with *Mheight* between 61 and 64 inches—the lower and upper quartile of the mother's heights rounded to the nearest inch—have been removed form the data. The OLS regression line appears similar, but the value of $R^2 = 0.37$ is about 50% larger. By removing the middle of the data, we have made

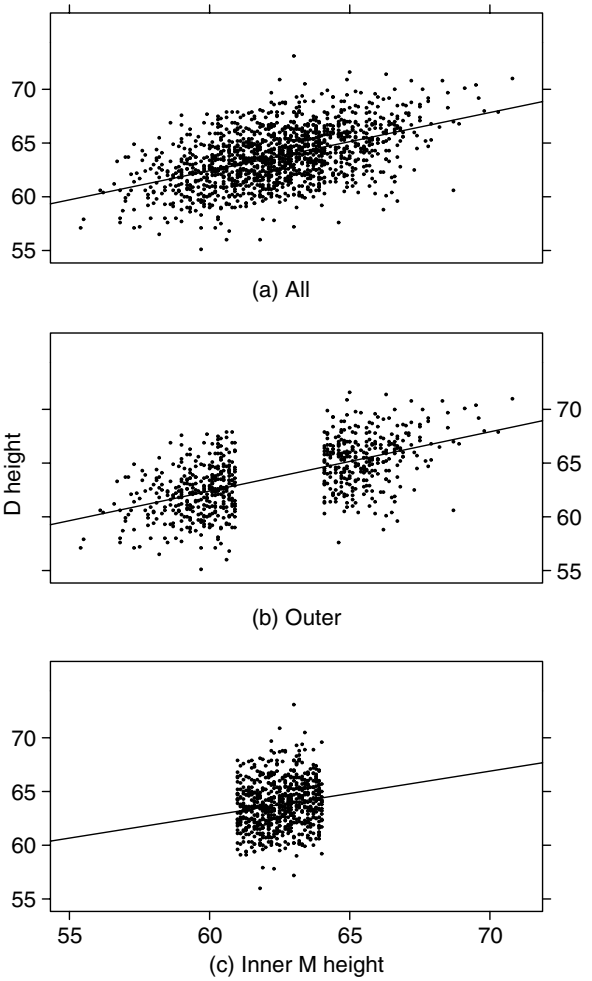

FIG. 4.2 Three views of the heights data.

R^2 larger, and it no longer estimates a population value. Similarly, in Figure 4.2c, we exclude all the cases with *Mheight* outside the quartiles, and get $R^2 = 0.027$, and the relationship between *Dheight* and *Mheight* virtually disappears.

This example points out that even in the unusual event of analyzing data drawn from a multivariate normal population, if sampling of the population is not random, the interpretation of R^2 may be completely misleading, as this statistic will be strongly influenced by the method of sampling. In particular, a few cases with unusual values for the predictors can largely determine the observed value of this statistic.

We have seen that we can manipulate the value of R^2 merely by changing our sampling plan for collecting data: if the values of the terms are widely dispersed,

then R^2 will tend to be too large, while if the values are over a very small range, then R^2 will tend to be too small. Because the notion of proportion of variability explained is so useful, a diagnostic method is needed to decide if it is a useful concept in any particular problem.

4.4.1 Simple Linear Regression and R^2

In simple regression linear problems, we can always determine the appropriateness of R^2 as a summary by examining the summary graph of the response versus the predictor. If the plot looks like a sample from a bivariate normal population, as in Figure 4.2a, then R^2 is a useful measure. The less the graph looks like this figure, the less useful is R^2 as a summary measure.

Figure 4.3 shows six summary graphs. Only for the first three of them is R^2 a useful summary of the regression problem. In Figure 4.3e, the mean function appears curved rather than straight so correlation is a poor measure of dependence. In Figure 4.3d the value of R^2 is virtually determined by one point, making R^2 necessarily unreliable. The regular appearance of the remaining plot suggests a different type of problem. We may have several identifiable groups of points caused by a lurking variable not included in the mean function, such that the mean function for each group has a negative slope, but when groups are combined the slope becomes positive. Once again R^2 is not a useful summary of this graph.

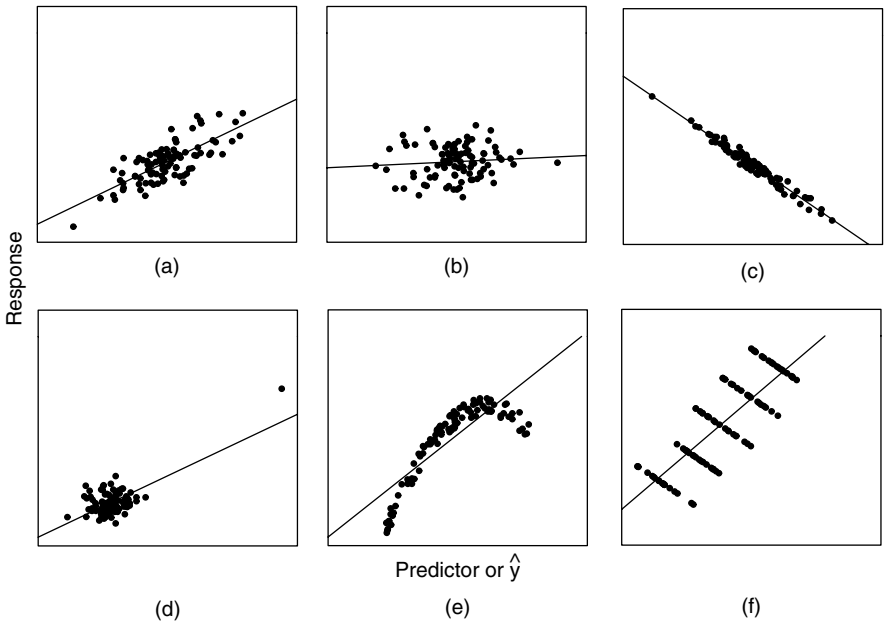

FIG. 4.3 Six summary graphs. R^2 is an appropriate measure for a–c, but inappropriate for d–f.

4.4.2 Multiple Linear Regression

In multiple linear regression, R^2 can also be interpreted as the square of the correlation in a summary graph, this time of Y versus fitted values $\hat{Y}$. This plot can be interpreted exactly the same way as the plot of the response versus the single term in simple linear regression to decide on the usefulness of R^2 as a summary measure.

For other regression methods such as nonlinear regression, we can define R^2 to be the square of the correlation between the response and the fitted values, and use this summary graph to decide if R^2 is a useful summary.

4.4.3 Regression through the Origin

With regression through the origin, the proportion of variability explained is given by $1 - SSreg/\sum y_i^2$, using uncorrected sums of squares. This quantity is *not invariant under location change*, so, for example, if units are changed from Fahrenheit to Celsius, you will get a different value for the proportion of variability explained. For this reason, use of an R^2-like measure for regression through the origin is not recommended.

4.5 MISSING DATA

In many problems, some variables will be unrecorded for some cases. The methods we study in this book generally assume and require complete data, without any missing values. The literature on analyzing incomplete data problems is very large, and our goal here is more to point out the issues than to provide solutions. Two recent books on this topic are Little and Rubin (1987) and Schafer (1997).

4.5.1 Missing at Random

The most common solution to missing data problems is to delete either cases or variables so the resulting data set is complete. Many software packages delete partially missing cases by default, and fit regression models to the remaining, complete, cases. This is a reasonable approach as long as the fraction of cases deleted is small enough, and the cause of values being unobserved is unrelated to the relationships under study. This would include data lost through an accident like dropping a test tube, or making an illegible entry in a logbook. If the reason for not observing values depends on the values that would have been observed, then the analysis of data may require modeling the cause of the failure to observe values. For example, if values of a measurement are unrecorded if the value is less than the minimum detection limit of an instrument, then the value is missing because the value that should have been observed is too small. A simple expedient in this case that is sometimes helpful is to substitute a value less than or equal to the detection limit for the unobserved values. This expedient is not always entirely satisfactory because substituting, or imputing, a fixed value for the unobserved quantity can reduce the variation on the filled-in variable, and yield misleading inferences.

As a second example, suppose we have a clinical trial that enrolls subjects with a particular medical condition, assigns each subject a treatment, and then the subjects are followed for a period of time to observe their response, which may be time until a particular landmark occurs, such as improvement of the medical condition. Subjects who do not respond well to the treatment may drop out of the study early, while subjects who do well may be more likely to remain in the study. Since the probability of observing a value depends on the value that would have been observed, simply deleting subjects who drop out early can easily lead to incorrect inferences because the successful subjects will be overrepresented among those who complete the study.

In many clinical trials, the response variable is not observed because the study ends, not because of patient characteristics. In this case, we call the response times *censored*; so for each patient, we know either the time to the landmark or the time to censoring. This is a different type of missing data problem, and analysis needs to include both the uncensored and censored observations. Book-length treatments of censored survival data are given by Kalbfleisch and Prentice (1980) and Cox and Oakes (1984), among others.

As a final example, consider a cross-cultural demographic study. Some demographic variables are harder to measure than others, and some variables, such as the rate of employment for women over the age of 15, may not be available for less-developed countries. Deleting countries that do not have this variable measured could change the population that is studied by excluding less-developed countries.

Rubin (1976) defined data to be *missing at random* (MAR) if the failure to observe a value does not depend on the value that would have been observed. With MAR data, case deletion can be a useful option. Determining whether an assumption of MAR is appropriate for a particular data set is an important step in the analysis of incomplete data.

4.5.2 Alternatives

All the alternatives we briefly outline here require strong assumptions concerning the data that may be impossible to check in practice.

Suppose first that we combine the response and predictors into a single vector Z. We assume that the distribution of Z is fully known, apart from unknown parameters. The simplest assumption is that $Z \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$. If we had reasonable estimates of $\boldsymbol{\mu}$ and Σ, then we could use (4.15) to estimate parameters for the regression of the response on the other terms. The *EM algorithm* (Dempster, Laird, and Rubin , 1977) is a computational method that is used to estimate the parameters of the known joint distribution based on data with missing values.

Alternatively, given a model for the data like multivariate normality, one could impute values for the missing data and then analyze the completed data as if it were fully observed. *Multiple imputation* carries this one step further by creating several imputed data sets that, according to the model used, are plausible, filled-in data sets, and then "average" the analyses of the filled-in data sets. Software for both imputation and the EM algorithm for maximum likelihood estimate is available

in several standard statistical packages, including the "missing" package in S-plus and the "MI" procedure in SAS.

The third approach is more comprehensive, as it requires building a model for the process of interest and the missing data process simultaneously. Examples of this approach are given by Ibrahim, Lipsitz, and Horton (2001), and Tang, Little, and Raghunathan (2003).

The data described in Table 4.2 provides an example. Allison and Cicchetti (1976) presented data on sleep patterns of 62 mammal species along with several other possible predictors of sleep. The data were in turn compiled from several other sources, and not all values are measured for all species. For example, *PS*, the number of hours of paradoxical sleep, was measured for only 50 of the 62 species in the data set, and *GP*, the gestation period, was measured for only 58 of the species. If we are interested in the dependence of hours of sleep on the other predictors, then we have at least three possible responses, *PS*, *SWS*, and *TS*, all observed on only a subset of the species. To use case deletion and then standard methods to analyze the conditional distributions of interest, we need to assume that the chance of a value being missing does not depend on the value. For example, the four missing values of *GP* are missing because no one had (as of 1976) published this value for these species. Using the imputation or the maximum likelihood methods are alternatives for these data, but they require making assumptions like normality, which might be palatable for many of the variables if transformed to logarithmic scale. Some of the variables, like *P* and *SE* are categorical, so other assumptions beyond multivariate normality might be needed.

TABLE 4.2 The Sleep Data[a]

Variable	Type	Number Observed	Percent Missing	Description
BodyWt	Variate	62	0	Body weight in kg
BrainWt	Variate	62	0	Brain weight in g
D	Factor	62	0	Danger index, 1 = least danger, . . . , 5 = most
GP	Variate	58	6	Gestation time, days
Life	Variate	58	6	Maximum life span, years
P	Factor	62	0	Predation index, 1 = lowest , . . . , 5 = highest
SE	Factor	62	0	Sleep exposure index, 1 = more exposed, . . . , 5 = most protected
PS	Response	50	19	Paradoxical dreaming sleep, hrs/day
SWS	Response	48	23	Slow wave nondreaming sleep, hrs/day
TS	Response	58	6	Total sleep, hrs/day
Species	Labels	62	0	Species of mammal

[a] 10 variables, 62 observations, 8 patterns of missing values; 5 variables (50%) have at least one missing value; 20 observations (32%) have at least one missing value.

4.6 COMPUTATIONALLY INTENSIVE METHODS

Suppose we have a sample $y_1, \ldots, y_n$ from a particular distribution G, for example a standard normal distribution. What is a confidence interval for the population median?

We can obtain an approximate answer to this question by computer simulation, set up as follows:

1. Obtain a simulated random sample $y_1^*, \ldots, y_n^*$ from the known distribution G. Most statistical computing languages include functions for simulating random deviates (see Thisted, 1988 for computational methods).
2. Compute and save the median of the sample in step 1.
3. Repeat steps 1 and 2 a large number of times, say B times. The larger the value of B, the more precise the ultimate answer.
4. If we take $B = 999$, a simple *percentile-based* 95% confidence interval for the median is the interval between the 25th smallest value and the 975th largest value, which are the sample 2.5 and 97.5 percentiles, respectively.

In most interesting problems, we will not actually know G and so this simulation is not available. Efron (1979) pointed out that the observed data can be used to estimate G, and then we can sample from the estimate $\hat{G}$. The algorithm becomes:

1. Obtain a random sample $y_1^*, \ldots, y_n^*$ from $\hat{G}$ by sampling *with replacement* from the observed values $y_1, \ldots, y_n$. In particular, the i-th element of the sample y_i^* is equally likely to be any of the original $y_1, \ldots, y_n$. Some of the y_i will appear several times in the random sample, while others will not appear at all.
2. Continue with steps 2–4 of the first algorithm. A test at the 5% level concerning the population median can be rejected if the hypothesized value of the median does not fall in the confidence interval computed at step 4.

Efron called this method the *bootstrap*, and we call B the number of bootstrap samples. Excellent references for the bootstrap are the books by Efron and Tibshirani (1993), and Davison and Hinkley (1997).

4.6.1 Regression Inference without Normality

Bootstrap methods can be applied in more complex problems like regression. Inferences and accurate standard errors for parameters and mean functions require either normality of regression errors or large sample sizes. In small samples without normality, standard inference methods can be misleading, and in these cases a bootstrap can be used for inference.

Transactions Data

The data in this example consists of a sample of branches of a large Australian bank (Cunningham and Heathcote, 1989). Each branch makes transactions of two types, and for each of the branches we have recorded the number of transactions T_1 and T_2, as well as *Time*, the total number of minutes of labor used by the branch in type 1 and type 2 transactions. If β_j is the average number of minutes for a transaction of type j, $j = 1, 2$, then the total number of minutes in a branch for transaction type j is $\beta_j T j$, and the total number of minutes is expected to be

$$\mathrm{E}(Time|T_1, T_2) = \beta_0 + \beta_1 T_1 + \beta_2 T_2 \tag{4.16}$$

possibly with $\beta_0 = 0$ because zero transactions should imply zero time spent. The data are displayed in Figure 4.4, and are given in the data file `transact.txt`. The key features of the scatterplot matrix are: (1) the marginal response plots in the last row appear to have reasonably linear mean functions; (2) there appear to be a number of branches with no T_1 transactions but many T_2 transactions; and (3) in the plot of *Time* versus T_2, variability appears to increase from left to right.

The errors in this problem probably have a skewed distribution. Occasional transactions take a very long time, but since transaction time is bounded below by

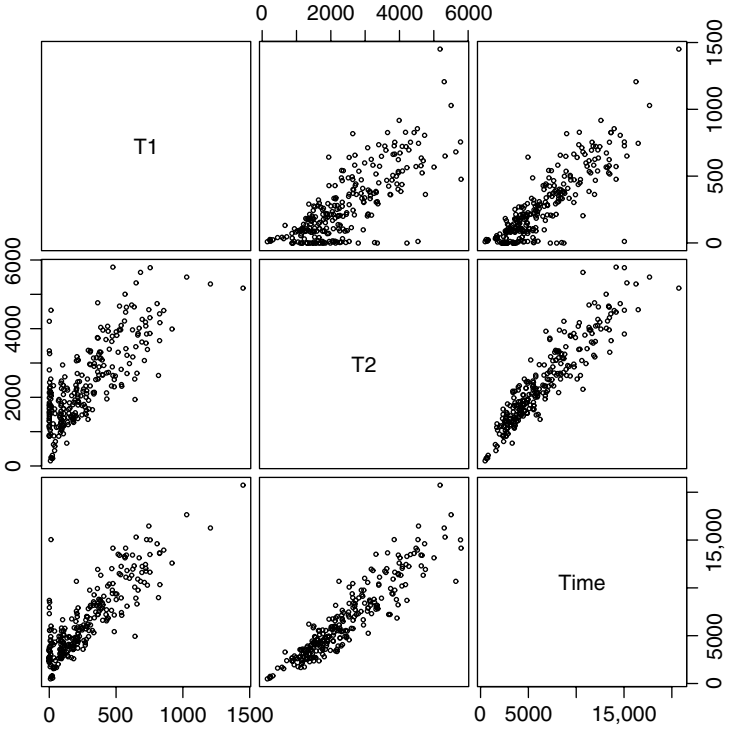

FIG. 4.4 Scatterplot matrix for the transactions data.

TABLE 4.3 Summary for $B = 999$ Case Bootstraps for the Transactions Data, Giving 95% Confidence Intervals, Lower to Upper, Based on Standard Normal Theory and on the Percentile Bootstrap

	Normal Theory			Bootstrap		
	Estimate	Lower	Upper	Estimate	Lower	Upper
Intercept	144.37	−191.47	480.21	136.09	−254.73	523.36
T_1	5.46	4.61	6.32	5.48	4.08	6.77
T_2	2.03	1.85	2.22	2.04	1.74	2.36

zero, there can not be any really extreme "quick" transactions. Inferences based on normal theory are therefore questionable.

Following the suggestion of Pardoe and Weisberg (2001) for this example, a bootstrap is computed as follows:

1. Number the cases in the data set from 1 to n. Take a random sample *with replacement* of size n from these case numbers. Thus, the i-th case number in the sample is equally likely to be any of the n cases in the original data.
2. Create a data set from the original data, but repeating each row in the data set the number of times that row was selected in the random sample in step 1. Some cases will appear several times and others will not appear at all. Compute the regression using this data set, and save the values of the coefficient estimates.
3. Repeat steps 1 and 2 a large number of times, say, B times.
4. Estimate a 95% confidence interval for each of the estimates by the 2.5 and 97.5 percentiles of the sample of B bootstrap samples.

Table 4.3 summarizes the percentile bootstrap for the transactions data. The column marked Estimate gives the OLS estimate under "Normal theory" and the average of the B bootstrap simulations under "Bootstrap." The difference between these two is called the *bootstrap bias*, which is quite small for all three terms relative to the size of the confidence intervals. The 95% bootstrap intervals are consistently wider than the corresponding normal intervals, indicating that the normal-theory confidence intervals are probably overly optimistic. The bootstrap intervals given in Table 4.3 are random, since if the bootstrap is repeated, the answers will be a little different. The variability in the end-points of the interval can be decreased by increasing the number B of bootstrap samples.

4.6.2 Nonlinear Functions of Parameters

One of the important uses of the bootstrap is to get estimates of error variability in problems where standard theory is either missing, or, equally often, unknown to the analyst. Suppose, for example, we wanted to get a confidence interval for the ratio β_1/β_2 in the transactions data. This is the ratio of the time for a type 1

transaction to the time for a type 2 transaction. The point estimate for this ratio is just $\hat{\beta}_1/\hat{\beta}_2$, but we will not learn how to get a normal-theory confidence interval for a nonlinear function of parameters like this until Section 6.1.2. Using the bootstrap, this computation is easy: just compute the ratio in each of the bootstrap samples and then use the percentiles of the bootstrap distribution to get the confidence interval. For these data, the point estimate is 2.68 with 95% bootstrap confidence interval from 1.76 to 3.86, so with 95% confidence, type 1 transactions take on average from about 1.76 to 3.86 times as long as do type 2 transactions.

4.6.3 Predictors Measured with Error

Predictors and the response are often measured with error. While we might have a theory that tells us the mean function for the response, given the true values of the predictors, we must fit with the response, given the imperfectly measured values of the predictors. We can sometimes use simulation to understand how the measurement error affects our answers.

Here is the basic setup. We have a true response Y^* and a set of terms X^* and a true mean function

$$\mathrm{E}(Y^*|X^* = \mathbf{x}^*) = \boldsymbol{\beta}'\mathbf{x}^*$$

In place of Y^* and X^* we observe $Y = Y^* + \delta$ and $X = X^* + \eta$, where δ and η are measurement errors. If we fit the mean function

$$\mathrm{E}(Y|X = \mathbf{x}) = \boldsymbol{\gamma}'\mathbf{x}$$

what can we say about the relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$? While there is a substantial theoretical literature on this problem (for example, Fuller, 1987), we shall attempt to get an answer to this question using simulation. To do so, we need to know something about δ and η.

Catchability of Northern Pike

One of the questions of interest to fisheries managers is the difficulty of catching a fish. A useful concept is the idea of *catchability*. Suppose that Y^* is the catch for an angler for a fixed amount of effort, and X^* is the abundance of fish available in the population that the angler is fishing. Suppose further that

$$\mathrm{E}(Y^*|X^* = x^*) = \beta_1 x^* \tag{4.17}$$

If this mean function were to hold, then we could define β_1 to be the catchability of this particular fish species.

The data we use comes from a study of Northern Pike, a popular game fish in inland lakes in the United States. Data were collected on 16 lakes by Rob Pierce of the Minnesota Department of Natural Resources. On each lake we have a measurement called *CPUE* or catch per unit effort, which is the catch for a

specific amount of fishing effort. Abundance on the lake is measured using the fish *Density* that is defined to be the number of fish in the lake divided by the surface area of the lake. While surface area can be determined with reasonable accuracy, the number of fish in the lake is estimated using a capture–recapture experiment (Seber, 2002). Since both *CPUE* and *Density* are experimentally estimated, they both have standard errors attached to them. In terms of (4.17), we have observed $CPUE = Y^* + \delta$ and $Density = x^* + \eta$. In addition, we can obtain estimates of the standard deviations of the δs and ηs from the properties of the methods used to find *CPUE* and *Density*. The data file `npdata.txt` includes both the *CPUE* and *Density* and their standard errors *SECPUE* and *SEdens*.

Figure 4.5 is the plot of the estimated *CPUE* and *Density*. Ignoring the lines on the graph, a key characteristic of this graph is the large variability in the points. A straight line mean function seems plausible for these data, but many other curves are equally plausible. We continue under the assumption that a straight-line mean function is sensible.

The two lines on Figure 4.5 are the OLS simple regression fits through the origin (solid line) and not through the origin (dashed line). The F-test comparing them has a p-value of about 0.13, so we are encouraged to use the simpler through-the-origin model that will allow us to interpret the slope as the catchability. The estimate is $\hat{\beta}_1 = 0.34$ with standard error 0.035, so a 95% confidence interval for β_1 ignoring measurement errors is $(0.250, 0.0399)$.

To assess the effect of measurement error on the estimate and on the confidence interval, we first make some assumptions. First, we suppose that the estimated standard errors of the measurements are the actual standard errors of the measurements. Second, we assume that the measurement errors are independently and normally

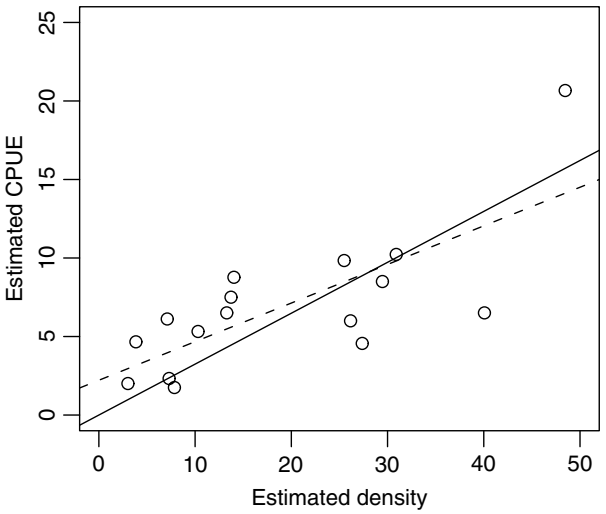

FIG. 4.5 Scatterplot of estimated *CPUE* versus *Density* for the northern pike data. Solid line is the OLS mean function through the origin, and the dashed line is the OLS line allowing an intercept.

TABLE 4.4 Simulation Summary for the Northern Pike Data

	Point Estimate	95% ConfidenceInterval
Normal theory	0.324	(0.250, 0.399)
Simulation	0.309	(0.230, 0.387)

distributed. Neither of these assumptions are checkable from these data, but for the purposes of a simulation these seem like reasonable assumptions.

The simulation proceeds as follows:

1. Generate a pseudo-response vector given by $\tilde{Y} = CPUE + \tilde{\delta}$, where the i-th element of $\tilde{\delta}$ is a normal random number with mean zero and variance given by the square of the estimated standard error for the i-th $CPUE$ value. In this problem, each observation has its own estimated error variance, but in other problems there may be a common estimate for all elements of the response.
2. Repeat step 1, but for the predictor to get $\tilde{x} = Density + \tilde{\eta}$.
3. Fit the simple regression model of $\tilde{Y}$ on $\tilde{x}$ and save the estimated slope.
4. Repeat steps 1–3 B times. The average of the B values of the slope estimate is an estimate of the slope in the problem with no measurement error. A confidence interval for the slope is found using the percentile method discussed with the bootstrap.

The samples generated in steps 1–2 are not quite from the right distribution, as they are centered at the observed values of $CPUE$ and $Density$ rather than the unobserved values of Y^* and x^*, but the observed values estimate the unobserved true values, so this substitution adds variability to the results, but does not affect the validity of the methodology.

The results for $B = 999$ simulations are summarized in Table 4.4. The results of the normal theory and the simulation that allows for measurement error are remarkably similar. In this problem, we judge the measurement error to be unimportant.

PROBLEMS

4.1. Fit the regression of *Soma* on *AVE*, *LIN* and *QUAD* as defined in Section 4.1 for the girls in the Berkeley Guidance Study data, and compare to the results in Section 4.1.

4.2.

4.2.1. Starting with (4.10), we can write

$$y_i = \mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x) + \varepsilon_i$$

Ignoring the error term ε_i, solve this equation for x_i as a function of y_i and the parameters.

4.2.2. Find the conditional distribution of $x_i|y_i$. Under what conditions is the equation you obtained in Problem 4.2.1, which is computed by inverting the regression of y on x, the same as the regression of x on y?

4.3. For the transactions data described in Section 4.6.1, define $A = (T_1 + T_2)/2$ to be the average transaction time, and $D = T_1 - T_2$, and fit the following four mean functions

$$M1 : \mathrm{E}(Y|T_1, T_2) = \beta_{01} + \beta_{11}T_1 + \beta_{21}T_2$$
$$M2 : \mathrm{E}(Y|T_1, T_2) = \beta_{02} + \beta_{32}A + \beta_{42}D$$
$$M3 : \mathrm{E}(Y|T_1, T_2) = \beta_{03} + \beta_{23}T_2 + \beta_{43}D$$
$$M4 : \mathrm{E}(Y|T_1, T_2) = \beta_{04} + \beta_{14}T_1 + \beta_{24}T_2 + \beta_{34}A + \beta_{44}D$$

4.3.1. In the fit of M4, some of the coefficients estimates are labelled as either "aliased" or as missing. Explain what this means.

4.3.2. What aspects of the fitted regressions are the same? What is different?

4.3.3. Why is the estimate for T_2 different in M1 and M3?

4.4. Interpreting coefficients with logarithms

4.4.1. For the simple regression with mean function $\mathrm{E}(\log(Y)|X = x) = \beta_0 + \beta_1 \log(x)$, provide an interpretation for β_1 as a rate of change in Y for a small change in x.

4.4.2. Show that the results of Section 4.1.7 do not depend on the base of the logarithms.

4.5. Use the bootstrap to estimate confidence intervals of the coefficients in the fuel data.

4.6. Windmill data For the windmill data in the data file `wm1.txt` discussed in Problem 2.13, page 45, use $B = 999$ replications of the bootstrap to estimate a 95% confidence interval for the long-term average wind speed at the candidate site and compare this to the prediction interval in Problem 2.13.5. See the comment at the end of Problem 2.13.4 to justify using a bootstrap confidence interval for the mean as a prediction interval for the long-term mean.

4.7. Suppose we fit a regression with the true mean function

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = 3 + 4x_1 + 2x_2$$

Provide conditions under which the mean function for $\mathrm{E}(Y|X_1 = x_1)$ is linear but has a negative coefficient for x_1.

4.8. In a study of faculty salaries in a small college in the Midwest, a linear regression model was fit, giving the fitted mean function

$$\widehat{\mathrm{E}(Salary|Sex)} = 24697 - 3340Sex \tag{4.18}$$

where *Sex* equals one if the faculty member was female and zero if male. The response *Salary* is measured in dollars (the data are from the 1970s).

4.8.1. Give a sentence that describes the meaning of the two estimated coefficients.

4.8.2. An alternative mean function fit to these data with an additional term, *Years*, the number of years employed at this college, gives the estimated mean function

$$\widehat{\mathrm{E}(Salary|Sex, Years)} = 18065 + 201Sex + 759Years \tag{4.19}$$

The important difference between these two mean functions is that the coefficient for *Sex* has changed signs. Using the results of this chapter, explain how this could happen. (Data consistent with these equations are presented in Problem 6.13).

4.9. Sleep data

4.9.1. For the sleep data described in Section 4.5, describe conditions under which the missing at random assumption is reasonable. In this case, deleting the partially observed species and analyzing the complete data can make sense.

4.9.2. Describe conditions under which the missing at random assumption for the sleep data is not reasonable. In this case, deleting partially observed species can change the inferences by changing the definition of the sampled population.

4.9.3. Suppose that the sleep data were fully observed, meaning that values for all the variables were available for all 62 species. Assuming that there are more than 62 species of mammals, provide a situation where examining the missing at random assumption could still be important.

4.10. The data given in `longley.txt` were first given by Longley (1967) to demonstrate inadequacies of regression computer programs then available. The variables are:

$$\begin{aligned} Def &= \text{GNP price deflator, in percent} \\ GNP &= \text{GNP, in millions of dollars} \\ Unemployed &= \text{Unemployment, in thousands of persons} \\ Armed.Forces &= \text{Size of armed forces, in thousands} \\ Population &= \text{Population 14 years of age and over, in thousands} \end{aligned}$$

Employed = Total derived employment in thousands the response

Year = Year

4.10.1. Draw the scatterplot matrix for these data excluding *Year*, and explain from the plot why this might be a good example to illustrate numerical problems of regression programs. (*Hint*: Numerical problems arise through rounding errors, and these are most likely to occur when terms in the regression model are very highly correlated.)

4.10.2. Fit the regression of *Employed* on the others excluding *Year*.

4.10.3. Suppose that the values given in this example were only accurate to three significant figures (two figures for *Def*). The effects of measurement errors can be assessed using a simulation study in which we add uniform random values to the observed values, and recompute estimates for each simulation. For example, *Unemp* for 1947 is given as 2356, which corresponds to 2,356,000. If we assume only three significant figures, we only believe the first three digits. In the simulation we would replace 2356 by $2356 + u$, where u is a uniform random number between -5 and $+5$. Repeat the simulation 1000 times, and on each simulation compute the coefficient estimates. Compare the standard deviation of the coefficient estimates from the simulation to the coefficient standard errors from the regression on the unperturbed data. If the standard deviations in the simulation are as large or larger than the standard errors, we would have evidence that rounding would have important impact on results.