CHAPTER 3

Linear regression: the basics

Linear regression is a method that summarizes how the average values of a numerical *outcome* variable vary over subpopulations defined by linear functions of *predictors*. Introductory statistics and regression texts often focus on how regression can be used to represent relationships between variables, rather than as a comparison of average outcomes. By focusing on regression as a comparison of averages, we are being explicit about its limitations for defining these relationships causally, an issue to which we return in Chapter 9. Regression can be used to predict an outcome given a linear function of these predictors, and regression coefficients can be thought of as comparisons across predicted values or as comparisons among averages in the data.

3.1 One predictor

We begin by understanding the coefficients without worrying about issues of estimation and uncertainty. We shall fit a series of regressions predicting cognitive test scores of three- and four-year-old children given characteristics of their mothers, using data from a survey of adult American women and their children (a subsample from the National Longitudinal Survey of Youth).

For a binary predictor, the regression coefficient is the difference between the averages of the two groups

We start by modeling the children's test scores given an indicator for whether the mother graduated from high school (coded as 1) or not (coded as 0). The fitted model is

$$\text{kid.score} = 78 + 12 \cdot \text{mom.hs} + \text{error}, \tag{3.1}$$

but for now we focus on the deterministic part,

$$\widehat{\text{kid.score}} = 78 + 12 \cdot \text{mom.hs}, \tag{3.2}$$

where $\widehat{\text{kid.score}}$ denotes either predicted or expected test score given the `mom.hs` predictor.

This model summarizes the difference in average test scores between the children of mothers who completed high school and those with mothers who did not. Figure 3.1 displays how the regression line runs through the mean of each subpopulation.

The intercept, 78, is the average (or predicted) score for children whose mothers did not complete high school. To see this algebraically, consider that to obtain predicted scores for these children we would just plug 0 into this equation. To obtain average test scores for children (or the predicted score for a single child) whose mothers were high school graduates, we would just plug 1 into this equation to obtain $78 + 12 \cdot 1 = 91$.

The difference between these two subpopulation means is equal to the coefficient

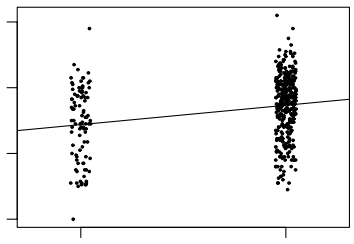

Figure 3.1 *Child's test score plotted versus an indicator for whether mother completed high school. Superimposed is the regression line, which runs through the average of each subpopulation defined by maternal education level. The indicator variable for high school completion has been* jittered*; that is, a random number has been added to each value so that the points do not lie on top of each other.*

on `mom.hs`. This coefficient tells us that children of mothers who have completed high school score 12 points higher on average than children of mothers who have not completed high school.

Regression with a continuous predictor

If we regress instead on a continuous predictor, mother's score on an IQ test, the fitted model is

$$\text{kid.score} = 26 + 0.6 \cdot \text{mom.iq} + \text{error}, \tag{3.3}$$

and is shown in Figure 3.2. We can think of the points on the line either as predicted test scores for children at each of several maternal IQ levels, or average test scores for subpopulations defined by these scores.

If we compare average child test scores for subpopulations that differ in maternal IQ by 1 point, we expect to see that the group with higher maternal IQ achieves 0.6 points more on average. Perhaps a more interesting comparison would be between groups of children whose mothers' IQ differed by 10 points—these children would be expected to have scores that differed by 6 points on average.

To understand the constant term in the regression we must consider a case with zero values of all the other predictors. In this example, the intercept of 26 reflects the predicted test scores for children whose mothers have IQ scores of zero. This is not the most helpful quantity—we don't observe any women with zero IQ. We will discuss a simple transformation in the next section that gives the intercept a more useful interpretation.

3.2 Multiple predictors

Regression coefficients are more complicated to interpret with multiple predictors because the interpretation for any given coefficient is, in part, contingent on the other variables in the model. Typical advice is to interpret each coefficient "with all the other predictors held constant." We illustrate with an example, followed by an elaboration in which the simple interpretation of regression coefficients does not work.

For instance, consider a linear regression predicting child test scores from mater-

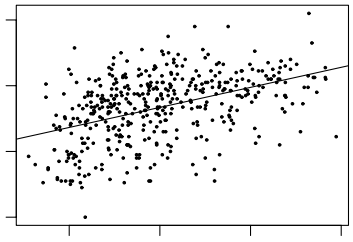

Figure 3.2 *Child's test score plotted versus maternal IQ with regression line superimposed. Each point on the line can be conceived of either as a predicted child test score for children with mothers who have the corresponding IQ, or as the average score for a subpopulation of children with mothers with that IQ.*

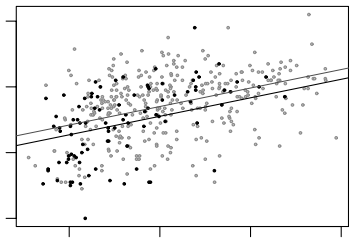

Figure 3.3 *Child's test score plotted versus maternal IQ. Light dots represent children whose mothers graduated from high school and dark dots represent children whose mothers did not graduate from high school. Superimposed are the regression lines from the regression of child's test score on maternal IQ and maternal high school indicator (the darker line for children whose mothers did not complete high school, the lighter line for children whose mothers did complete high school).*

nal education and maternal IQ. The fitted model is

$$\text{kid.score} = 26 + 6 \cdot \text{mom.hs} + 0.6 \cdot \text{mom.iq} + \text{error}, \tag{3.4}$$

and is displayed in Figure 3.3. This model forces the slope of the regression of child's test score on mother's IQ score to be the same for each maternal education subgroup. The next section considers models in which the slopes of the two lines differ. First, however, we interpret the coefficients in model (3.4):

1. *The intercept.* If a child had a mother with an IQ of 0 and who did not complete high school (thus, `mom.hs` $= 0$), then we would predict this child's test score to be 26. This is not a useful prediction, since no mothers have IQs of 0.
2. *The coefficient of maternal high school completion.* Comparing children whose mothers have the same IQ, but who differed in whether they completed high school, the model predicts an expected difference of 6 in their test scores.
3. *The coefficient of maternal IQ.* Comparing children with the same value of `mom.hs`, but whose mothers differ by 1 point in IQ, we would expect to see

a difference of 0.6 points in the child's test score (equivalently, a difference of 10 in mothers' IQs corresponds to a difference of 6 points for their children).

It's not always possible to change one predictor while holding all others constant

We interpret the regression slopes as comparisons of individuals that differ in one predictor while being *at the same levels of the other predictors*. In some settings, one can also imagine manipulating the predictors to change some or hold others constant—but such an interpretation is not necessary. This becomes clearer when we consider situations in which it is logically impossible to change the value of one predictor while keeping the value of another constant. For example, if a model includes both IQ and IQ^2 as predictors, it does not make sense to consider changes in IQ with IQ^2 held constant. Or, as we discuss in the next section, if a model includes `mom.hs`, `mom.IQ`, and their interaction, `mom.hs * mom.IQ`, it is not meaningful to consider any of these three with the other two held constant.

Counterfactual and predictive interpretations

In the more general context of multiple linear regression, it pays to be more explicit about how we interpret coefficients in general. We distinguish between two interpretations of regression coefficients.

- The *predictive interpretation* considers how the outcome variable differs, on average, when comparing two groups of units that differ by 1 in the relevant predictor while being identical in all the other predictors. Under the linear model, the coefficient is the expected difference in y between these two units. This is the sort of interpretation we have described thus far.
- The *counterfactual interpretation* is expressed in terms of changes within individuals, rather than comparisons between individuals. Here, the coefficient is the expected change in y caused by adding 1 to the relevant predictor, while leaving all the other predictors in the model unchanged. For example, "changing maternal IQ from 100 to 101 would lead to an expected increase of 0.6 in child's test score." This sort of interpretation arises in causal inference.

Most introductory statistics and regression texts warn against the latter interpretation but then allow for similar interpretations such as "a change of 10 in maternal IQ is *associated* with a change of 6 points in child's score." Thus, the counterfactual interpretation is probably more familiar to you—and is sometimes easier to understand. However, as we discuss in detail in Chapter 9, the counterfactual interpretation can be inappropriate without making some strong assumptions.

3.3 Interactions

In model (3.4), the slope of the regression of child's test score on mother's IQ was forced to be equal across subgroups defined by mother's high school completion, but inspection of the data in Figure 3.3 suggests that the slopes differ substantially. A remedy for this is to include an *interaction* between `mom.hs` and `mom.iq`—that is, a new predictor which is defined as the product of these two variables. This allows the slope to vary across subgroups. The fitted model is

$$\text{kid.score} = -11 + 51 \cdot \text{mom.hs} + 1.1 \cdot \text{mom.iq} - 0.5 \cdot \text{mom.hs} \cdot \text{mom.iq} + \text{error}$$

and is displayed in Figure 3.4a, where we see the separate regression lines for each subgroup defined by maternal education.

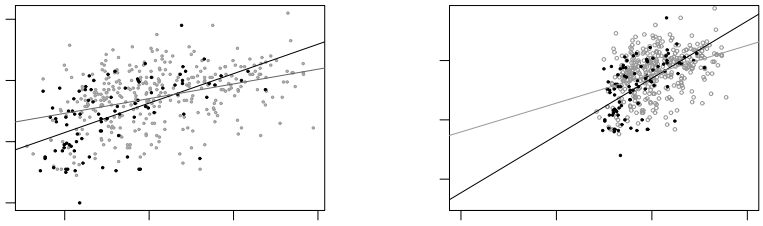

Figure 3.4 *(a) Regression lines of child's test score on mother's IQ with different symbols for children of mothers who completed high school (light circles) and those whose mothers did not complete high school (dark dots). The interaction allows for a different slope in each group, with light and dark lines corresponding to the light and dark points. (b) Same plot but with horizontal axis extended to zero to reveal the intercepts of the lines.*

Figure 3.4b shows the regression line and uncertainty on a scale with the x-axis extended to zero to display the intercepts—the points on the y-axis where the lines cross zero. This highlights the fact that not only is the value meaningless in terms of its interpretation, it is also so far out of the range of our data as to be highly unreliable as a subpopulation estimate.

Care must be taken in interpreting the coefficients in this model. We derive meaning from the coefficients (or, sometimes, functions of the coefficients) by examining average or predicted test scores within and across specific subgroups. Some coefficients are interpretable only for certain subgroups.

1. *The intercept* represents the predicted test scores for children whose mothers did not complete high school and had IQs of 0—not a meaningful scenario. (As we discuss in Sections 4.1–4.2, intercepts can be more interpretable if input variables are centered before including them as regression predictors.)

2. *The coefficient of* `mom.hs` can be conceived as the difference between the predicted test scores for children whose mothers did not complete high school and had IQs of 0, and children whose mothers did complete high school and had IQs of 0. You can see this by just plugging in the appropriate numbers and comparing the equations. Since it is implausible to imagine mothers with IQs of zero, this coefficient is not easily interpretable.

3. *The coefficient of* `mom.iq` can be thought of as the comparison of mean test scores across children whose mothers did not complete high school, but whose mothers differ by 1 point in IQ. This is the slope of the dark line in Figure 3.4.

4. *The coefficient on the interaction term* represents the *difference* in the slope for `mom.iq`, comparing children with mothers who did and did not complete high school: that is, the difference between the slopes of the light and dark lines in Figure 3.4.

An equivalent way to understand the model is to look at the separate regression lines for children of mothers who completed high school and those whose mothers

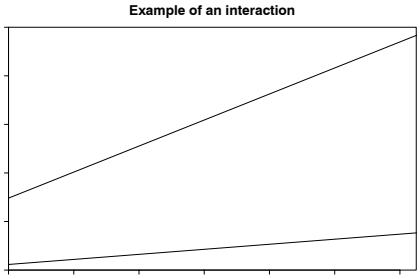

Figure 3.5 *Illustration of interactions between smoking and home radon level on the lifetime probability of lung cancer in men. The effects of radon are much more severe for smokers. The lines are estimated based on case-control studies; see Lin et al. (1999) for references.*

did not:

$$\begin{aligned}
\text{no hs: kid.score} &= -11 + 51 \cdot 0 + 1.1 \cdot \text{mom.iq} - 0.5 \cdot 0 \cdot \text{mom.iq} \\
&= -11 + 1.1 \cdot \text{mom.iq} \\
\text{hs: kid.score} &= -11 + 51 \cdot 1 + 1.1 \cdot \text{mom.iq} - 0.5 \cdot 1 \cdot \text{mom.iq} \\
&= 40 + 0.6 \cdot \text{mom.iq}.
\end{aligned}$$

The estimated slopes of 1.1 for children whose mothers did not complete high school and 0.6 for children of mothers who did are directly interpretable. The intercepts still suffer from the problem of only being interpretable at mothers' IQs of 0.

When should we look for interactions?

Interactions can be important. In practice, inputs that have large main effects also tend to have large interactions with other inputs (however, small main effects do not preclude the possibility of large interactions). For example, smoking has a huge effect on cancer. In epidemiologial studies of other carcinogens, it is crucial to adjust for smoking both as a main effect and as an interaction. Figure 3.5 illustrates with the example of home radon exposure: high levels of radon are associated with greater likelihood of cancer—but this difference is much greater for smokers than for nonsmokers.

Including interactions is a way to allow a model to be fit differently to different subsets of data. These two approaches are related, as we discuss later in the context of multilevel models.

Interpreting regression coefficients in the presence of interactions

Models with interactions can often be more easily interpreted if we first pre-process the data by centering each input variable about its mean or some other convenient reference point. We discuss this in Section 4.2 in the context of linear transformations.

3.4 Statistical inference

When illustrating specific examples, it helps to use descriptive variable names. In order to discuss more general theory and data manipulations, however, we shall adopt generic mathematical notation. This section introduces this notation and discusses the stochastic aspect of the model as well.

Units, outcome, predictors, and inputs

We refer to the individual data points as *units*—thus, the answer to the question, "What is the unit of analysis?" will be something like "persons" or "schools" or "congressional elections," *not* something like "pounds" or "miles." Multilevel models feature more than one set of units (for example, both persons and schools), as we discuss later on.

We refer to the X-variables in the regression as *predictors* or "predictor variables," and y as the *outcome* or "outcome variable." We do *not* use the terms "dependent" and "independent" variables, because we reserve those terms for their use in describing properties of probability distributions.

Finally, we use the term *inputs* for the information on the units that goes into the X-variables. Inputs are not the same as predictors. For example, consider the model that includes the interaction of maternal education and maternal IQ:

$$\text{kid.score} = 58 + 16 \cdot \text{mom.hs} + 0.5 \cdot \text{mom.iq} - 0.2 \cdot \text{mom.hs} \cdot \text{mom.iq} + \text{error}.$$

This regression has four *predictors*—maternal high school, maternal IQ, maternal high school $\times$ IQ, and the constant term—but only two *inputs*, maternal education and IQ.

Regression in vector-matrix notation

We follow the usual notation and label the outcome for the i^{th} individual as y_i and the deterministic prediction as $X_i\beta = \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$, indexing the persons in the data as $i = 1, \ldots, n = 1378$. In our most recent example, y_i is the i^{th} child's test score, and there are $k = 4$ predictors in the vector X_i (the i^{th} row of the matrix X): X_{i1}, a *constant term* that is defined to equal 1 for all persons; X_{i2}, the mother's high school completion status (coded as 0 or 1); X_{i3}, the mother's test score; and X_{i4}, the interaction between mother's test score and high school completion status. The vector β of coefficients has length $k = 4$ as well. The errors from the model are labeled as ϵ_i and assumed to follow a normal distribution with mean 0 and standard deviation σ, which we write as $\text{N}(0, \sigma^2)$. The parameter σ represents the variability with which the outcomes deviate from their predictions based on the model. We use the notation $\tilde{y}$ for unobserved data to be predicted from the model, given predictors $\tilde{X}$; see Figure 3.6.

Two ways of writing the model

The classical linear regression model can then be written mathematically as

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ &= \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \ldots, n, \end{aligned}$$

where the errors ϵ_i have independent normal distributions with mean 0 and standard deviation σ.

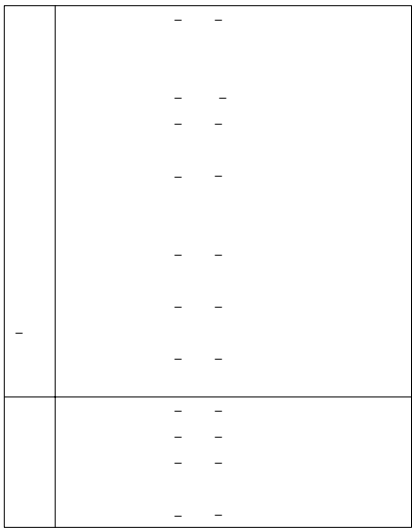

Figure 3.6 *Notation for regression modeling. The model is fit to the observed outcomes y given predictors X. As described in the text, the model can then be applied to predict unobserved outcomes $\tilde{y}$ (indicated by small question marks), given predictors on new data $\tilde{X}$.*

An equivalent representation is

$$y_i \sim \mathrm{N}(X_i\beta, \sigma^2), \text{ for } i = 1, \ldots, n,$$

where X is an n by k matrix with i^{th} row X_i, or, using multivariate notation,

$$y \sim \mathrm{N}(X\beta, \sigma^2 I),$$

where y is a vector of length n, X is a $n \times k$ matrix of predictors, β is a column vector of length k, and I is the $n \times n$ identity matrix. Fitting the model (in any of its forms) using least squares yields estimates $\hat{\beta}$ and $\hat{\sigma}$.

Fitting and summarizing regressions in R

We can fit regressions using the `lm()` function in R. We illustrate with the model including mother's high school completion and IQ as predictors, for simplicity not adding the interaction for now. We shall label this model as `fit.3` as it is the third model fit in this chapter:

R code

```
fit.3 <- lm (kid.score ~ mom.hs + mom.iq)
display (fit.3)
```

(The spaces in the R code are not necessary, but we include them to make the code more readable.) The result is

R output

```
lm(formula = kid.score ~ mom.hs + mom.iq)
             coef.est coef.se
(Intercept)     25.7     5.9
mom.hs           5.9     2.2
```

```
mom.iq           0.6     0.1
  n = 434, k = 3
  residual sd = 18.1, R-Squared = 0.21
```

The `display()` function was written by us (see Section C.2 for details) to give a clean printout focusing on the most pertinent pieces of information: the coefficients and their standard errors, the sample size, number of predictors, residual standard deviation, and R^2.

In contrast, the default R option,

```
print (fit.3)
```
R code

displays too little information, giving only the coefficient estimates with no standard errors and no information on the residual standard deviations:

```
Call:
lm(formula = kid.score ~ mom.hs + mom.iq)

Coefficients:
(Intercept)       mom.hs       mom.iq
   25.73154      5.95012      0.56391
```
R code

Another option in R is the `summary()` function:

```
summary (fit.3)
```
R code

but this produces a mass of barely digestible information displayed to many decimal places:

```
Call:
lm(formula = formula("kid.score ~ mom.hs + mom.iq"))

Residuals:
    Min      1Q  Median      3Q     Max
-52.873 -12.663   2.404  11.356  49.545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
mom.hs       5.95012    2.21181   2.690  0.00742 **
mom.iq       0.56391    0.06057   9.309  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-Squared: 0.2141,     Adjusted R-squared: 0.2105
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```
R output

We prefer our `display()` function, which consisely presents the most relevant information from the model fit.

Least squares estimate of the vector of regression coefficients, β

For the model $y = X\beta + \epsilon$, the least squares estimate is the $\hat{\beta}$ that minimizes the sum of squared errors, $\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2$, for the given data X, y. Intuitively, the least squares criterion seems useful because, if we are trying to predict an outcome using other variables, we want to do so in such a way as to minimize the error of our prediction.

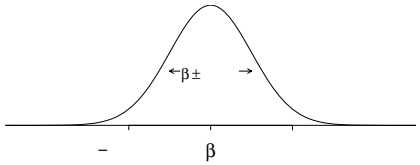

Figure 3.7 *Distribution representing uncertainty in an estimated regression coefficient. The range of this distribution corresponds to the possible values of β that are consistent with the data. When using this as an uncertainty distribution, we assign an approximate 68% chance that β will lie within 1 standard error of the point estimate, $\hat{\beta}$, and an approximate 95% chance that β will lie within 2 standard errors. Assuming the regression model is correct, it should happen only about 5% of the time that the estimate, $\hat{\beta}$, falls more than 2 standard errors away from the true β.*

The least squares estimate is also the maximum likelihood estimate if the errors ϵ_i are independent with equal variance and normally distributed (see Section 18.1). In any case, the least squares estimate can be expressed in matrix notation as $\hat{\beta} = (X^t X)^{-1} X^t y$. In practice, the computation is performed using various efficient matrix decompositions without ever fully computing $X^t X$ or inverting it. For our purposes, it is merely useful to realize that $\hat{\beta}$ is a linear function of the outcomes y.

Standard errors: uncertainty in the coefficient estimates

The estimates $\hat{\beta}$ come with standard errors, as displayed in the regression output. The standard errors represent estimation uncertainty. We can roughly say that coefficient estimates within 2 standard errors of $\hat{\beta}$ are consistent with the data. Figure 3.7 shows the normal distribution that approximately represents the range of possible values of β. For example, in the model on page 38, the coefficient of `mom.hs` has an estimate $\hat{\beta}$ of 5.9 and a standard error of 2.2; thus the data are roughly consistent with values of β in the range $[5.9 \pm 2 \cdot 2.2] = [1.5, 10.3]$. More precisely, one can account for the uncertainty in the standard errors themselves by using the t distribution with degrees of freedom set to the number of data points minus the number of estimated coefficients, but the normal approximation works fine when the degrees of freedom are more than 30 or so.

The uncertainty in the coefficient estimates will also be correlated (except in the special case of studies with balanced designs). All this information is encoded in the estimated covariance matrix $V_\beta \hat{\sigma}^2$, where $V_\beta = (X^t X)^{-1}$. The diagonal elements of $V_\beta \hat{\sigma}^2$ are the estimation variances of the individual components of β, and the off-diagonal elements represent covariances of estimation. Thus, for example, $\sqrt{V_{\beta\,11}}\,\hat{\sigma}$ is the standard error of $\hat{\beta}_1$, $\sqrt{V_{\beta\,22}}\,\hat{\sigma}$ is the standard error of $\hat{\beta}_2$, and $V_{\beta\,12}/\sqrt{V_{\beta\,11} V_{\beta\,22}}$ is the correlation of the estimates $\hat{\beta}_1, \hat{\beta}_2$.

We do not usually look at this covariance matrix; rather, we summarize inferences using the coefficient estimates and standard errors, and we use the covariance matrix for predictive simulations, as described in Section 7.2.

Residuals, r_i

The *residuals*, $r_i = y_i - X_i \hat{\beta}$, are the differences between the data and the fitted values. As a byproduct of the least squares estimation of β, the residuals r_i will be uncorrelated with all the predictors in the model. If the model includes a constant

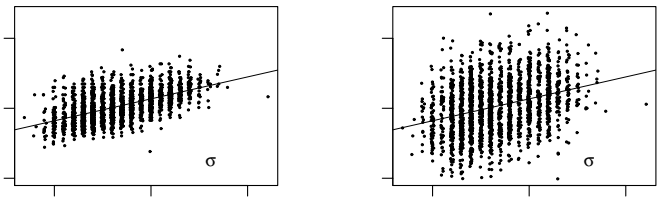

Figure 3.8 *Two hypothetical datasets with the same regression line, $y = a+bx$, but different values of the residual standard deviation, σ. The left plot shows actual data from a survey of adults; the right plot shows data with random noise added to y.*

term, then the residuals must be uncorrelated with a constant, which means they must have mean 0. This is a byproduct of how the model is estimated; it is *not* a regression assumption. We shall discuss later in the chapter how residuals can be used to diagnose problems with the model.

Residual standard deviation $\hat{\sigma}$ and explained variance R^2

The residual standard deviation, $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} r_i^2/(n-k)}$, summarizes the scale of the residuals. For example, in the test scores example, $\hat{\sigma} = 18$, which tells us that the linear model can predict children's test scores to about an accuracy of 18 points. Said another way, we can think of this standard deviation as a measure of the average distance each observation falls from its prediction from the model.

The fit of the model can be summarized by $\hat{\sigma}$ (the smaller the residual variance, the better the fit) and by R^2, the fraction of variance "explained" by the model. The "unexplained" variance is $\hat{\sigma}^2$, and if we label s_y as the standard deviation of the data, then $R^2 = 1 - \hat{\sigma}^2/s_y^2$. In the test scores regression, R^2 is a perhaps disappointing 22%. (However, in a deeper sense, it is presumably a good thing that this regression has a low R^2—that is, that a child's achievement cannot be accurately predicted given only these maternal characteristics.)

The quantity $n-k$, the number of data points minus the number of estimated coefficients, is called the *degrees of freedom* for estimating the residual errors. In classical regression, k must be less than n—otherwise, the data could be fit perfectly, and it would not be possible to estimate the regression errors at all.

Difficulties in interpreting residual standard deviation and explained variance

As we make clear throughout the book, we are generally more interested in the "deterministic" part of the model, $y = X\beta$, than in the variation, ϵ. However, when we do look at the residual standard deviation, $\hat{\sigma}$, we are typically interested in it for its own sake—as a measure of the unexplained variation in the data—or because of its relevance to the precision of inferences about the regression coefficients β. (As discussed already, standard errors for β are proportional to σ.) Figure 3.8 illustrates two regressions with the same deterministic model, $y = a + bx$, but different values of σ.

Interpreting the proportion of explained variance, R^2, can be tricky because its numerator and denominator can be changed in different ways. Figure 3.9 illustrates with an example where the regression model is identical, but R^2 decreases because

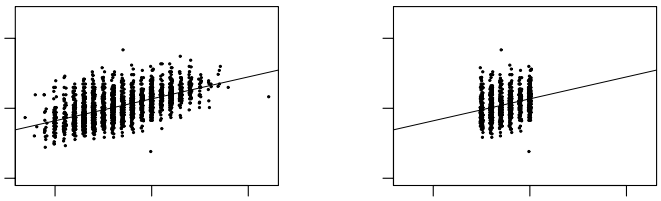

Figure 3.9 *Two hypothetical datasets with the same regression line, $y = a+bx$ and residual standard deviation, σ, but different values of the explained variance, R^2. The left plot shows actual data; the right plot shows data restricted to heights between 65 and 70 inches.*

the model is estimated on a subset of the data. (Going from the left to right plots in Figure 3.9, the residual standard deviation σ is unchanged but the standard deviation of the raw data, s_y, decreases when we restrict to this subset; thus, $R^2 = 1 - \hat{\sigma}^2/s_y^2$ declines.) Even though R^2 is much lower in the right plot, the model fits the data just as well as in the plot on the left.

Statistical significance

Roughly speaking, if a coefficient estimate is more than 2 standard errors away from zero, then it is called *statistically significant*. When an estimate is statistically significant, we are fairly sure that the sign (+ or −) of the estimate is stable, and not just an artifact of small sample size.

People sometimes think that if a coefficient estimate is not significant, then it should be excluded from the model. We disagree. It is fine to have nonsignificant coefficients in a model, as long as they make sense. We discuss this further in Section 4.6.

Uncertainty in the residual standard deviation

Under the model, the estimated residual variance, $\hat{\sigma}^2$, has a sampling distribution centered at the true value, σ^2, and proportional to a χ^2 distribution with $n-k$ degrees of freedom. We make use of this uncertainty in our predictive simulations, as described in Section 7.2.

3.5 Graphical displays of data and fitted model

Displaying a regression line as a function of one input variable

We displayed some aspects of our test scores model using plots of the data in Figures 3.1–3.3.

We can make a plot such as Figure 3.2 as follows:

R code

```
fit.2 <- lm (kid.score ~ mom.iq)
plot (mom.iq, kid.score, xlab="Mother IQ score", ylab="Child test score")
curve (coef(fit.2)[1] + coef(fit.2)[2]*x, add=TRUE)
```

The function `plot()` creates the scatterplot of observations, and `curve` superimposes the regression line using the saved coefficients from the `lm()` call (as extracted using the `coef()` function). The expression within `curve()` can also be written using matrix notation in R:

```
curve (cbind(1,x) %*% coef(fit.2), add=TRUE)
```
R code

Displaying two fitted regression lines

Model with no interaction. For the model with two inputs, we can create a graph with two sets of points and two regression lines, as in Figure 3.3:

```
fit.3 <- lm (kid.score ~ mom.hs + mom.iq)
colors <- ifelse (mom.hs==1, "black", "gray")
plot (mom.iq, kid.score, xlab="Mother IQ score", ylab="Child test score",
  col=colors, pch=20)
curve (cbind (1, 1, x) %*% coef(fit.3), add=TRUE, col="black")
curve (cbind (1, 0, x) %*% coef(fit.3), add=TRUE, col="gray")
```
R code

Setting `pch=20` tells the `plot()` function to display the data using small dots, and the `col` option sets the colors of the points, which we have assigned to black or gray according to the value of `mom.hs`.[1] Finally, the calls to `curve()` superimpose the regression lines for the two groups defined by maternal high school completion.

Model with interaction. We can set up the same sort of plot for the model with interactions, with the only difference being that the two lines have different slopes:

```
fit.4 <- lm (kid.score ~ mom.hs + mom.iq + mom.hs:mom.iq)
colors <- ifelse (mom.hs==1, "black", "gray")
plot (mom.iq, kid.score, xlab="Mother IQ score", ylab="Child test score",
  col=colors, pch=20)
curve (cbind (1, 1, x, 1*x) %*% coef(fit.4), add=TRUE, col="black")
curve (cbind (1, 0, x, 0*x) %*% coef(fit.4), add=TRUE, col="gray")
```
R code

The result is shown in Figure 3.4.

Displaying uncertainty in the fitted regression

As discussed in Section 7.2, we can use the `sim()` function in R to create simulations that represent our uncertainty in the estimated regression coefficients. Here we briefly describe how to use these simulations to display this inferential uncertainty. For simplicity we return to a model with just one predictor:

```
fit.2 <- lm (kid.score ~ mom.iq)
```
R code

yielding

```
            coef.est coef.se
(Intercept)    25.8     5.9
mom.iq          0.6     0.1
  n = 434, k = 2
  residual sd = 18.3, R-Squared = 0.2
```
R output

The following code creates Figure 3.10, which shows the fitted regression line along with several simulations representing uncertainty about the line:

[1] An alternative sequence of commands is
```
plot (mom.iq, kid.score, xlab="Mother IQ score", ylab="Child test score", type="n")
points (mom.iq[mom.hs==1], kid.score[mom.hs==1], pch=20, col="black")
points (mom.iq[mom.hs==0], kid.score[mom.hs==0], pch=20, col="gray")
```
Here, `plot()`, called with the `type="n"` option, sets up the axes but without plotting the points. Then each call to `points()` superimposes the observations for each group (defined by maternal high school completion) separately—each using a different symbol.

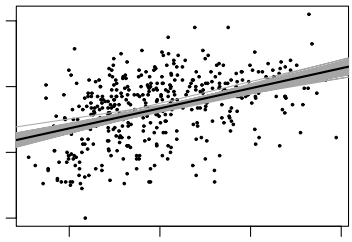

Figure 3.10 *Data and regression of child's test score on maternal IQ, with the solid line showing the fitted regression model and light lines indicating uncertainty in the fitted regression.*

R code

```
fit.2.sim <- sim (fit.2)
plot (mom.iq, kid.score, xlab="Mother IQ score", ylab="Child test score")
for (i in 1:10){
  curve (fit.2.sim$beta[i,1] + fit.2.sim$beta[i,2]*x, add=TRUE,col="gray")
}
curve (coef(fit.2)[1] + coef(fit.2)[2]*x, add=TRUE, col="black")
```

The `for (i in i:10)` loop allows us to display 10 different simulations.[2] Figure 3.10 also illustrates the uncertainty we have about *predictions* from our model. This uncertainty increases with greater departures from the mean of the predictor variable.

Displaying using one plot for each input variable

Now consider the regression model with the indicator for maternal high school completion included:

R code

```
fit.3 <- lm (kid.score ~ mom.hs + mom.iq)
```

We display this model in Figure 3.11 as two plots, one for each of the two input variables with the other held at its average value:

R code

```
beta.hat <- coef (fit.3)
beta.sim <- sim (fit.3)$beta
par (mfrow=c(1,2))

plot (mom.iq, kid.score, xlab="Mother IQ score", ylab="Child test score")
for (i in 1:10){
  curve (cbind (1, mean(mom.hs), x) %*% beta.sim[i,], lwd=.5,
     col="gray", add=TRUE)
}
curve (cbind (1, mean(mom.hs), x) %*% beta.hat, col="black", add=TRUE)

plot (mom.hs, kid.score, xlab="Mother completed high school",
```

[2] Another way to code this loop in R is to use the `apply()` function, for example,

```
Oneline <- function (beta) {curve (beta[1]+beta[2]*x, add=TRUE, col="gray")}
apply (fit.2.sim$beta, 1, Oneline)
```

Using `apply()` in this way is cleaner for experienced R users; the looped form as shown in the text is possibly easier for R novices to understand.

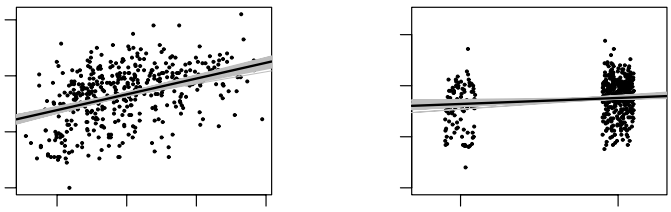

Figure 3.11 *Data and regression of child's test score on maternal IQ and high school completion, shown as a function of each of the two input variables (with light lines indicating uncertainty in the regressions). Values for high school completion have been jittered to make the points more distinct.*

```
  ylab="Child test score")
for (i in 1:10){
  curve (cbind (1, x, mean(mom.iq)) %*% beta.sim[i,], lwd=.5,
     col="gray", add=TRUE)
}
curve (cbind (1, x, mean(mom.iq)) %*% beta.hat, col="black", add=TRUE)
```

3.6 Assumptions and diagnostics

We now turn to the assumptions of the regression model, along with diagnostics that can be used to assess whether some of these assumptions are reasonable. Some of the most important assumptions, however, rely on the researcher's knowledge of the subject area and may not be directly testable from the available data alone.

Assumptions of the regression model

We list the assumptions of the regression model in *decreasing* order of importance.

1. *Validity.* Most importantly, the data you are analyzing should map to the research question you are trying to answer. This sounds obvious but is often overlooked or ignored because it can be inconvenient. Optimally, this means that the outcome measure should accurately reflect the phenomenon of interest, the model should include all relevant predictors, and the model should generalize to the cases to which it will be applied.

   For example, with regard to the outcome variable, a model of earnings will not necessarily tell you about patterns of total assets. A model of test scores will not necessarily tell you about child intelligence or cognitive development.

   Choosing inputs to a regression is often the most challenging step in the analysis. We are generally encouraged to include all "relevant" predictors, but in practice it can be difficult to determine which are necessary and how to interpret coefficients with large standard errors. Chapter 9 discusses the choice of inputs for regressions used in causal inference.

   A sample that is representative of all mothers and children may not be the most appropriate for making inferences about mothers and children who participate in the Temporary Assistance for Needy Families program. However, a carefully

selected subsample may reflect the distribution of this population well. Similarly, results regarding diet and exercise obtained from a study performed on patients at risk for heart disease may not be generally applicable to generally healthy individuals. In this case assumptions would have to be made about how results for the at-risk population might relate to those for the healthy population.

Data used in empirical research rarely meet all (if any) of these criteria precisely. However, keeping these goals in mind can help you be precise about the types of questions you can and cannot answer reliably.

2. *Additivity and linearity.* The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors: $y = \beta_1 x_1 + \beta_2 x_2 + \cdots$.

   If additivity is violated, it might make sense to transform the data (for example, if $y = abc$, then $\log y = \log a + \log b + \log c$) or to add interactions. If linearity is violated, perhaps a predictor should be put in as $1/x$ or $\log(x)$ instead of simply linearly. Or a more complicated relationship could be expressed by including both x and x^2 as predictors.

   For example, it is common to include both `age` and `age`^{2} as regression predictors. In medical and public health examples, this allows a health measure to decline with higher ages, with the rate of decline becoming steeper as age increases. In political examples, including both `age` and `age`^{2} allows the possibility of increasing slopes with age and also U-shaped patterns if, for example, the young and old favor taxes more than the middle-aged.

   In such analyses we usually prefer to include age as a categorical predictor, as discussed in Section 4.5. Another option is to use a nonlinear function such as a spline or other generalized additive model. In any case, the goal is to add predictors so that the linear and additive model is a reasonable approximation.

3. *Independence of errors.* The simple regression model assumes that the errors from the prediction line are independent. We will return to this issue in detail when discussing multilevel models.

4. *Equal variance of errors.* If the variance of the regression errors are unequal, estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance (see Section 18.4). In most cases, however, this issue is minor. Unequal variance does not affect the most important aspect of a regression model, which is the form of the predictor $X\beta$.

5. *Normality of errors.* The regression assumption that is generally *least* important is that the errors are normally distributed. In fact, for the purpose of estimating the regression line (as compared to predicting individual data points), the assumption of normality is barely important at all. Thus, in contrast to many regression textbooks, we do *not* recommend diagnostics of the normality of regression residuals.

   If the distribution of residuals is of interest, perhaps because of predictive goals, this should be distinguished from the distribution of the data, y. For example, consider a regression on a single discrete predictor, x, which takes on the values 0, 1, and 2, with one-third of the population in each category. Suppose the true regression line is $y = 0.2 + 0.5x$ with normally distributed errors with standard deviation 0.1. Then a graph of the data y will show three fairly sharp modes centered at 0.2, 0.7, and 1.2. Other examples of such mixture distributions arise in economics, when including both employed and unemployed people, or

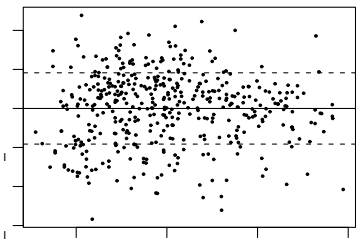

Figure 3.12 *Residual plot for child test score data when regressed on maternal IQ, with dotted lines showing ±1 standard-deviation bounds. The residuals show no striking patterns.*

the study of elections, when comparing districts with incumbent legislators of different parties.

Further assumptions are necessary if a regression coefficient is to be given a causal interpretation, as we discuss in Chapters 9 and 10.

Plotting residuals to reveal aspects of the data not captured by the model

A good way to diagnose violations of some of the assumptions just considered (importantly, linearity) is to plot the residuals r_i versus fitted values $X_i\hat{\beta}$ or simply individual predictors x_i; Figure 3.12 illustrates for the test scores example where child's test score is regressed simply on mother's IQ. The plot looks fine; there do not appear to be any strong patterns. In other settings, residual plots can reveal systematic problems with model fit, as is illustrated, for example, in Chapter 6.

3.7 Prediction and validation

Sometimes the goal of our model is to make predictions using new data. In the case of predictions of future time points, these data may eventually become available, allowing the researcher to see how well the model works for this purpose. Sometimes out-of-sample predictions are made for the explicit purpose of model checking, as we illustrate next.

Prediction

From model (3.4) on page 33, we would predict that a child of a mother who graduated from high school and with IQ of 100 would achieve a test score of $26 + 6 \cdot 1 + 0.6 \cdot 100 = 92$. If this equation represented the true model, rather than an estimated model, then we could use $\hat{\sigma} = 18$ as an estimate of the standard error for our prediction. Actually, the estimated error standard deviation is slightly higher than $\hat{\sigma}$, because of uncertainty in the estimate of the regression parameters—a complication that gives rise to those special prediction standard errors seen in most

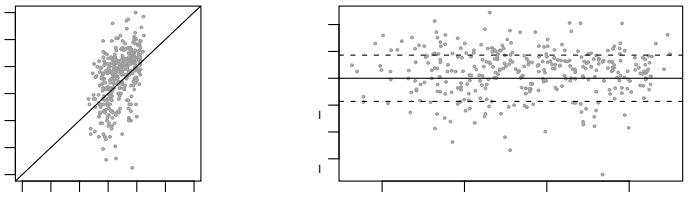

Figure 3.13 *Plots assessing how well the model fit to older children works in making predictions for younger children. The first panel compares predictions for younger children from a model against their actual values. The second panel compares residuals from these predictions against the predicted values.*

regression texts.[3] In R we can create a data frame for the new data and then use the `predict()` function. For example, the following code gives a point prediction and 95% predictive interval:

R code

```
x.new <- data.frame (mom.hs=1, mom.iq=100)
predict (fit.3, x.new, interval="prediction", level=0.95)
```

More generally, we can propagate predictive uncertainty using simulation, as explained in Section 7.2.

We use the notation $\tilde{y}_i$ for the outcome measured on a new data point and $\tilde{X}_i$ for the vector of predictors (in this example, $\tilde{X}_i = (1, 1, 100)$). The predicted value from the model is $\tilde{X}_i\hat{\beta}$, with a predictive standard error slightly higher than $\hat{\sigma}$. The normal distribution then implies that approximately 50% of the actual values should be within $\pm 0.67\hat{\sigma}$ of the predictions, 68% should be within $\pm\hat{\sigma}$, and 95% within $\pm 2\hat{\sigma}$.

We can similarly predict a vector of $\tilde{n}$ new outcomes, $\tilde{y}$, given a $\tilde{n} \times k$ matrix of predictors, $\tilde{X}$; see Figure 3.13.

External validation

The most fundamental way to test a model, in any scientific context, is to use it to make predictions and then compare to actual data.

Figure 3.13 illustrates with the test score data model, which was fit to data collected from 1986 and 1994 for children who were born before 1987. We apply the model to predict the outcomes of children born in 1987 or later (data collected from 1990 to 1998). This is not an ideal example for prediction because we would not necessarily expect the model for the older children to be appropriate for the younger children, even though tests for all children were taken at age 3 or 4. However, we can use it to demonstrate the methods for computing and evaluating predictions. We look at point predictions here and simulation-based predictions in Section 7.2.

The new data, $\tilde{y}$, are the outcomes for the 336 new children predicted from

[3] For example, in linear regression with one predictor, the "forecast standard error" around the prediction from a new data point with predictor value $\tilde{x}$ is

$$\hat{\sigma}_{\text{forecast}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

`mom.iq` and `mom.hs`, using the model fit using the data from the older children. The first panel of Figure 3.13 plots actual values $\tilde{y}_i$ versus predicted values $\tilde{X}_i\hat{\beta}$, and the second panel plots residuals versus predicted values with dotted lines at $\pm\hat{\sigma}$ (approximate 68% error bounds; see Section 2.3). The error plot shows no obvious problems with applying the older-child model to the younger children, though from the scale we detect that the predictions have wide variability.

Even if we had detected clear problems with these predictions, this would not mean necessarily that there is anything wrong with the model as fit to the original dataset. However, we would need to understand it further before generalizing to other children.

3.8 Bibliographic note

Linear regression has been used for centuries in applications in the social and physical sciences; see Stigler (1986). Many introductory statistics texts have good discussions of simple linear regression, for example Moore and McCabe (1998) and De Veaux et al. (2006). Fox (2002) teaches R in the context of applied regression. In addition, the R website links to various useful free literature.

Carlin and Forbes (2004) provide an excellent introduction to the concepts of linear modeling and regression, and Pardoe (2006) is an introductory text focusing on business examples. For fuller treatments, Neter et al. (1996) and Weisberg provide accessible introductions to regression, and Ramsey and Schafer (2001) is a good complement, with a focus on issues such as model understanding, graphical display, and experimental design. Woolridge (2001) presents regression modeling from an econometric perspective. The R^2 summary of explained variance is analyzed by Wherry (1931); see also King (1986) for examples of common mistakes in reasoning with regression and Section 21.9 for more advanced references on R^2 and other methods for summarizing fitted models. Berk (2004) discusses the various assumptions implicit in regression analysis.

For more on children's test scores and maternal employment, see Hill et al. (2005). See Appendix B and Murrell (2005) for more on how to make the sorts of graphs shown in this chapter and throughout the book. The technique of jittering (used in Figure 3.1 and elsewhere in this book) comes from Chambers et al. (1983).

3.9 Exercises

1. The folder `pyth` contains outcome y and inputs x_1, x_2 for 40 data points, with a further 20 points with the inputs but no observed outcome. Save the file to your working directory and read it into R using the `read.table()` function.

   (a) Use R to fit a linear regression model predicting y from x_1, x_2, using the first 40 data points in the file. Summarize the inferences and check the fit of your model.

   (b) Display the estimated model graphically as in Figure 3.2.

   (c) Make a residual plot for this model. Do the assumptions appear to be met?

   (d) Make predictions for the remaining 20 data points in the file. How confident do you feel about these predictions?

   After doing this exercise, take a look at Gelman and Nolan (2002, section 9.4) to see where these data came from.

2. Suppose that, for a certain population, we can predict log earnings from log height as follows:

- A person who is 66 inches tall is predicted to have earnings of $30,000.
- Every increase of 1% in height corresponds to a predicted increase of 0.8% in earnings.
- The earnings of approximately 95% of people fall within a factor of 1.1 of predicted values.

(a) Give the equation of the regression line and the residual standard deviation of the regression.

(b) Suppose the standard deviation of log heights is 5% in this population. What, then, is the R^2 of the regression model described here?

3. In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other.

(a) First generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing `var1 <- rnorm(1000,0,1)` in R. Generate another variable in the same way (call it `var2`). Run a regression of one variable on the other. Is the slope coefficient statistically significant?

(b) Now run a simulation repeating this process 100 times. This can be done using a loop. From each simulation, save the z-score (the estimated coefficient of `var1` divided by its standard error). If the absolute value of the z-score exceeds 2, the estimate is statistically significant. Here is code to perform the simulation:[4]

R code

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}
```

How many of these 100 z-scores are statistically significant?

4. The `child.iq` folder contains a subset of the children and mother data discussed earlier in the chapter. You have access to children's test scores at age 3, mother's education, and the mother's age at the time she gave birth for a sample of 400 children. The data are a Stata file which you can read into R by saving in your working directory and then typing the following:

R code

```
library ("foreign")
iq.data <- read.dta ("child.iq.dta")
```

(a) Fit a regression of child test scores on mother's age, display the data and fitted model, check assumptions, and interpret the slope coefficient. When do you recommend mothers should give birth? What are you assuming in making these recommendations?

(b) Repeat this for a regression that further includes mother's education, interpreting both slope coefficients in this model. Have your conclusions about the timing of birth changed?

[4] We have initialized the vector of z-scores with missing values (NAs). Another approach is to start with `z.scores <- numeric(length=100)`, which would initialize with a vector of zeroes. In general, however, we prefer to initialize with NAs, because then when there is a bug in the code, it sometimes shows up as NAs in the final results, alerting us to the problem.

(c) Now create an indicator variable reflecting whether the mother has completed high school or not. Consider interactions between the high school completion and mother's age in family. Also, create a plot that shows the separate regression lines for each high school completion status group.

(d) Finally, fit a regression of child test scores on mother's age and education level for the first 200 children and use this model to predict test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children.

5. The folder `beauty` contains data from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

(a) Run a regression using beauty (the variable `btystdave`) to predict course evaluations (`courseevaluation`), controlling for various other inputs. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the residual standard deviation. Plot the residuals versus fitted values.

(b) Fit some other models, including beauty and also other input variables. Consider at least one model with interactions. For each model, state what the *predictors* are, and what the *inputs* are (see Section 2.1), and explain the meaning of each of its coefficients.

See also Felton, Mitchell, and Stinson (2003) for more on this topic.

CHAPTER 4

Linear regression: before and after fitting the model

It is not always appropriate to fit a classical linear regression model using data in their raw form. As we discuss in Sections 4.1 and 4.4, linear and logarithmic transformations can sometimes help in the interpretation of the model. Nonlinear transformations of the data are sometimes necessary to more closely satisfy additivity and linearity assumptions, which in turn should improve the fit and predictive power of the model. Section 4.5 presents some other univariate transformations that are occasionally useful. We have already discussed interactions in Section 3.3, and in Section 4.6 we consider other techniques for combining input variables.

4.1 Linear transformations

Linear transformations do not affect the fit of a classical regression model, and they do not affect predictions: the changes in the inputs and the coefficients cancel in forming the predicted value $X\beta$.[1] However, well-chosen linear transformation can improve interpretability of coefficients and make a fitted model easier to understand. We saw in Chapter 3 how linear transformations can help with the interpretation of the intercept; this section provides examples involving the interpretation of the other coefficients in the model.

Scaling of predictors and regression coefficients. The regression coefficient β_j represents the average difference in y comparing units that differ by 1 unit on the j^{th} predictor and are otherwise identical. In some cases, though, a difference of 1 unit on the x-scale is not the most relevant comparison. Consider, for example, a model fit to data we downloaded from a survey of adult Americans in 1994 that predicts their earnings (in dollars) given their height (in inches) and sex (coded as 1 for men and 2 for women):

$$\text{earnings} = -61000 + 1300 \cdot \text{height} + \text{error}, \tag{4.1}$$

with a residual standard deviation of 19000. (A linear model is not really appropriate for these data, as we shall discuss soon, but we'll stick with the simple example for introducing the concept of linear transformations.)

Figure 4.1 shows the regression line and uncertainty on a scale with the x-axis extended to zero to display the intercept—the point on the y-axis where the line crosses zero. The estimated intercept of -61000 has little meaning since it corresponds to the predicted earnings for a person of zero height.

Now consider the following alternative forms of the model:

$$\begin{aligned} \text{earnings} &= -61000 + 51 \cdot \text{height (in millimeters)} + \text{error} \\ \text{earnings} &= -61000 + 81000000 \cdot \text{height (in miles)} + \text{error}. \end{aligned}$$

How important is height? While \$51 does not seem to matter very much, \$81,000,000

[1] In contrast, in a multilevel model, linear transformations can change the fit of a model and its predictions, as we explain in Section 13.6.

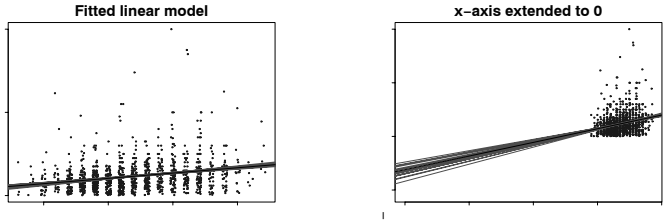

Figure 4.1 *Regression of earnings on height, earnings* $= -61000 + 1300 \cdot$ *height, with solid line showing the fitted regression model and light lines indicating uncertainty in the fitted regression. In the plot on the right, the* x*-scale is extended to zero to reveal the intercept of the regression line.*

is a lot. Yet, both these equations reflect the same underlying information. To understand these coefficients better, we need some sense of the variation in height in the population to which we plan to apply the model. One approach is to consider the standard deviation of heights in the data, which is 3.8 inches (or 97 millimeters, or 0.000061 miles). The expected difference in earnings corresponding to a 3.8-inch difference in height is $\$1300 \cdot 3.8 = \$51 \cdot 97 = \$81000000 \cdot 0.000061 = \$4900$, which is reasonably large but much smaller than the residual standard deviation of $19000 unexplained by the regression.

Standardization using z-scores

Another way to scale the coefficients is to *standardize* the predictor by subtracting the mean and dividing by the standard deviation to yield a "z-score." In this example, `height` would be replaced by `z.height` $=$ (`height` $- 66.9)/3.8$, and the coefficient for `z.height` will be 4900. Then coefficients are interpreted in units of standard deviations with respect to the corresponding predictor just as they were, after the fact, in the previous example. In addition, standardizing predictors using z-scores will change our interpretation of the intercept to the mean of y when all predictor values are at their mean values.

We actually prefer to divide by 2 standard deviations to allow inferences to be more consistent with those for binary inputs, as we discuss in Section 4.2.

Standardization using reasonable scales

It is often useful to keep inputs on familiar scales such as inches, dollars, or years, but making convenient rescalings to aid in the interpretability of coefficients. For example, we might work with income/$10000 or age/10.

For another example, in some surveys, party identification is on a 1–7 scale, from strong Republican to strong Democrat. The rescaled variable $(\text{PID} - 4)/2$, equals -1 for Republicans, 0 for moderates, and $+1$ for Democrats, and so the coefficient on this variable is directly interpretable.

4.2 Centering and standardizing, especially for models with interactions

Figure 4.1b illustrates the difficulty of interpreting the intercept term in a regression in a setting where it does not make sense to consider predictors set to zero. More generally, similar challenges arise in interpreting coefficients in models with interactions, as we saw in Section 3.3 with the following model:

R output

```
lm(formula = kid.score ~ mom.hs + mom.iq + mom.hs:mom.iq)
             coef.est coef.se
(Intercept)      -11.5    13.8
mom.hs            51.3    15.3
mom.iq             1.1     0.2
mom.hs:mom.iq     -0.5     0.2
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

The coefficient on `mom.hs` is 51.3—does this mean that children with mothers who graduated from high school do, on average, 51.3 points better on their tests? No. The model includes an interaction, and 51.3 is the predicted difference for kids that differ in `mom.hs`, *among those with* `mom.iq` $= 0$. Since `mom.iq` is never even close to zero (see Figure 3.4 on page 35), the comparison at zero, and thus the coefficient of 51.3, is essentially meaningless.

Similarly, the coefficient of 1.1 for "main effect" of `mom.iq` is the slope for this variable, among those children for whom `mom.hs` $= 0$. This is less of a stretch (since `mom.hs` actually does equal zero for many of the cases in the data; see Figure 3.1 on page 32) but still can be somewhat misleading since `mom.hs` $= 0$ is at the edge of the data.

Centering by subtracting the mean of the data

We can simplify the interpretation of the regression model by first subtracting the mean of each input variable:

R code

```
c.mom.hs <- mom.hs - mean(mom.hs)
c.mom.iq <- mom.iq - mean(mom.iq)
```

The resulting regression is easier to interpret, with each main effect corresponding to a predictive difference with the other input at its average value:

R output

```
lm(formula = kid.score ~ c.mom.hs + c.mom.iq + c.mom.hs:c.mom.iq)
                 coef.est coef.se
(Intercept)          87.6     0.9
c.mom.hs              2.8     2.4
c.mom.iq              0.6     0.1
c.mom.hs:c.mom.iq    -0.5     0.2
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

The residual standard deviation and R^2 do not change—linear transformation of the predictors does not affect the fit of a classical regression model—and the coefficient and standard error of the interaction do not change, but the main effects and the intercept move a lot and are now interpretable based on comparison to the mean of the data.

Using a conventional centering point

Another option is to center based on an understandable reference point, for example, the midpoint of the range for `mom.hs` and the population average IQ:

R code

```
c2.mom.hs <- mom.hs - 0.5
c2.mom.iq <- mom.iq - 100
```

In this parameterization, the coefficient of `c2.mom.hs` is the average predictive difference between a child with `mom.hs` = 1 and `mom.hs` = 0, for those children with `mom.iq` = 100. Similarly, the coefficient of `c2.mom.iq` corresponds to a comparison for the case `mom.hs` = 0.5, which includes no actual data but represents a midpoint of the range.

R output

```
lm(formula = kid.score ~ c2.mom.hs + c2.mom.iq + c2.mom.hs:c2.mom.iq)
                    coef.est coef.se
(Intercept)            86.8     1.2
c2.mom.hs               2.8     2.4
c2.mom.iq               0.7     0.1
c2.mom.hs:c2.mom.iq    -0.5     0.2
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

Once again, the residual standard deviation, R^2, and coefficient for the interaction have not changed. The intercept and main effect have changed very little, because the points 0.5 and 100 happen to be close to the mean of `mom.hs` and `mom.iq` in the data.

Standardizing by subtracting the mean and dividing by 2 standard deviations

Centering helped us interpret the main effects in the regression, but it still leaves us with a scaling problem. The coefficient of `mom.hs` is much larger than that of `mom.iq`, but this is misleading, considering that we are comparing the complete change in one variable (mother completed high school or not) to a mere 1-point change in mother's IQ, which is not much at all (see Figure 3.4 on page 35).

A natural step is to scale the predictors by dividing by 2 standard deviations—we shall explain shortly why we use 2 rather than 1—so that a 1-unit change in the rescaled predictor corresponds to a change from 1 standard deviation below the mean, to 1 standard deviation above. Here are the rescaled predictors in the child testing example:

R code

```
z.mom.hs <- (mom.hs - mean(mom.hs))/(2*sd(mom.hs))
z.mom.iq <- (mom.iq - mean(mom.iq))/(2*sd(mom.iq))
```

We can now interpret all the coefficients on a roughly common scale (except for the intercept, which now corresponds to the average predicted outcome with all inputs at their mean):

R output

```
lm(formula = kid.score ~ z.mom.hs + z.mom.iq + z.mom.hs:z.mom.iq)
                    coef.est coef.se
(Intercept)            87.6     0.9
z.mom.hs                2.3     2.0
z.mom.iq               17.7     1.8
z.mom.hs:z.mom.iq     -11.9     4.0
  n = 434, k = 4
  residual sd = 18.0, R-Squared = 0.23
```

Why scale by 2 standard deviations?

We divide by 2 standard deviations rather than 1 to maintain coherence when considering binary input variables. To see this, consider the simplest binary x variable which takes on the values 0 and 1, each with probability 0.5. The standard deviation of x is then $\sqrt{0.5 \cdot 0.5} = 0.5$, and so the standardized variable, $(x - \mu_x)/(2\sigma_x)$, takes on the values ± 0.5, and its coefficient reflects comparisons between $x = 0$ and $x = 1$. In contrast, if we had divided by 1 standard deviation, the rescaled variable takes on the values ± 1, and its coefficient corresponds to half the difference between the two possible values of x. This identity is close to precise for binary inputs even when the frequencies are not exactly equal, since $\sqrt{p(1-p)} \approx 0.5$ when p is not too far from 0.5.

In a complicated regression with many predictors, it can make sense to leave binary inputs as is, and linearly transform continuous inputs, possibly by scaling using the standard deviation. In this case, dividing by 2 standard deviations ensures a rough comparability in the coefficients. In our children's testing example, the predictive difference corresponding to 2 standard deviations of mother's IQ is clearly much higher than the comparison of mothers with and without a high school education.

Multiplying each regression coefficient by 2 standard deviations of its predictor

For models with no interactions, a procedure that is equivalent to centering and rescaling is to leave the regression predictors as is, and then create rescaled regression coefficients by multiplying each β by two times the standard deviation of its corresponding x. This gives a sense of the importance of each variable, controlling for all the others in the linear model. As noted, scaling by 2 (rather than 1) standard deviations allows these scaled coefficients to be comparable to unscaled coefficients for binary predictors.

4.3 Correlation and "regression to the mean"

Consider a regression with a single predictor (in addition to the constant term); thus, $y = a + bx + \text{error}$. If both x and y are standardized—that is, if they are defined as `x <- (x-mean(x))/sd(x)` and `y <- (y-mean(y))/sd(y)`—then the regression intercept is zero and the slope is simply the correlation between x and y. Thus, the slope of a regression of two standardized variables must always be between -1 and 1, or, to put it another way, if a regression slope is more than 1 or less than -1, the variance of y must exceed that of x. In general, the slope of a regression with one predictor is $b = \rho\sigma_y/\sigma_x$, where ρ is the correlation between the two variables and σ_x and σ_y are the standard deviations of x and y.

The principal components line and the regression line

Some of the confusing aspects of regression can be understood in the simple case of standardized variables. Figure 4.2 shows a simulated-data example of standardized variables with correlation (and thus regression slope) 0.5. The left plot shows the *principal component line*, which goes closest through the cloud of points, in the sense of minimizing the sum of squared Euclidean distances between the points and the line. The principal component line in this case is simply $y = x$.

The right plot in Figure 4.2 shows the *regression line*, which minimizes the sum of the squares of the *vertical* distances between the points and the line—it is the

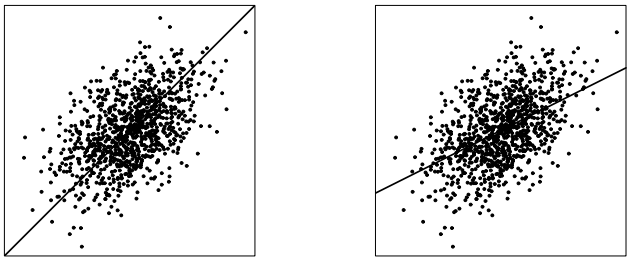

Figure 4.2 *Data simulated from a bivariate normal distribution with correlation 0.5. The regression line, which represents the best prediction of y given x, has half the slope of the principal component line, which goes closest through the cloud of points.*

familiar least squares line, $y = \hat{a} + \hat{b}x$, with $\hat{a}, \hat{b}$ chosen to minimize $\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2$. In this case, $\hat{a} = 0$ and $\hat{b} = 0.5$; the regression line thus has slope 0.5.

When given this sort of scatterplot (without any lines superimposed) and asked to draw the regression line of y on x, students tend to draw the principal component line shown in Figure 4.2a. However, for the goal of predicting y from x, or for estimating the average of y for any given value of x, the regression line is in fact better—even if it does not appear so at first.

The superiority of the regression line for estimating the average of y given x can be seen from a careful study of Figure 4.2. For example, consider the points at the extreme left of either graph. They all lie above the principal components line but are roughly half below and half above the regression line. Thus, the principal component line underpredicts y for low values of x. Similarly, a careful study of the right side of each graph shows that the principal component line overpredicts y for high values of x. In contrast, the regression line again gives unbiased predictions, in the sense of going through the average value of y given x.

Regression to the mean

Recall that when x and y are standardized (that is, placed on a common scale, as in Figure 4.2), the regression line always has slope less than 1. Thus, when x is 1 standard deviations above the mean, the predicted value of y is somewhere between 0 and 1 standard deviations above the mean. This phenomenon in linear models—that y is predicted to be closer to the mean (in standard-deviation units) than x—is called *regression to the mean* and occurs in many vivid contexts.

For example, if a woman is 10 inches taller than the average for her sex, and the correlation of mothers' and (adult) sons' heights is 0.5, then her son's predicted height is 5 inches taller than the average for men. He is expected to be taller than average, but not so much taller—thus a "regression" (in the nonstatistical sense) to the average.

A similar calculation can be performed for any pair of variables that are not perfectly correlated. For example, let x_i and y_i be the number of games won by baseball team i in two successive seasons. They will not be correlated 100%; thus, we would expect the teams that did the best in season 1 (that is, with highest values of x) to do not as well in season 2 (that is, with values of y that are closer

to the average for all the teams). Similarly, we would expect a team with a poor record in season 1 to improve in season 2.

A naive interpretation of regression to the mean is that heights, or baseball records, or other variable phenomena necessarily become more and more "average" over time. This view is mistaken because it ignores the error in the regression predicting y from x. For any data point x_i, the point prediction for its y_i will be regressed toward the mean, but the actual y_i that is observed will not be exactly where it is predicted. Some points end up falling closer to the mean and some fall further. This can be seen in Figure 4.2b.

4.4 Logarithmic transformations

When additivity and linearity (see Section 3.6) are not reasonable assumptions, a nonlinear transformation can sometimes remedy the situation. It commonly makes sense to take the logarithm of outcomes that are all-positive. For outcome variables, this becomes clear when we think about making predictions on the original scale. The regression model imposes no constraints that would force these predictions to be positive as well. However, if we take the logarithm of the variable, run the model, make predictions on the log scale, and then transform back (by exponentiating), the resulting predictions are necessarily positive because for any real a, $\exp(a) > 0$.

Perhaps more importantly, a linear model on the logarithmic scale corresponds to a multiplicative model on the original scale. Consider the linear regression model

$$\log y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i$$

Exponentiating both sides yields

$$\begin{aligned} y_i &= e^{b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i} \\ &= B_0 \cdot B_1^{X_{i1}} \cdot B_2^{X_{i2}} \cdots E_i \end{aligned}$$

where $B_0 = e^{b_0}$, $B_1 = e^{b_1}$, $B_2 = e^{b_2}$, ... are exponentiated regression coefficients (and thus are positive), and $E_i = e^{\epsilon_i}$ is the exponentiated error term (also positive). On the scale of the original data y_i, the predictors $X_{i1}, X_{i2}, \ldots$ come in multiplicatively.

Height and earnings example

We illustrate logarithmic regression by considering models predicting earnings from height. Expression (4.1) on page 53 shows a linear regression of earnings on height. However, it really makes more sense to model earnings on the logarithmic scale (our model here excludes those people who reported zero earnings). We can fit a regression to log earnings and then take the exponential to get predictions on the original scale.

Direct interpretation of small coefficients on the log scale. We take the logarithm of earnings and regress on height,

R code

```
log.earn <- log (earn)
earn.logmodel.1 <- lm (log.earn ~ height)
display (earn.logmodel.1)
```

yielding the following estimate:

R output

```
lm(formula = log.earn ~ height)
            coef.est coef.se
(Intercept)     5.74    0.45
```

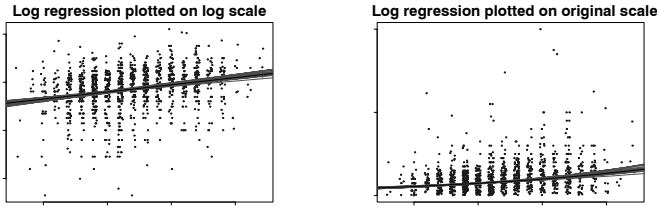

Figure 4.3 *Plot of regression of earnings on height, with solid line showing the fitted log regression model, log(earnings)* $= 5.78 + 0.06 \cdot$ *height, plotted on the logarithmic and untransformed scales. Compare to the linear model (Figure 4.1a).*
scatterplot!data and regression lines superimposed

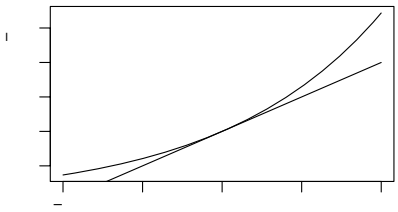

Figure 4.4 *Interpretation of exponentiated coefficients in a logarithmic regression model as relative difference (curved upper line), and the approximation* $exp(x) = 1 + x$*, which is valid for small coefficients* x *(straight line).*

```
height          0.06    0.01
  n = 1192, k = 2
  residual sd = 0.89, R-Squared = 0.06
```

The estimated coefficient $\beta_1 = 0.06$ implies that a difference of 1 inch in height corresponds to an expected positive difference of 0.06 in log(earnings), so that earnings are multiplied by exp(0.06). But exp(0.06) ≈ 1.06 (more precisely, it is 1.062). Thus, a difference of 1 in the predictor corresponds to an expected positive difference of about 6% in the outcome variable. Similarly, if β_1 were -0.06, then a positive difference of 1 inch of height would correspond to an expected *negative* difference of about 6% in earnings.

This correspondence does grow weaker as the magnitude of the coefficient increases. Figure 4.4 displays the deterioration of the correspondence as the coefficient size increases. The plot is restricted to coefficients in the range $(-1, 1)$ because, on the log scale, regression coefficients are typically (though not always) less than 1. A coefficient of 1 on the log scale implies that a change of one unit in the predictor is associated with a change of exp(1) = 2.7 in the outcome, and if predictors are parameterized in a reasonable way, it is unusual to see effects of this magnitude.

Why we use natural log rather than log-base-10

We prefer natural logs (that is, logarithms base e) because, as described above, coefficients on the natural-log scale are directly interpretable as approximate pro-

portional differences: with a coefficient of 0.06, a difference of 1 in x corresponds to an approximate 6% difference in y, and so forth.[2]

Another approach is to take logarithms base 10, which we write as $\log_{10}$. The connection between the two different scales is that $\log_{10}(x) = \log(x)/\log(10) = \log(x)/2.30$. The advantage of $\log_{10}$ is that the predicted values themselves are easier to interpret; for example, when considering the earnings regressions, $\log_{10}(10{,}000) = 4$ and $\log_{10}(100{,}000) = 5$, and with some experience we can also quickly read off intermediate values—for example, if $\log_{10}(\text{earnings}) = 4.5$, then earnings $\approx 30{,}000$.

The disadvantage of $\log_{10}$ is that the resulting coefficients are harder to interpret. For example, if we define

```
log10.earn <- log10 (earn)
```
R code

the regression on height looks like

```
lm(formula = log10.earn ~ height)
            coef.est coef.se
(Intercept) 2.493    0.197
height      0.026    0.003
  n = 1187, k = 2
  residual sd = 0.388, R-Squared = 0.06
```
R output

The coefficient of 0.026 tells us that a difference of 1 inch in height corresponds to a difference of 0.026 in $\log_{10}$(earnings); that is, a multiplicative difference of $10^{0.026} = 1.062$. This is the same 6% change as before, but it cannot be seen by simply looking at the coefficient as could be done on the natural-log scale.

Building a regression model on the log scale

Adding another predictor. Each inch of height corresponds to a 6% increase in earnings—that seems like a lot! But men are mostly taller than women and also tend to have higher earnings. Perhaps the 6% predictive difference can be "explained" by differences between the sexes. Do taller people earn more, on average, than shorter people of the same sex? We can answer this question by including sex into the regression model—in this case, a predictor called `male` that equals 1 for men and 0 for women:

```
lm(formula = log.earn ~ height + male)
            coef.est coef.se
(Intercept)     8.15    0.60
height          0.02    0.01
male            0.42    0.07
  n = 1192, k = 3
  residual sd = 0.88, R-Squared = 0.09
```
R output

After controlling for sex, an inch of height corresponds to estimated predictive difference of 2%: under this model, two persons of the same sex but differing by 1 inch in height will differ, on average, by 2% in earnings. The predictive comparison of sex, however, is huge: comparing a man and a woman of the same height, the man's earnings are $\exp(0.42) = 1.52$ times the woman's; that is, 52% more. (We cannot simply convert the 0.42 to 42% because this coefficient is not so close to zero; see Figure 4.4.)

[2] Natural log is sometimes written as "ln" or "$\log_e$" but we simply write "log" since this is our default.

Naming inputs. Incidentally, we named this new input variable `male` so that it could be immediately interpreted. Had we named it `sex`, for example, we would always have to go back to the coding to check whether 0 and 1 referred to men and women, or vice versa.[3]

Checking statistical significance. The difference between the sexes is huge and well known, but the height comparison is interesting too—a 2% difference, for earnings of $50,000, comes to a nontrivial $1000 per inch. To judge statistical significance, we can check to see if the estimated coefficient is more than 2 standard errors from zero. In this case, with an estimate of 0.02 and standard error of 0.01, we would need to display to three decimal places to be sure (using the `digits` option in the `display()` function):

R output

```
lm(formula = log.earn ~ height + male)
            coef.est coef.se
(Intercept)    8.153   0.603
height         0.021   0.009
male           0.423   0.072
  n = 1192, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

The coefficient for height indeed is statistically significant. Another way to check significance is to directly compute the 95% confidence interval based on the inferential simulations, as we discuss in Section 7.2.

Residual standard deviation and R^2. Finally, the regression model has a residual standard deviation of 0.88, implying that approximately 68% of log earnings will be within 0.88 of the predicted value. On the original scale, approximately 68% of earnings will be within a factor of $\exp(0.88) = 2.4$ of the prediction. For example, a 70-inch person has predicted earnings of $8.153 + 0.021 \cdot 70 = 9.623$, with a predictive standard deviation of approximately 0.88. Thus, there is an approximate 68% chance that this person has log earnings in the range $[9.623 \pm 0.88] = [8.74, 10.50]$, which corresponds to earnings in the range $[\exp(8.74), \exp(10.50)] = [6000, 36000]$. This very wide range tells us that the regression model does not predict earnings well—it is not very impressive to have a prediction that can be wrong by a factor of 2.4—and this is also reflected in the R^2, which is only 0.09, indicating that only 9% of the variance in the data is explained by the regression model. This low R^2 manifests itself graphically in Figure 4.3, where the range of the regression predictions is clearly much narrower than the range of the data.

Including an interaction. We now consider a model with an interaction between height and sex, so that the predictive comparison for height can differ for men and women:

R code

```
earn.logmodel.3 <- lm (log.earn ~ height + male + height:male)
```

which yields

R output

```
            coef.est coef.se
(Intercept)    8.388   0.844
height         0.017   0.013
male          -0.079   1.258
height:male    0.007   0.019
  n = 1192, k = 4
  residual sd = 0.88, R-Squared = 0.09
```

[3] Another approach would be to consider `sex` variable as a factor with two named levels, `male` and `female`; see page 68. Our point here is that, if the variable is coded numerically, it is convenient to give it the name `male` corresponding to the coding of 1.

That is,

$$\log(\text{earnings}) = 8.4 + 0.017 \cdot \text{height} - 0.079 \cdot \text{male} + 0.007 \cdot \text{height} \cdot \text{male}. \quad (4.2)$$

We shall interpret each of the four coefficients in this model.

- The *intercept* is the predicted log earnings if `height` and `male` both equal zero. Because heights are never close to zero, the intercept has no direct interpretation.
- The coefficient for `height` is the predicted difference in log earnings corresponding to a 1-inch difference in height, if `male` equals zero. Thus, the estimated predictive difference per inch of height is 1.7% for women. The estimate is less than 2 standard errors from zero, indicating that the data are consistent with a zero or negative predictive difference also.
- The coefficient for `male` is the predicted difference in log earnings between women and men, if `height` equals 0. Heights are never close to zero, and so the coefficient for `male` has no direct interpretation in this model. (We have already encountered this problem; for example, consider the difference between the intercepts of the two lines in Figure 3.4b on page 35.)
- The coefficient for `height:male` is the difference in slopes of the lines predicting log earnings on height, comparing men to women. Thus, an inch of height corresponds to 0.7% more of an increase in earnings among men than among women, and the estimated predictive difference per inch of height among men is $1.7\% + 0.7\% = 2.4\%$.

The interaction coefficient is not statistically significant, but it is plausible that the correlation between height and earnings is stronger for men and women, and so we keep it in the model, following general principles we discuss more fully in Section 4.6.

Linear transformation to make coefficients more interpretable. We can make the parameters in the interaction model clearer to interpret by rescaling the height predictor to have a mean of 0 and standard deviation 1:

R code
```
z.height <- (height - mean(height))/sd(height)
```

For these data, `mean(height)` and `sd(height)` are 66.9 inches and 3.8 inches, respectively. Fitting the model to `z.height`, `male`, and their interaction yields

R output
```
lm(formula = log.earn ~ z.height + male + z.height:male)
             coef.est coef.se
(Intercept)      9.53    0.05
z.height         0.07    0.05
male             0.42    0.07
z.height:male    0.03    0.07
  n = 1192, k = 4
  residual sd = 0.88, R-Squared = 0.09
```

We can now interpret all four of the coefficients:

- The *intercept* is the predicted log earnings if `z.height` and `male` both equal zero. Thus, a 66.9-inch tall woman is predicted to have log earnings of 9.53, and thus earnings of $\exp(9.53) = 14000$.
- The coefficient for `z.height` is the predicted difference in log earnings corresponding to a 1 standard-deviation difference in height, if `male` equals zero. Thus, the estimated predictive difference for a 3.8-inch increase in height is 7% for women.

- The coefficient for `male` is the predicted difference in log earnings between women and men, if `z.height` equals 0. Thus, a 66.9-inch man is predicted to have log earnings that are 0.42 higher than that of a 66.9-inch woman. This corresponds to a ratio of $\exp(0.42) = 1.52$, so the man is predicted to have 52% higher earnings than the woman.

- The coefficient for `z.height:male` is the difference in slopes between the predictive differences for height among women and men. Thus, a 3.8-inch difference of height corresponds to 3% more of an increase in earnings for men than for women, and the estimated predictive comparison among men is $7\% + 3\% = 10\%$.

One might also consider centering the predictor for sex, but here it is easy enough to interpret `male` = 0, which corresponds to the baseline category (in this case, women).

Further difficulties in interpretation

For a glimpse into yet another difficulty in interpreting regression coefficients, consider the simpler log earnings regression without the interaction term. The predictive interpretation of the height coefficient is simple enough: comparing two adults of the same sex, the taller person will be expected to earn 2% more per inch of height (see the model on page 61). This seems to be a reasonable comparison.

For the coefficient for sex, we would say: comparing two adults of the same height but different sex, the man will be expected to earn 52% more. But is this a relevant comparison? For example, if we are comparing a 66-inch woman to a 66-inch man, then we are comparing a tall woman to a short man. So, in some sense, they do not differ only in sex. Perhaps a more reasonable comparison would be of an "average woman" to an "average man."

The ultimate solution to this sort of problem must depend on why the model is being fit in the first place. For now we shall focus on the technical issues of fitting reasonable models to data. We return to issues of interpretation in Chapters 9 and 10.

Log-log model: transforming the input and outcome variables

If the log transformation is applied to an input variable as well as the outcome, the coefficient can be interpreted as the expected proportional change in y per proportional change in x. For example:

R output

```
lm(formula = log.earn ~ log.height + male)
            coef.est coef.se
(Intercept)     3.62    2.60
log.height      1.41    0.62
male            0.42    0.07
  n = 1192, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

For each 1% difference in height, the predicted difference in earnings is 1.41%. The other input, `male`, is categorical so it does not make sense to take its logarithm.

In economics, the coefficient in a log-log model is sometimes called an "elasticity"; see Exercise 4.6 for an example.

Taking logarithms even when not necessary

If a variable has a narrow dynamic range (that is, if the ratio between the high and low values is close to 1), then it will not make much of a difference in fit if the regression is on the logarithmic or the original scale. For example, the standard deviation of `log.height` in our survey data is 0.06, meaning that heights in the data vary by only approximately a factor of 6%.

In such a situation, it might seem to make sense to stay on the original scale for reasons of simplicity. However, the logarithmic transformation can make sense even here, because coefficients are often more easily understood on the log scale. The choice of scale comes down to interpretability: whether it is easier to understand the model as proportional increase in earnings per inch, or per proportional increase in height.

For an input with a larger amount of relative variation (for example, heights of children, or weights of animals), it would make sense to work with its logarithm immediately, both as an aid in interpretation and likely an improvement in fit too.

4.5 Other transformations

Square root transformations

The square root is sometimes useful for compressing high values more mildly than is done by the logarithm. Consider again our height and earnings example.

Fitting a linear model to the raw, untransformed scale seemed inappropriate. Expressed in a different way than before, we would expect the differences between people earning nothing versus those earning $10,000 to be far greater than the differences between people earning, say, $80,000 versus $90,000. But under the linear model, these are all equal increments (as in model (4.1)), where an extra inch is worth $1300 more in earnings at all levels.

On the other hand, the log transformation seems too severe with these data. With logarithms, the differences between populations earning $5000 versus $10,000 is equivalent to the differences between those earning $40,000 versus those earning $80,000. On the square root scale, however, the differences between the 0 earnings and $10,000 earnings groups are about the same as comparisons between $10,000 and $40,000 or between $40,000 and $90,000. (These move from 0 to 100, 200, and 300 on the square root scale.) See Chapter 25 for more on this example.

Unfortunately, models on the square root scale lack the clean interpretation of the original-scale and log-transformed models. For one thing, large negative predictions on this scale get squared and become large positive values on the original scale, thus introducing a nonmonotonicity in the model. We are more likely to use the square root model for prediction than with models whose coefficients we want to understand.

Idiosyncratic transformations

Sometimes it is useful to develop transformations tailored for specific problems. For example, with the original height-earnings data it would have not been possible to simply take the logarithm of earnings as many observations had zero values. Instead, a model can be constructed in two steps: (1) model the probability that earnings exceed zero (for example, using a logistic regression; see Chapter 5); (2) fit a linear regression, conditional on earnings being positive, which is what we did

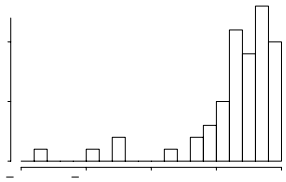

Figure 4.5 *Histogram of handedness scores of a sample of students. Scores range from* -1 *(completely left-handed) to* $+1$ *(completely right-handed) and are based on the responses to ten questions such as "Which hand do you write with?" and "Which hand do you use to hold a spoon?" The continuous range of responses shows the limitations of treating handedness as a dichotomous variable. From Gelman and Nolan (2002).*

in the example above. One could also model total income, but economists are often interested in modeling earnings alone.

In any case, plots and simulation should definitely be used to summarize inferences, since the coefficients of the two parts of the model combine nonlinearly in their joint prediction of earnings. We discuss this sort of model further in Sections 6.7 and 7.4.

What sort of transformed scale would be appropriate for a variable such as "assets" that can be negative, positive, or zero? One possibility is a discrete coding that compresses the high range, for example, 0 for assets in the range [$-\$100, \$100$], 1 for assets between \$100 and \$1000, 2 for assets between \$1000 and \$10,000, and so forth, and -1 for assets between $-\$100$ and $-\$10{,}000$, and so forth. Such a mapping could be expressed more fully as a continuous transformation, but for explanatory purposes it can be convenient to use a discrete scale.

Using continuous rather than discrete predictors

Many variables that appear binary or discrete can usefully be viewed as continuous. For example, rather than define "handedness" as -1 for left-handers and $+1$ for right-handers, one can use a standard ten-question handedness scale that gives an essentially continuous scale from -1 to 1 (see Figure 4.5).

We avoid discretizing continuous variables (except as a way of simplifying a complicated transformation, as described previously, or to model nonlinearity, as described later). A common mistake is to take a numerical measure and replace it with a binary "pass/fail" score. For example, suppose we tried to predict election winners, rather than continuous votes. Such a model would not work well, as it would discard much of the information in the data (for example, the distinction between a candidate receiving 51% or 65% of the vote). The model would be "wasting its effort" in the hopeless task of predicting the winner in very close cases. Even if our only goal is to predict the winners, we are better off predicting continuous vote shares and then transforming them into predictions about winners, as in our example with congressional elections in Section 7.3.

Using discrete rather than continuous predictors

In some cases, however, it is appropriate to discretize a continuous variable if a simple monotonic or quadratic relation does not seem appropriate. For example, in

modeling political preferences, it can make sense to include age with four indicator variables: 18–29, 29–44, 45–64, and 65+, to allow for different sorts of generational patterns. Furthermore, variables that assign numbers to categories that are ordered but for which the gaps between neighboring categories are not always equivalent are often good candidates for discretization.

As an example, Chapter 3 described models for children's test scores given information about their mothers. Another input variable that can be used in these models is maternal employment, which is defined on a four-point ordered scale:

- `mom.work` = 1: mother did not work in first three years of child's life
- `mom.work` = 2: mother worked in second or third year of child's life
- `mom.work` = 3: mother worked part-time in first year of child's life
- `mom.work` = 4: mother worked full-time in first year of child's life.

Fitting a simple model using discrete predictors yields

R output

```
lm(formula = kid.score ~ as.factor(mom.work), data = kid.iq)
                     coef.est coef.se
(Intercept)              82.0     2.3
as.factor(mom.work)2      3.8     3.1
as.factor(mom.work)3     11.5     3.6
as.factor(mom.work)4      5.2     2.7
  n = 434, k = 4
  residual sd = 20.2, R-Squared = 0.02
```

This parameterization of the model allows for different averages for the children of mothers corresponding to each category of maternal employment. The "baseline" category (`mom.work` = 1) corresponds to children whose mothers do not go back to work at all in the first three years after the child is born; the average test score for these children is estimated by the intercept, 82.0. The average test scores for the children in the other categories is found by adding the corresponding coefficient to this baseline average. This parameterization allows us to see that the children of mothers who work part-time in the first year after the child is born achieve the highest average test scores, 82.0 + 11.5. These families also tend to be the most advantaged in terms of many other sociodemographic characteristics as well, so a causal interpretation is not warranted.

Index and indicator variables

Index variables divide a population into categories. For example:

- `male` = 1 for males and 0 for females
- `age` = 1 for ages 18–29, 2 for ages 30–44, 3 for ages 45–64, 4 for ages 65+
- `state` = 1 for Alabama, ..., 50 for Wyoming
- `county` indexes for the 3082 counties in the United States.

Indicator variables are 0/1 predictors based on index variables. For example:

- `sex.1` = 1 for females and 0 otherwise
  `sex.2` = 1 for males and 0 otherwise
- `age.1` = 1 for ages 18–29 and 0 otherwise
  `age.2` = 1 for ages 30–44 and 0 otherwise
  `age.3` = 1 for ages 45–64 and 0 otherwise
  `age.4` = 1 for ages 65+ and 0 otherwise

- 50 indicators for `state`
- 3082 indicators for `county`.

As demonstrated in the previous section, including these variables as regression predictors allows for different means for the populations corresponding to each of the categories delineated by the variable.

When to use index or indicator variables. When an input has only two levels, we prefer to code it with a single variable and name it appropriately; for example, as discussed earlier with the earnings example, the name `male` is more descriptive than `sex.1` and `sex.2`.

R also allows variables to be included as *factors* with named *levels*; for example, `sex` would have the levels `male` and `female`. In this book, however, we restrict ourselves to numerically defined variables, which is convenient for mathematical notation and also when setting up models in Bugs.

When an input has multiple levels, we prefer to create an index variable (thus, for example, `age`, which can take on the levels 1, 2, 3, 4), which can then be given indicators if necessary. As discussed in Chapter 11, multilevel modeling offers a general approach to such categorical predictors.

Identifiability

A model is said to be *nonidentifiable* if it contains parameters that cannot be estimated uniquely—or, to put it another way, that have standard errors of infinity. The offending parameters are called *nonidentified*. The most familiar and important example of nonidentifiability arises from collinearity of regression predictors. A set of predictors is collinear if there is a linear combination of them that equals 0 for all the data.

If an index variable takes on J values, then there are J associated indicator variables. A classical regression can include only $J-1$ of any set of indicators—if all J were included, they would be collinear with the constant term. (You could include a full set of J by excluding the constant term, but then the same problem would arise if you wanted to include a new set of indicators. For example, you could not include both of the sex categories and all four of the age categories. It is simpler just to keep the constant term and all but one of each set of indicators.)

For each index variable, the indicator that is excluded from the regression is known as the default, reference, or baseline condition because it is the implied category if all the $J-1$ indicators are set to zero. The default in R is to set the first level of a factor as the reference condition; other options include using the last level as baseline, selecting the baseline, and constraining the coefficients to sum to zero. There is some discussion in the regression literature on how best to set reference conditions, but we will not worry about it, because in multilevel models we can include all J indicator variables at once.

In practice, you will know that a regression is nonidentified because your computer program will give an error or return "NA" for a coefficient estimate (or it will be dropped by the program from the analysis and nothing will be reported except that it has been removed).

4.6 Building regression models for prediction

A model must be created before it can be fit and checked, and yet we put "model building" near the end of this chapter. Why? It is best to have a theoretical model laid out before any data analyses begin. But in practical data analysis it is usually

easiest to start with a simple model and then build in additional complexity, taking care to check for problems along the way.

There are typically many reasonable ways in which a model can be constructed. Models may differ depending on the inferential goals or the way the data were collected. Key choices include how the input variables should be combined in creating predictors, and which predictors should be included in the model. In classical regression, these are huge issues, because if you include too many predictors in a model, the parameter estimates become so variable as to be useless. Some of these issues are less important in multilevel regression but they certainly do not disappear completely.

This section focuses on the problem of building models for prediction. Building models that can yield causal inferences is a related but separate topic that is addressed in Chapters 9 and 10.

General principles

Our general principles for building regression models for prediction are as follows:

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.
2. It is not always necessary to include these inputs as separate predictors—for example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.
3. For inputs that have large effects, consider including their interactions as well.
4. We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance (typically measured at the 5% level; that is, a coefficient is "statistically significant" if its estimate is more than 2 standard errors from zero):
   (a) If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them.
   (b) If a predictor is not statistically significant and does not have the expected sign (for example, incumbency having a negative effect on vote share), consider removing it from the model (that is, setting its coefficient to zero).
   (c) If a predictor *is* statistically significant and does not have the expected sign, then think hard if it makes sense. (For example, perhaps this is a country such as India in which incumbents are generally unpopular; see Linden, 2006.) Try to gather data on potential lurking variables and include them in the analysis.
   (d) If a predictor is statistically significant and has the expected sign, then by all means keep it in the model.

These strategies do not completely solve our problems but they help keep us from making mistakes such as discarding important information. They are predicated on having thought hard about these relationships *before* fitting the model. It's always easier to justify a coefficient's sign after the fact than to think hard ahead of time about what we expect. On the other hand, an explanation that is determined after running the model can still be valid. We should be able to adjust our theories in light of new information.

Example: predicting the yields of mesquite bushes

We illustrate some ideas of model checking with a real-data example that is nonetheless somewhat artificial in being presented in isolation from its applied context. Partly because this example is not a "success story" and our results are inconclusive, it represents the sort of analysis a student might perform in exploring a new dataset.

Data were collected in order to develop a method of estimating the total production (biomass) of mesquite leaves using easily measured parameters of the plant, before actual harvesting takes place. Two separate sets of measurements were taken, one on a group of 26 mesquite bushes and the other on a different group of 20 mesquite bushes measured at a different time of year. All the data were obtained in the same geographical location (ranch), but neither constituted a strictly random sample.

The outcome variable is the total weight (in grams) of photosynthetic material as derived from actual harvesting of the bush. The input variables are:

diam1:	diameter of the canopy (the leafy area of the bush) in meters, measured along the longer axis of the bush
diam2:	canopy diameter measured along the shorter axis
canopy.height:	height of the canopy
total.height:	total height of the bush
density:	plant unit density (# of primary stems per plant unit)
group:	group of measurements (0 for the first group, 1 for the second group)

It is reasonable to predict the leaf weight using some sort of regression model. Many formulations are possible. The simplest approach is to regress `weight` on all of the predictors, yielding the estimates:

R output

```
lm(formula = weight ~ diam1 + diam2 + canopy.height + total.height +
    density + group, data = mesquite)
                coef.est coef.se
(Intercept)       -729     147
diam1              190     113
diam2              371     124
canopy.height      356     210
total.height      -102     186
density            131      34
group             -363     100
  n = 46, k = 7
  residual sd = 269, R-Squared = 0.85
```

To get a sense of the importance of each predictor, it is useful to know the range of each variable:

R output

```
                 min   q25 median   q75    max   IQR
diam1            0.8   1.4    2.0   2.5    5.2   1.1
diam2            0.4   1.0    1.5   1.9    4.0   0.9
canopy.height    0.5   0.9    1.1   1.3    2.5   0.4
total.height     0.6   1.2    1.5   1.7    3.0   0.5
density          1.0   1.0    1.0   2.0    9.0   1.0
group            0.0   0.0    0.0   1.0    1.0   1.0

weight            60   220    360   690   4050   470
```

"IQR" in the last column refers to the *interquartile range*—the difference between the 75^{th} and 25^{th} percentile points of each variable.

But perhaps it is more reasonable to fit on the logarithmic scale, so that effects are multiplicative rather than additive:

R output

```
lm(formula = log(weight) ~ log(diam1) + log(diam2) + log(canopy.height) +
    log(total.height) + log(density) + group, data = mesquite)
                  coef.est coef.se   IQR of predictor
(Intercept)           5.35    0.17    --
log(diam1)            0.39    0.28    0.6
log(diam2)            1.15    0.21    0.6
log(canopy.height)    0.37    0.28    0.4
log(total.height)     0.39    0.31    0.4
log(density)          0.11    0.12    0.3
group                -0.58    0.13    1.0
  n = 46, k = 7
  residual sd = 0.33, R-Squared = 0.89
```

Instead of, "each meter difference in canopy height is associated with an additional 356 grams of leaf weight," we have, "a difference of x% in canopy height is associated with an (approximate) positive difference of $0.37x$% in leaf weight" (evaluated at the same levels of all other variables across comparisons).

So far we have been throwing all the predictors directly into the model. A more "minimalist" approach is to try to come up with a simple model that makes sense. Thinking geometrically, we can predict leaf weight from the volume of the leaf canopy, which we shall roughly approximate as

$$\text{canopy.volume} = \text{diam1} \cdot \text{diam2} \cdot \text{canopy.height}.$$

This model is an oversimplification: the leaves are mostly on the surface of a bush, not in its interior, and so some measure of surface area is perhaps more appropriate. We shall return to this point shortly.

It still makes sense to work on the logarithmic scale:

R output

```
lm(formula = log(weight) ~ log(canopy.volume))
                  coef.est coef.se
(Intercept)           5.17    0.08
log(canopy.volume)    0.72    0.05
  n = 46, k = 2
  residual sd = 0.41, R-Squared = 0.80
```

Thus, leaf weight is approximately proportional to `canopy.volume` to the 0.72 power. It is perhaps surprising that this power is not closer to 1. The usual explanation for this is that there is variation in `canopy.volume` that is unrelated to the weight of the leaves, and this tends to *attenuate* the regression coefficient—that is, to decrease its absolute value from the "natural" value of 1 to something lower. Similarly, regressions of "after" versus "before" typically have slopes of less than 1. (For another example, Section 7.3 has an example of forecasting congressional elections in which the vote in the previous election has a coefficient of only 0.58.)

The regression with only `canopy.volume` is satisfyingly simple, with an impressive R-squared of 80%. However, the predictions are still much worse than the model with all the predictors. Perhaps we should go back and put in the other predictors. We shall define:

$$\begin{aligned} \text{canopy.area} &= \text{diam1} \cdot \text{diam2} \\ \text{canopy.shape} &= \text{diam1}/\text{diam2}. \end{aligned}$$

The set (canopy.volume, canopy.area, canopy.shape) is then just a different parameterization of the three canopy dimensions. Including them all in the model yields:

R output

```
lm(formula = log(weight) ~ log(canopy.volume) + log(canopy.area) +
    log(canopy.shape) + log(total.height) + log(density) + group)
                  coef.est coef.se
(Intercept)           5.35    0.17
log(canopy.volume)    0.37    0.28
log(canopy.area)      0.40    0.29
log(canopy.shape)    -0.38    0.23
log(total.height)     0.39    0.31
log(density)          0.11    0.12
group                -0.58    0.13
  n = 46, k = 7
  residual sd = 0.33, R-Squared = 0.89
```

This fit is identical to that of the earlier log-scale model (just a linear transformation of the predictors), but to us these coefficient estimates are more directly interpretable:

- Canopy volume and area are both positively associated with weight. Neither is statistically significant, but we keep them in because they both make sense: (1) a larger-volume canopy should have more leaves, and (2) conditional on volume, a canopy with larger cross-sectional area should have more exposure to the sun.
- The negative coefficient of `canopy.shape` implies that bushes that are more circular in cross section have more leaf weight (after controlling for volume and area). It is not clear whether we should "believe" this. The coefficient is not statistically significant; we could keep this predictor in the model or leave it out.
- Total height is positively associated with weight, which could make sense if the bushes are planted close together—taller bushes get more sun. The coefficient is not statistically significant, but it seems to make sense to "believe" it and leave it in.
- It is not clear how to interpret the coefficient for `density`. Since it is not statistically significant, maybe we can exclude it.
- For whatever reason, the coefficient for `group` is large and statistically significant, so we must keep it in. It would be a good idea to learn how the two groups differ so that a more relevant measurement could be included for which `group` is a proxy.

This leaves us with a model such as

R output

```
lm(formula = log(weight) ~ log(canopy.volume) + log(canopy.area) +
    group)
                  coef.est coef.se
(Intercept)           5.22    0.09
log(canopy.volume)    0.61    0.19
log(canopy.area)      0.29    0.24
group                -0.53    0.12
  n = 46, k = 4
  residual sd = 0.34, R-Squared = 0.87
```

or

R output

```
lm(formula = log(weight) ~ log(canopy.volume) + log(canopy.area) +
    log(canopy.shape) + log(total.height) + group)
                  coef.est coef.se
(Intercept)           5.31    0.16
log(canopy.volume)    0.38    0.28
```

```
log(canopy.area)       0.41    0.29
log(canopy.shape)     -0.32    0.22
log(total.height)      0.42    0.31
group                 -0.54    0.12
  n = 46, k = 6
  residual sd = 0.33, R-Squared = 0.88
```

We want to include both volume and area in the model, since for geometrical reasons we expect both to be positively predictive of leaf volume. It would also make sense to look at some residual plots to look for any patterns in the data beyond what has been fitted by the model.

Finally, it would seem like a good idea to include interactions of `group` with the other predictors. Unfortunately, with only 46 data points, it turns out to be impossible to estimate these interactions accurately: none of them are statistically significant.

To conclude this example: we have had some success in transforming the outcome and input variables to obtain a reasonable predictive model. However, we do not have any clean way of choosing among the models (or combining them). We also do not have any easy way of choosing between the linear and log-transformation models, or bridging the gap between them. For this problem, the log model seems to make much more sense, but we would also like a data-based reason to prefer it, if it is indeed preferable.

4.7 Fitting a series of regressions

It is common to fit a regression model repeatedly, either for different datasets or to subsets of an existing dataset. For example, one could estimate the relation between height and earnings using surveys from several years, or from several countries, or within different regions or states within the United States.

As discussed in Part 2 of this book, multilevel modeling is a way to estimate a regression repeatedly, partially pooling information from the different fits. Here we consider the more informal procedure of estimating the regression separately—with no pooling between years or groups—and then displaying all these estimates together, which can be considered as an informal precursor to multilevel modeling.[4]

Predicting party identification

Political scientists have long been interested in party identification and its changes over time. We illustrate here with a series of cross-sectional regressions modeling party identification given political ideology and demographic variables.

We use the National Election Study, which asks about party identification on a 1–7 scale (1 = strong Democrat, 2 = Democrat, 3 = weak Democrat, 4 = independent, ..., 7 = strong Republican), which we treat as a continuous variable. We include the following predictors: political ideology (1 = strong liberal, 2 = liberal, ..., 7 = strong conservative), ethnicity (0 = white, 1 = black, 0.5 = other), age (as categories: 18–29, 30–44, 45–64, and 65+ years, with the lowest age category as a baseline), education (1 = no high school, 2 = high school graduate, 3 = some college, 4 =

[4] The method of repeated modeling, followed by time-series plots of estimates, is sometimes called the "secret weapon" because it is so easy and powerful but yet is rarely used as a data-analytic tool. We suspect that one reason for its rarity of use is that, once one acknowledges the time-series structure of a dataset, it is natural to want to take the next step and model that directly. In practice, however, there is a broad range of problems for which a cross-sectional analysis is informative, and for which a time-series display is appropriate to give a sense of trends.

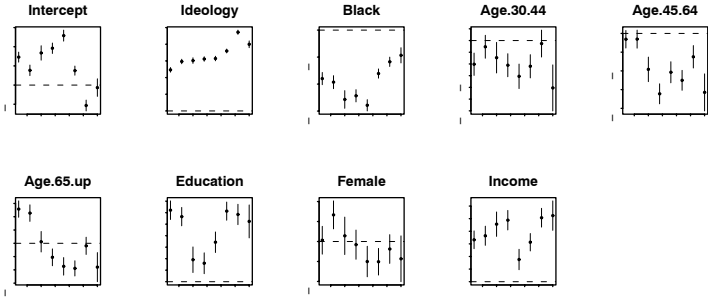

Figure 4.6 *Estimated coefficients (and 50% intervals) for the regression of party identification on political ideology, ethnicity, and other predictors, as fit separately to poll data from each presidential election campaign from 1976 through 2000. The plots are on different scales, with the input variables ordered roughly in declining order of the magnitudes of their coefficients. The set of plots illustrates the display of inferences from a series of regressions.*

college graduate), sex (0 = male, 1 = female), and income (1 = 0–16th percentile, 2 = 17–33rd percentile, 3 = 34–67th percentile, 4 = 68–95th percentile, 5 = 96–100th percentile).

Figure 4.6 shows the estimated coefficients tracked over time. Ideology and ethnicity are the most important,[5] and they remain fairly stable over time. The predictive differences for age and sex change fairly dramatically during the thirty-year period.

4.8 Bibliographic note

For additional reading on transformations, see Atkinson (1985), Mosteller and Tukey (1977), Box and Cox (1964), and Carroll and Ruppert (1981). Bring (1994) has a thoroough discussion on standardizing regression coefficients; see also Blalock (1961) and Greenland, Schlessman, and Criqui (1986). Harrell (2001) discusses strategies for regression modeling.

For more on the earnings and height example, see Persico, Postlewaite, and Silverman (2004) and Gelman and Nolan (2002). For more on the handedness example, see Gelman and Nolan (2002, sections 2.5 and 3.3.2). The historical background of regression to the mean is covered by Stigler (1986), and its connections to multilevel modeling are discussed by Stigler (1983).

The mesquite bushes example in Section 4.6 comes from an exam problem from the 1980s; we have not been able to track down the original data. For more on the ideology example in Section 4.7, see Bafumi (2005).

4.9 Exercises

1. Logarithmic transformation and regression: consider the following regression:

$$\log(\text{weight}) = -3.5 + 2.0\log(\text{height}) + \text{error},$$

[5] Ideology is on a seven-point scale, so that its coefficients must be multiplied by 4 to get the expected change when comparing a liberal (ideology=2) to a conservative (ideology=6).

with errors that have standard deviation 0.25. Weights are in pounds and heights are in inches.

(a) Fill in the blanks: approximately 68% of the persons will have weights within a factor of __ and __ of their predicted values from the regression.

(b) Draw the regression line and scatterplot of log(weight) versus log(height) that make sense and are consistent with the fitted model. Be sure to label the axes of your graph.

2. The folder `earnings` has data from the Work, Family, and Well-Being Survey (Ross, 1990). Pull out the data on earnings, sex, height, and weight.

(a) In R, check the dataset and clean any unusually coded data.

(b) Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model as average earnings for people with average height?

(c) Fit some regression models with the goal of predicting earnings from some combination of sex, height, and weight. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.

(d) Interpret all model coefficients.

3. Plotting linear and nonlinear regressions: we downloaded data with weight (in pounds) and age (in years) from a random sample of American adults. We first created new variables: `age10` = age/10 and `age10.sq` = $(\text{age}/10)^2$, and indicators `age18.29`, `age30.44`, `age45.64`, and `age65up` for four age categories. We then fit some regressions, with the following results:

R output

```
lm(formula = weight ~ age10)
            coef.est coef.se
(Intercept)    161.0     7.3
age10            2.6     1.6
  n = 2009, k = 2
  residual sd = 119.7, R-Squared = 0.00

lm(formula = weight ~ age10 + age10.sq)
            coef.est coef.se
(Intercept)     96.2    19.3
age10           33.6     8.7
age10.sq        -3.2     0.9
  n = 2009, k = 3
  residual sd = 119.3, R-Squared = 0.01

lm(formula = weight ~ age30.44 + age45.64 + age65up)
             coef.est coef.se
(Intercept)     157.2     5.4
age30.44TRUE     19.1     7.0
age45.64TRUE     27.2     7.6
age65upTRUE       8.5     8.7
  n = 2009, k = 4
  residual sd = 119.4, R-Squared = 0.01
```

(a) On a graph of weights versus age (that is, weight on y-axis, age on x-axis), draw the fitted regression line from the first model.

(b) On the same graph, draw the fitted regression line from the second model.

(c) On another graph with the same axes and scale, draw the fitted regression line from the third model. (It will be discontinuous.)

4. Logarithmic transformations: the folder `pollution` contains mortality rates and various environmental factors from 60 U.S. metropolitan areas (see McDonald and Schwing, 1973). For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transformations in regression.

   (a) Create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

   (b) Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

   (c) Interpret the slope coefficient from the model you chose in (b).

   (d) Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

   (e) Cross-validate: fit the model you chose above to the first half of the data and then predict for the second half. (You used all the data to construct the model in (d), so this is not really cross-validation, but it gives a sense of how the steps of cross-validation can be implemented.)

5. Special-purpose transformations: for a study of congressional elections, you would like a measure of the relative amount of money raised by each of the two major-party candidates in each district. Suppose that you know the amount of money raised by each candidate; label these dollar values D_i and R_i. You would like to combine these into a single variable that can be included as an input variable into a model predicting vote share for the Democrats.

   (a) Discuss the advantages and disadvantages of the following measures:

   - The simple difference, $D_i - R_i$
   - The ratio, D_i/R_i
   - The difference on the logarithmic scale, $\log D_i - \log R_i$
   - The relative proportion, $D_i/(D_i + R_i)$.

   (b) Propose an idiosyncratic transformation (as in the example on page 65) and discuss the advantages and disadvantages of using it as a regression input.

6. An economist runs a regression examining the relations between the average price of cigarettes, P, and the quantity purchased, Q, across a large sample of counties in the United States, assuming the following functional form, $\log Q = \alpha + \beta \log P$. Suppose the estimate for β is 0.3. Interpret this coefficient.

7. Sequence of regressions: find a regression problem that is of interest to you and can be performed repeatedly (for example, data from several years, or for several countries). Perform a separate analysis for each year, or country, and display the estimates in a plot as in Figure 4.6 on page 74.

8. Return to the teaching evaluations data from Exercise 3.5. Fit regression models predicting evaluations given many of the inputs in the dataset. Consider interactions, combinations of predictors, and transformations, as appropriate. Consider

several models, discuss in detail the final model that you choose, and also explain why you chose it rather than the others you had considered.