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

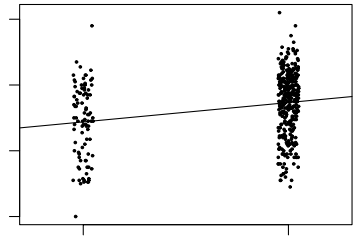

Figure 3.1 *Child's test score plotted versus an indicator for whether mother completed high school. Superimposed is the regression line, which runs through the average of each subpopulation defined by maternal education level. The indicator variable for high school completion has been* jittered*; that 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-

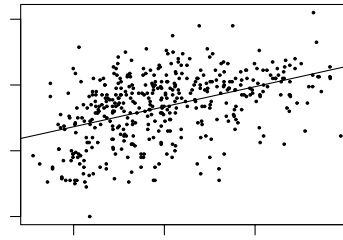

Figure 3.2 *Child's test score plotted versus maternal IQ with regression line superimposed. Each point on the line can be conceived of either as a predicted child test score for children with mothers who have the corresponding IQ, or as the average score for a subpopulation of children with mothers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

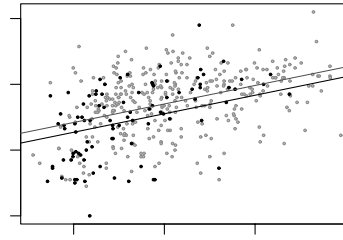

Figure 3.3 *Child's test score plotted versus maternal IQ. Light dots represent children whose mothers graduated from high school and dark dots represent children whose mothers did not graduate from high school. Superimposed are the regression lines from the regression of child's test score on maternal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

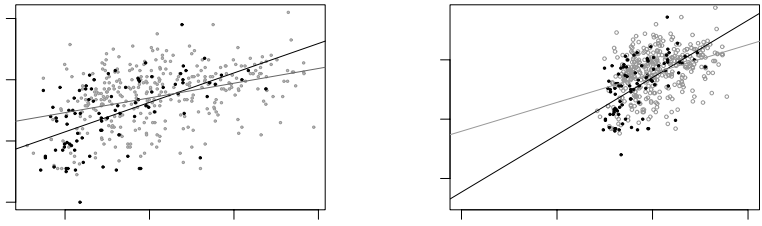

Figure 3.4 *(a) Regression lines of child's test score on mother's IQ with different symbols for children of mothers who completed high school (light circles) and those whose mothers did not complete high school (dark dots). The interaction allows for a different slope in each group, with light and dark 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

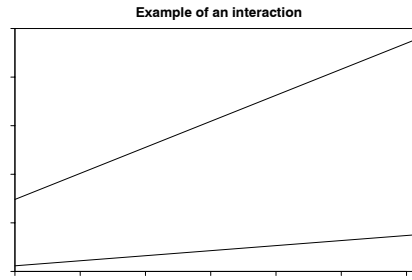

Figure 3.5 *Illustration of interactions between smoking and home radon level on the lifetime probability of lung cancer in men. The effects of radon are much more severe for smokers. The lines are estimated based on case-control studies; see Lin et al. (1999) for references.*

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 × 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.

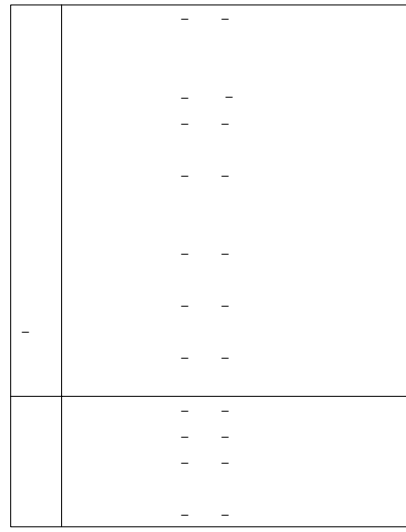

Figure 3.6 *Notation for regression modeling. The model is fit to the observed outcomes y given predictors X. As described in the text, the model can then be applied to predict unobserved outcomes $\tilde{y}$ (indicated by small question marks), given predictors on new data $\tilde{X}$.*

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.

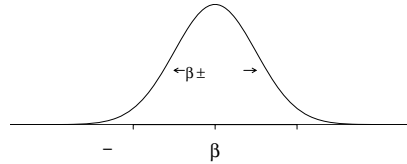

Figure 3.7 *Distribution representing uncertainty in an estimated regression coefficient. The range of this distribution corresponds to the possible values of β that are consistent with the data. When using this as an uncertainty distribution, we assign an approximate 68% chance that β will lie within 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

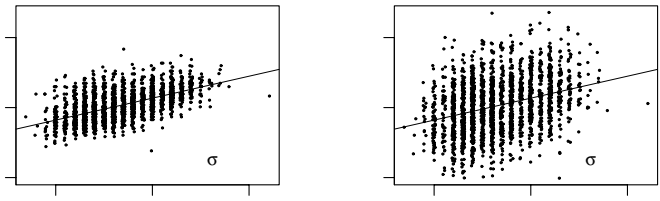

Figure 3.8 *Two hypothetical datasets with the same regression line, $y = a+bx$, but different values of the residual standard deviation, σ. The left plot shows actual data from a survey of adults; the right plot shows data with random noise added to y.*

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

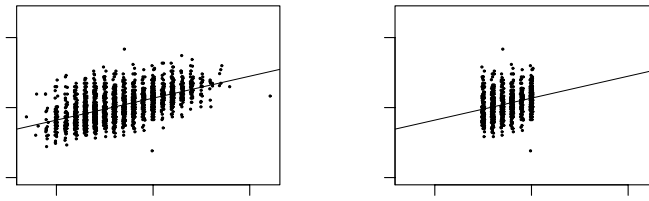

Figure 3.9 *Two hypothetical datasets with the same regression line, $y = a+bx$ and residual standard deviation, σ, but different values of the explained variance, R^2. The left plot shows actual data; the right plot shows data restricted to heights between 65 and 70 inches.*

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.

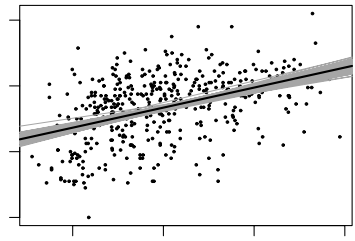

Figure 3.10 *Data and regression of child's test score on maternal IQ, with the solid line showing the fitted regression model and light lines indicating uncertainty in the fitted regression.*

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.

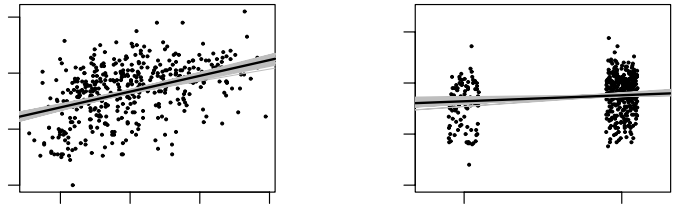

Figure 3.11 *Data and regression of child's test score on maternal IQ and high school completion, shown as a function of each of the two input variables (with light lines indicating uncertainty in the regressions). Values for high school completion have been jittered to make the points more distinct.*

```
  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

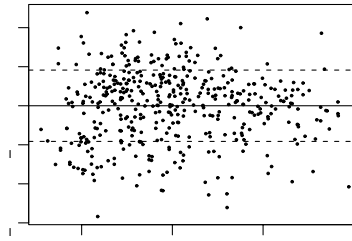

Figure 3.12 *Residual plot for child test score data when regressed on maternal IQ, with dotted lines showing ±1 standard-deviation bounds. The residuals show no striking patterns.*

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

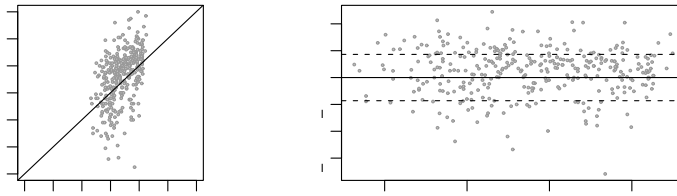

Figure 3.13 *Plots assessing how well the model fit to older children works in making predictions for younger children. The first panel compares predictions for younger children from a model against their actual values. The second panel compares residuals from these predictions against the predicted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