Chapter 10

Linear regression with multiple predictors

Note:
The code examples throughout this chapter use R instead of Python.
However, model formulas are the same as what we would provide to the ols function in Python!

As we move from the simple model, $y = a + bx +$ error to the more general $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots +$ error, complexities arise, involving choices of what predictors x to include in the model, interpretations of the coefficients and how they interact, and construction of new predictors from existing variables to capture discreteness and nonlinearity. We need to learn how to build and understand models as new predictors are added. We discuss these challenges through a series of examples illustrated with R code and graphs of data and fitted models.

10.1 Adding predictors to a model

Example: Children's IQ tests

Regression coefficients are typically 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. The coefficient β_k is the average or expected difference in outcome y, comparing two people who differ by one unit in the predictor x_k while being equal in all the other predictors. This is sometimes stated in shorthand as comparing two people (or, more generally, two observational units) that differ in x_k with all the other predictors held constant. We illustrate with an example, starting with single predictors and then putting them together. We fit a series of regressions predicting cognitive test scores of preschoolers 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).

Starting with a binary predictor

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). We fit the model in R as `stan_glm(kid_score ~ mom_hs, data=kidiq)`, and the result is,[1]

$$\text{kid\_score} = 78 + 12 * \text{mom\_hs} + \text{error}. \tag{10.1}$$

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 10.1 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 set `mom_hs` to 0. To get average test scores for children (or the predicted score for a single child) whose mothers were high school graduates, we would just plug in 1 to obtain $78 + 12 * 1 = 90$.

The difference between these two subpopulation means is equal to the coefficient on `mom_hs`, and it 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.

[1] Data and code for this example are in the folder `KidIQ`.

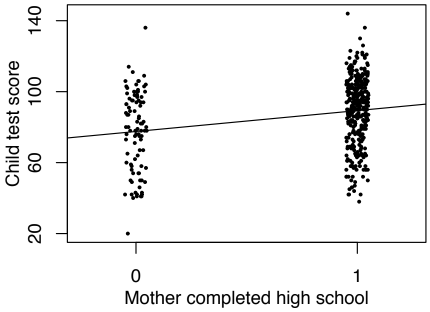


Figure 10.1 *Child's test score plotted versus an indicator for whether mother completed high school. Superimposed is the regression line, which runs through the average of each subpopulation defined by maternal education level. The indicator variable for high school completion has been* jittered*; that is, a random number has been added to each x-value so that the points do not lie on top of each other.*

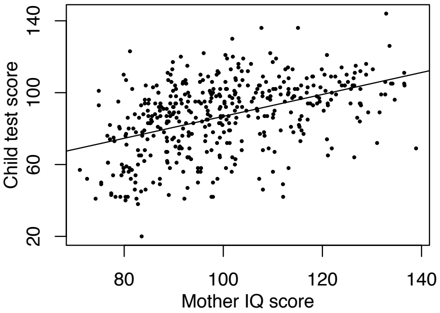


Figure 10.2 *Child's test score plotted versus maternal IQ with regression line superimposed. Each point on the line can be conceived of either as a predicted child test score for children with mothers who have the corresponding IQ, or as the average score for a subpopulation of children with mothers with that IQ.*

A single continuous predictor

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

$$\text{kid\_score} = 26 + 0.6 * \text{mom\_iq} + \text{error} \tag{10.2}$$

and is shown in Figure 10.2. We can think of the line as representing 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.

Including both predictors

Now consider a linear regression of child test scores on two predictors: the maternal high school indicator and maternal IQ. In R, we fit and display a regression with two predictors like this:

```
fit_3 <- stan_glm(kid_score ~ mom_hs + mom_iq, data=kidiq)
print(fit_3)
```

The formula inside the stan_glm(), is the same as what we would provide to ols in Python!

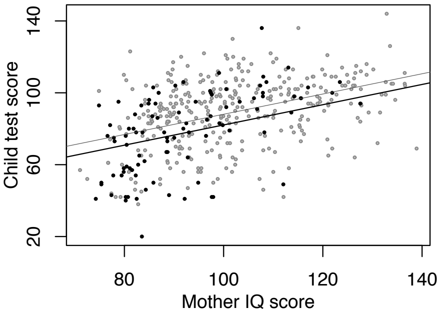


Figure 10.3 *Child's test score plotted versus maternal IQ. Light dots represent children whose mothers graduated from high school and dark dots represent children whose mothers did not graduate from high school. Superimposed are the lines from the regression of child's test score on maternal IQ and 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).*

And here is the result:

```
            Median MAD_SD
(Intercept) 25.7    5.9
mom_hs       6.0    2.4
mom_iq       0.6    0.1

Auxiliary parameter(s):
      Median MAD_SD
sigma 18.2   0.6
```

Understanding the fitted model

The fitted line from the above regression is shown in Figure 10.3 and has the form,

$$\text{kid\_score} = 26 + 6 * \text{mom\_hs} + 0.6 * \text{mom\_iq} + \text{error} \qquad (10.3)$$

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. In Section 10.3 we consider an *interaction* model in which the slopes of the two lines differ. First, however, we interpret the coefficients in model (10.3):

- *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. In Sections 12.1–12.2 we discuss ways to make the intercept more interpretable.
- *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.
- *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).

10.2 Interpreting regression coefficients

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

We interpret 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 items 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 items. 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.

Introductory statistics and regression texts sometimes warn against the latter interpretation but then allow for similar phrasings such as "a change of 10 in maternal IQ is *associated* with a change of 6 points in child's score." The latter expression is not necessarily correct either. From the data alone, a regression only tells us about *comparisons between units*, not about *changes within units*.

Thus, the most careful interpretation of regression coefficients is in terms of comparisons, for example, "When comparing two children whose mothers have the same level of education, the child whose mother is x IQ points higher is predicted to have a test score that is $6x$ higher, on average." Or, "Comparing two items i and j that differ by an amount x on predictor k but are identical on all other predictors, the predicted difference $y_i - y_j$ is $\beta_k x$, on average." This is an awkward way to put things, which helps explain why people often prefer simpler formulations such as "a change of 1 in x_k causes, or is associated with, a change of β in y"—but those sorts of expressions can be terribly misleading. You just have to accept that regression, while a powerful data-analytic tool, can be difficult to interpret. We return in Chapters 18–21 to conditions under which regressions can be interpreted causally.

10.3 Interactions

In model (10.3), 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 10.3 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 defined as the product of these two variables. This allows the slope to vary across subgroups. In, R, we fit and display the model with,

```
fit_4 <- stan_glm(kid_score ~ mom_hs + mom_iq + mom_hs:mom_iq, data=kidiq)
print(fit_4)
```

thus including the *main effects* and their interaction, `mom_hs:mom_iq`. The fitted model,

$$\text{kid\_score} = -11 + 51 * \text{mom\_hs} + 1.1 * \text{mom\_iq} - 0.5 * \text{mom\_hs} * \text{mom\_iq} + \text{error},$$

is displayed in Figure 10.4a, with separate lines for each subgroup defined by maternal education.

The formula inside the stan_glm(), is the same as what we would provide to ols in Python!

Figure 10.4b shows the regression lines on a scale with the x-axis extended to zero to display the

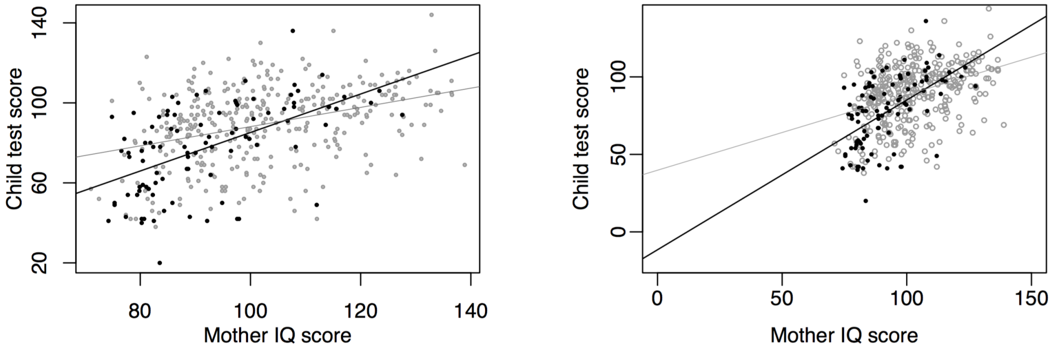


Figure 10.4 *(a) Regression lines of child's test score on mother's IQ with different symbols for children of mothers who completed high school (light circles) and those whose mothers did not complete high school (dark dots). The interaction allows for a different slope in each group, with light and dark lines corresponding to the light and dark points. (b) Same plot but with horizontal and vertical axes extended to zero to reveal the intercepts.*

intercepts—the points on the y-axis where the lines cross zero. This highlights that the intercept has no direct interpretation when the range of possible values of the predictor is far from zero.

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

- *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 12.1–12.2, intercepts can be more interpretable if input variables are centered before including them as regression predictors.
- *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 0, this coefficient is not easily interpretable.
- *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 10.4.
- *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 10.4.

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 did not:

$$\begin{aligned}
\text{mom\_hs} = 0: \quad \text{kid\_score} &= -11 + 51 * 0 + 1.1 * \text{mom\_iq} - 0.5 * 0 * \text{mom\_iq} \\
&= -11 + 1.1 * \text{mom\_iq} \\
\text{mom\_hs} = 1: \quad \text{kid\_score} &= -11 + 51 * 1 + 1.1 * \text{mom\_iq} - 0.5 * 1 * \text{mom\_iq} \\
&= 40 + 0.6 * \text{mom\_iq}.
\end{aligned}$$

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 mother's IQs of 0.

When should we look for interactions?

Example: Radon, smoking, and lung cancer

Interactions can be important, and the first place we typically look for them is with predictors that have large coefficients when not interacted. For a familiar example, smoking is strongly associated with cancer. In epidemiological studies of other carcinogens, it is crucial to adjust for smoking both as an uninteracted predictor and as an interaction, because the strength of association between other risk factors and cancer can depend on whether the individual is a smoker. We illustrated this interaction in Figure 1.7 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—fitting models separately within distinct subgroups versus fitting an interacted model to the full sample—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 preprocess the data by centering each input variable about its mean or some other convenient reference point. We discuss this in Section 12.2 in the context of linear transformations.

10.4 Indicator variables

Example: Height and weight

In Section 7.3 we discussed how to express comparisons using regression with indicator ("dummy") variables. We further explore this idea here, fitting models to predict height from weight and other variables based on survey data. Here are data from the first rows of the data frame `earnings`:[2]

```
  height weight male  earn ethnicity education walk exercise smokenow tense angry age
1     74    210    1 50000     White        16    3        3        2     0     0  45
2     66    125    0 60000     White        16    6        5        1     0     0  58
3     64    126    0 30000     White        16    8        1        2     1     1  29
4     65    200    0 25000     White        17    8        1        2     0     0  57
```

We start by predicting weight (in pounds) from height (in inches):

```
fit_1 <- stan_glm(weight ~ height, data=earnings)
print(fit_1)
```

which yields,

```
            Median MAD_SD
(Intercept) -172.9   11.6
height         4.9    0.2

Auxiliary parameter(s):
      Median MAD_SD
sigma 29.1   0.5
```

The fitted regression line is $\text{weight} = -172.9 + 4.9 * \text{height}$:

- Comparing two people who differ in height by one inch, their expected difference in weight is 4.9 pounds.
- The predicted weight for a 0-inch-tall person is -172.9 pounds . . . hmmm, this doesn't mean very much. The average height of American adults is about 66 inches, so let's state it this way: the predicted weight for a 66-inch-tall person is $-172.9 + 4.9 * 66$ pounds:

```
coefs_1 <- coef(fit_1)
predicted_1 <- coefs_1[1] + coefs_1[2]*66
```

[2] Data and code for this example are in the folder `Earnings`.

Centering a predictor

To improve interpretation of the fitted models, we use a centered version of height as a predictor:

```
earnings$c_height <- earnings$height - 66
fit_2 <- stan_glm(weight ~ c_height, data=earnings)
```

yielding,

```
            Median MAD_SD
(Intercept) 153.4    0.6
c_height      4.9    0.2

Auxiliary parameter(s):
      Median MAD_SD
sigma 29.1   0.5
```

Including a binary variable in a regression

Next we expand the model by including an indicator variable for sex:

```
fit_3 <- stan_glm(weight ~ c_height + male, data=earnings)
```

which yields,

```
            Median MAD_SD
(Intercept) 149.6    1.0
c_height      3.9    0.3
male         12.0    2.0

Auxiliary parameter(s):
      Median MAD_SD
sigma 28.8    0.5
```

The coefficient of 12.0 on `male` tells us that, in these data, comparing a man to a woman of the same height, the man will be predicted to be 12 pounds heavier.

To compute the predicted weight for a 70-inch-tall woman, say:

```
coefs_3 <- coef(fit_3)
predicted <- coefs_3[1] + coefs_3[2]*4.0 + coefs_3[3]*0
```

the result is 165 pounds. The corresponding point prediction for a 70-inch-tall man comes to 177 pounds, which is indeed 12 pounds higher than the prediction for the woman.

Using indicator variables for multiple levels of a categorical predictor

Next we shall include ethnicity in the regression. In our data this variable takes on four levels, as we can see by typing `table(earnings$ethnicity)` in the R console:

```
   Black Hispanic    Other    White 
     177      103       37     1473 
```

We can include ethnicity in our regression as a *factor*:

```
fit_4 <- stan_glm(weight ~ c_height + male + factor(ethnicity), data=earnings)
print(fit_4)
```

which yields,

We would use C(ethnicity) instead of factor(ethnicity)

```
                           Median MAD_SD
(Intercept)                154.1    2.2 
c_height                     3.8    0.3 
male                        12.2    2.0 
factor(ethnicity)Hispanic   -5.9    3.6 
factor(ethnicity)Other     -12.6    5.2 
factor(ethnicity)White      -5.0    2.3 

Auxiliary parameter(s):
      Median MAD_SD
sigma 28.7    0.5  
```

Ethnicity has four levels in our data, but looking carefully at the output, we see only three coefficients for ethnicity, for Hispanics, Others, and Whites. The missing group is Blacks. In computing the regression, R took `Black` to be the *baseline* category against which all other groups are measured.

Thus, the above coefficient of −5.9 implies that, when comparing a Hispanic person and a Black person with the same height and sex, the fitted model predicts the Hispanic person to be 5.9 pounds lighter, on average. Similarly, the model predicts an Other person to be 12.6 pounds lighter and a White person to be 5.0 pounds lighter than a Black person of the same height and sex.

Changing the baseline factor level

When including a factor variable in a regression, any of the levels can be used as the baseline. By default, R orders the factors in alphabetical order, hence in this case `Black` is the first category and is used as the baseline.

We can change the baseline category by directly setting the levels when constructing the factor:

```
earnings$eth <- factor(earnings$ethnicity,
  levels=c("White", "Black", "Hispanic", "Other"))
fit_5 <- stan_glm(weight ~ c_height + male + eth, data=earnings)
print(fit_5)
```

which yields,

It’s a little easier in Python, would use C(ethnicity, reference=‘Black’)

```
             Median MAD_SD
(Intercept) 149.1    1.0  
c_height      3.8    0.2  
male         12.2    2.0  
ethBlack      5.0    2.2  
ethHispanic  -0.9    2.9  
ethOther     -7.6    4.6  

Auxiliary parameter(s):
      Median MAD_SD
sigma 28.3   0.4   
```

This model uses White as the baseline category because we listed it first when setting up the factor variable eth. Going through the coefficients:

- In the earlier fit, the intercept of 154.1 was the predicted weight for a person with c_height = 0, male = 0, and ethnicity = Black. In the new version, the intercept of 149.1 is the predicted value with c_height = 0, male = 0, and ethnicity = White, the new baseline category. The change of 5.0 corresponds to negative of the coefficient of White in the original regression.
- The coefficients for height and male do not change.
- The coefficient for Black, which earlier was 0 by implication, as Black was the baseline category and thus not included in the regression, is now 5.0, which is the difference relative to the new baseline category of White.
- The coefficient for Hispanic has increased from −5.9 to −0.9, a change of 5.0 corresponding to the shift in the baseline from Black to White.
- The coefficient for Other also increases by this same 5.0.
- The coefficient for White has increased from −5.0 to the implicit value of 0.

An alternative approach is to create indicators for the four ethnic groups directly:

```
earnings$eth_White <- ifelse(earnings$ethnicity=="White", 1, 0)
earnings$eth_Black <- ifelse(earnings$ethnicity=="Black", 1, 0)
earnings$eth_Hispanic <- ifelse(earnings$ethnicity=="Hispanic", 1, 0)
earnings$eth_Other <- ifelse(earnings$ethnicity=="Other", 1, 0)
```

It is not necessary to name the new variables in this way, but this sort of naming can sometimes make it easier to keep track. In any case, once we have created these numerical variables we can include them in the usual way, for example:

```
fit_6 <- stan_glm(weight ~ height + male + eth_Black + eth_Hispanic +
  eth_Other, data=earnings)
```

10.7 Mathematical notation and statistical inference

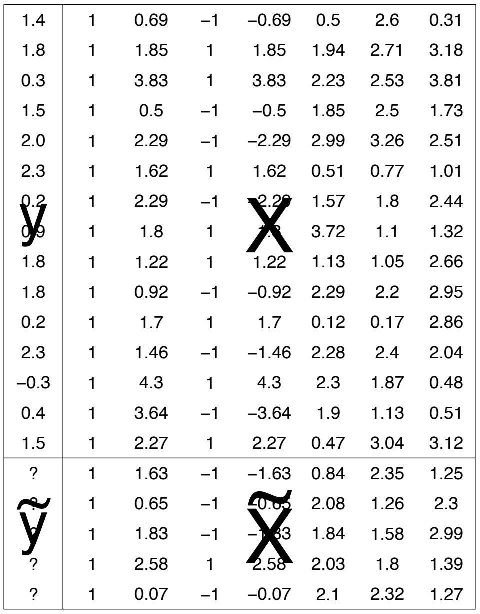
When illustrating specific examples, it helps to use descriptive variable names. In order to discuss more general theory and data manipulations, however, we shall adopt generic mathematical notation. This section introduces this notation and discusses the stochastic aspect of the model as well.

Predictors

Example: Children's IQ tests

We use the term *predictors* for the columns in the X matrix (other than the constant term), and we also sometimes use the term when we want to emphasize the information that goes into the predictors. For example, consider the model that includes the interaction of maternal education and maternal IQ:

$$\text{kid\_score} = 58 + 16 * \text{mom\_hs} + 0.5 * \text{mom\_iq} - 0.2 * \text{mom\_hs} * \text{mom\_iq} + \text{error}.$$



1.4	1	0.69	−1	−0.69	0.5	2.6	0.31
1.8	1	1.85	1	1.85	1.94	2.71	3.18
0.3	1	3.83	1	3.83	2.23	2.53	3.81
1.5	1	0.5	−1	−0.5	1.85	2.5	1.73
2.0	1	2.29	−1	−2.29	2.99	3.26	2.51
2.3	1	1.62	1	1.62	0.51	0.77	1.01
0.2	1	2.29	−1	−2.29	1.57	1.8	2.44
0.9	1	1.8	1	1.8	3.72	1.1	1.32
1.8	1	1.22	1	1.22	1.13	1.05	2.66
1.8	1	0.92	−1	−0.92	2.29	2.2	2.95
0.2	1	1.7	1	1.7	0.12	0.17	2.86
2.3	1	1.46	−1	−1.46	2.28	2.4	2.04
−0.3	1	4.3	1	4.3	2.3	1.87	0.48
0.4	1	3.64	−1	−3.64	1.9	1.13	0.51
1.5	1	2.27	1	2.27	0.47	3.04	3.12
?	1	1.63	−1	−1.63	0.84	2.35	1.25
?	1	0.65	−1	−0.65	2.08	1.26	2.3
?	1	1.83	−1	−1.83	1.84	1.58	2.99
?	1	2.58	1	2.58	2.03	1.8	1.39
?	1	0.07	−1	−0.07	2.1	2.32	1.27

Figure 10.8 *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 $\tilde{X}$.*

This regression has three *predictors*: maternal high school, maternal IQ, and maternal high school * IQ. Depending on context, the constant term is also sometimes called a predictor.

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 people in the data as $i = 1, \ldots, n$. In our most recent example, y_i is the i^{th} child's test score, and there are $n = 1378$ data points and $k = 4$ items 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 people; 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 deviations of the outcomes from the model, called *errors*, are labeled as ϵ_i and assumed to follow a normal distribution with mean 0 and standard deviation σ, which we write as normal$(0, \sigma)$. The term *residual* is used for the differences between the outcomes and predictions from the estimated model. Thus, $y - X\beta$ and $y - X\hat{\beta}$ are the vectors of errors and residuals, respectively. We use the notation $\tilde{y}$ for predictions from the model, given new data $\tilde{X}$; see Figure 10.8.

Conventions vary across disciplines regarding what terms to use for the variables we refer to as predictors and outcomes (or responses). Some use the terms "independent variable" for predictors and "dependent variable" for the outcome. These terms came from a time when regression models were used to model outcomes from experiments where the manipulation of the input variables might have led to predictors that were independent of each other. This is rarely the case in social science, however, so we avoid these terms. Other times the predictors and outcome are called the "left-hand side" and "right-hand side" variables.

Two ways of writing the model

The classical linear regression model can then be written mathematically as

$$y_i = \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad \text{for } i = 1, \dots, n, \tag{10.4}$$

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

An equivalent representation is,

$$y_i = X_i \beta + \epsilon_i, \quad \text{for } i = 1, \dots, n,$$

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

$$y_i \sim \text{normal}(X_i \beta, \sigma), \ \text{for } i = 1, \dots, n.$$

For even more compact notation we can use,

$$y \sim \text{multivariate normal}(X\beta, \sigma^2 I),$$

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 $\hat{\sigma}$, as we demonstrated in Section 8.1 for simple regression with just one predictor and a constant term.

Least squares, maximum likelihood, and Bayesian inference

The steps of estimation and statistical inference in linear regression with multiple predictors are the same as with one predictor, as described in Sections 8.1 and 9.5. The starting point is the least squares estimate, that is, the vector $\hat{\beta}$ that minimizes the sum of the squared residuals, $\text{RSS} = \sum_{i=1}^{n} (y_i - X_i \hat{\beta})^2$. For the standard linear regression model with predictors that are measured accurately and errors that are independent, of equal variance, and normally distributed, the least squares solution is also the maximum likelihood estimate. The only slight difference from Section 8.1 is that the standard estimate of the residual standard deviation is

$$\hat{\sigma} = \sqrt{\text{RSS}/(n-k)}\,, \tag{10.5}$$

where k is the number of regression coefficients. This reduces to formula (8.5) on page 104 for a regression with one predictor, in which case $k = 2$.

Nonidentified parameters, collinearity, and the likelihood function

In maximum likelihood, parameters are nonidentified if they can be changed without altering the likelihood. Continuing with the "hill" analogy from Section 8.1, nonidentifiability corresponds to a "ridge" in the likelihood—a direction in parameter space in which the likelihood is flat.

To put it another way, 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* (also called multicollinearity) of regression predictors. A set of predictors is collinear if there is a linear combination of them that equals 0 for all the data. We discuss this problem in the context of indicator variables at the end of Section 12.5.

A simple example of collinearity is a model predicting family outcomes given the number of boys in the family, the number of girls, and the total number of children. Labeling these predictors as x_{2i}, x_{3i}, and x_{4i}, respectively, for family i (with $x_{1i} = 1$ being reserved for the constant term in the regression), we have $x_{4i} = x_{2i} + x_{3i}$, thus $-x_{2i} - x_{3i} + x_{4i} = 0$ and the set of predictors is collinear.

There can also be problems with *near*-collinearity, which leads to poor identification. For example, suppose you try to predict people's heights from the lengths of their left and right feet. These two predictors are nearly (but not exactly) collinear, and as a result it will be difficult to untangle the two coefficients, and we can characterize the fitted regression as unstable, in the sense that if it were re-fitted with new data, a new sample from the same population, we could see much different results. This instability should be reflected in large standard errors for the estimated coefficients.

Hypothesis testing: why we do not like *t* tests and *F* tests

One thing that we do *not* recommend is traditional null hypothesis significance tests. For reference, we review here the two most common such procedures.

The t test is used to demonstrate that a regression coefficient is statistically significantly different from zero, and it is formally a test of the null hypothesis that a particular coefficient β_j equals zero. Under certain assumptions, the standardized regression coefficient $\hat{\beta}_j/\text{s.e.}_j$ will approximately follow a t_{n-k} distribution under the null hypothesis, and so the null hypothesis can be rejected at a specified significance level if $|\hat{\beta}_j/\text{s.e.}_j|$ exceeds the corresponding quantile of the t_{n-k} distribution.

The F test is used to demonstrate that there is evidence that an entire regression model—not just any particular coefficient—adds predictive power, and it is formally a test of the null hypothesis that *all* the coefficients in the model, except the constant term, equal zero. Under certain assumptions, the ratio of total to residual sum of squares, suitably scaled, follows something called the F distribution, and so the null hypothesis can be rejected if the ratio of sums of squares exceeds some level that depends on both the sample size n and the number of predictors k.

We have essentially no interest in using hypothesis tests for regression because we almost never encounter problems where it would make sense to think of coefficients as being exactly zero. Thus, rejection of null hypotheses is irrelevant, since this just amounts to rejecting something we never took seriously in the first place. In the real world, with enough data, any hypothesis can be rejected.

That said, uncertainty in estimation is real, and we do respect the deeper issue being addressed by hypothesis testing, which is assessing when an estimate is overwhelmed by noise, so that some particular coefficient or set of coefficients could just as well be zero, as far as the data are concerned. We recommend addressing such issues by looking at standard errors as well as parameter estimates, and by using Bayesian inference when estimates are noisy (see Section 9.5), as the use of prior information should stabilize estimates and predictions. When there is the goal of seeing whether good predictions can be made without including some variables in the model, we recommend comparing models using cross validation, as discussed in Section 11.8.