Chapter 12

Transformations and regression

Note:
The code examples in this chapter use R. But we've see how to do similar transformations using polars expressions in Python.
We haven't discussed all of them, so feel free to try them out yourself

It is not always best to fit a regression using data in their raw form. In this chapter we start by discussing linear transformations for standardizing predictors and outcomes in a regression, which connects to "regression to the mean," earlier discussed in Chapter 6, and how it relates to linear transformations and correlation. We then discuss logarithmic and other transformations with a series of examples in which input and outcome variables are transformed and combined in various ways in order to get more understandable models and better predictions. This leads us to more general thoughts about building and comparing regression models in applications, which we develop in the context of an additional example.

12.1 Linear transformations

Scaling of predictors and regression coefficients

Example: Earnings and height

The coefficient β_j represents the average difference in y, comparing items that differ by 1 unit on the j^{th} predictor and are otherwise identical. In some cases, though, a difference of 1 unit in x is not the most relevant comparison. Consider, from page 12, a model fit to data we downloaded from a survey of adult Americans in 1990 that predicts their earnings (in dollars) given their height (in inches):

$$\text{earnings} = -85\,000 + 1600 * \text{height} + \text{error}, \tag{12.1}$$

with a residual standard deviation of 22 000.

A linear model is not really appropriate for this problem, as we shall discuss soon, but we'll stick with the simple example for introducing the concept of linear transformations.

Figure 12.1a shows the regression line and uncertainty along with the data, and Figure 12.1b extends the x-axis to zero to display the intercept—the point on the y-axis where the line crosses zero. The estimated intercept of −85 000 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:

$$\text{earnings} = -85\,000 + 63 * \text{height (in millimeters)} + \text{error},$$
$$\text{earnings} = -85\,000 + 101\,000\,000 * \text{height (in miles)} + \text{error}.$$

How important is height? While \$63 does not seem to matter much, \$101 000 000 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 scale by the standard deviation of heights in the data, which is 3.8 inches (or 97 millimeters, or 0.000 060 miles). The expected difference in earnings corresponding to a 3.8-inch difference in height is \$1600 ∗ 3.8 = \$63 ∗ 97 = \$101 000 000 ∗ 0.000 060 = \$6100, which is reasonably large but much smaller than the residual standard deviation of \$22 000 unexplained by the regression.

Linear transformations of the predictors X or the outcome y do not affect the fit of a classical

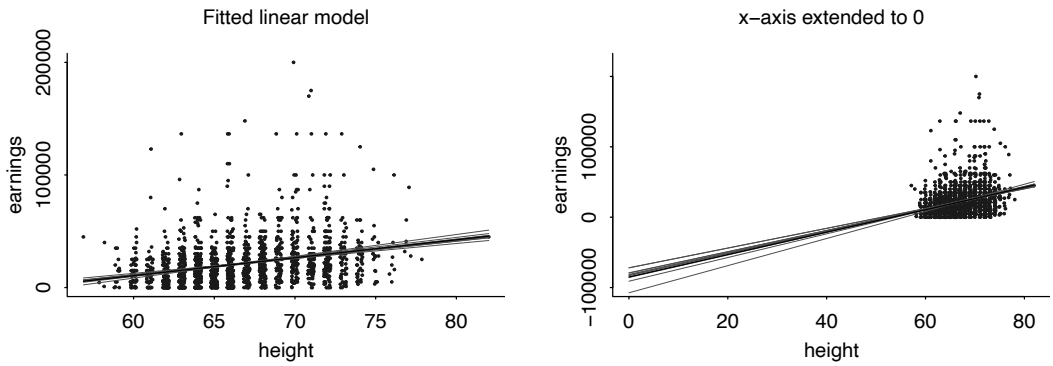

Figure 12.1 *(a) Regression of earnings on height, earnings* $= -85\,000 + 1600 * \textit{height}$*, with lines indicating uncertainty in the fitted regression. (b) Extending the x-scale to zero reveals the estimate and uncertainty for the intercept of the regression line. To improve resolution, a data point at earnings of* \$400 000 *has been excluded from the graphs.*

regression model, and they do not affect predictions; the changes in the inputs and the coefficients cancel in forming the predicted value $X\beta$. However, well-chosen linear transformations can improve interpretability of coefficients and make a fitted model easier to understand. We saw in Chapters 4 and 10 how linear transformations can help with the interpretation of the intercept; this section and the next provide examples involving the interpretation of the other coefficients in the model.

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." For these `height` would be replaced by `z_height` = (`height` − 66.6)/3.8, and the coefficient for `z_height` becomes 6100. 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. This is helpful because standard deviations can be seen as a measure of practical significance; in this case, a difference in one standard deviation on the input scale is a meaningful difference in that it roughly reflects a typical difference between the mean and a randomly drawn observation. 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.

It can often be preferable, however, to divide by 2 standard deviations to allow inferences to be more consistent with those for binary inputs, as we discuss in Section 12.2.

Standardization using the sample mean and standard deviation of the predictors uses raw estimates from the data and thus should be used only when the number of observations is big enough that these estimates are stable. When sample size is small, we recommend standardizing using an externally specified population distribution or other externally specified reasonable scales.

Standardization using an externally specified population distribution

A related approach is to rescale based on some standard set outside the data. For example, in analyses of test scores it is common to express estimates on the scale of standard deviations of test scores across all students in a grade. A test might be on a 0–100 scale, with fourth graders having a national mean score of 55 and standard deviation of 18. Then if the analysis is done on the scale of points on the exam, all coefficient estimates and standard errors from analyses of fourth graders are divided by 18 so that they are on this universal scale. Equivalently, one could first define $z = (y - 55)/18$ for all fourth graders and then run all regressions on z. The virtue of using a fixed scaling, rather than standardizing each dataset separately, is that estimates are all directly comparable.

Standardization using reasonable scales

Sometimes it is useful to keep inputs on familiar scales such as inches, dollars, or years, but make convenient rescalings to aid in the interpretability of coefficients. For example, we might work with income/$10 000 or age in decades.

For another example, in Section 10.9 we analyzed party identification, a variable on a 1–7 scale: 1 = strong Democrat, 2 = Democrat, 3 = weak Democrat, 4 = independent, 5 = weak Republican, 6 = Republican, 7 = strong Republican. Rescaling to (pid − 4)/4 gives us a variable that equals −0.5 for Democrats, 0 for moderates, and +0.5 for Republicans, and so the coefficient on this variable is directly interpretable, with a change of 1 comparing a Democrat to a Republican.

12.2 Centering and standardizing for models with interactions

Example: Children's IQ tests

Figure 12.1b 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 10.3 with the following model:[1]

```
             Median MAD_SD
(Intercept)   -8.0   13.2
mom_hs        47.0   14.5
mom_iq         0.9    0.1
mom_hs:mom_iq -0.5    0.2

Auxiliary parameter(s):
      Median MAD_SD
sigma 18.0   0.6
```

The coefficient on mom_hs is 47.0—does this mean that children with mothers who graduated from high school perform, on average, 47.0 points better on their tests? No. The model includes an interaction, and 47.0 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 10.4), the comparison at zero, and thus the coefficient of 47.0, is essentially meaningless.

Similarly, the coefficient of 0.9 for the "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, as mom_hs actually does equal zero for many of the cases in the data (see Figure 10.1) but still can be somewhat misleading since mom_hs = 0 is at the edge of the data so that this coefficient cannot be interpreted as an average over the general population.

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:

```
kidiq$c_mom_hs <- kidiq$mom_hs - mean(kidiq$mom_hs)
kidiq$c_mom_iq <- kidiq$mom_iq - mean(kidiq$mom_iq)
```

Each main effect now corresponds to a predictive difference with the other input at its average value:

```
                  Median MAD_SD
(Intercept)       87.6    0.9
c_mom_hs           2.9    2.4
c_mom_iq           0.6    0.1
c_mom_hs:c_mom_iq -0.5    0.2
```

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

```
Auxiliary parameter(s):
      Median MAD_SD
sigma 18.0   0.6
```

The residual standard deviation does not change—linear transformation of the predictors does not affect the fit of the model—and the coefficient and standard error of the interaction did not change, but the main effects and the intercept change 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:

```
kidiq$c2_mom_hs <- kidiq$mom_hs - 0.5
kidiq$c2_mom_iq <- kidiq$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 a child with mom_hs = 0, among those children with mom_iq = 100. Similarly, the coefficient of c2_mom_iq corresponds to a comparison under the condition mom_hs = 0.5, which includes no actual data but represents a midpoint of the range.

```
                    Median MAD_SD
(Intercept)         86.8    1.2
c2_mom_hs            2.9    2.3
c2_mom_iq            0.7    0.1
c2_mom_hs:c2_mom_iq -0.5    0.2

Auxiliary parameter(s):
      Median MAD_SD
sigma 18.0   0.6
```

Once again, the residual standard deviation 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 10.4.

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:

```
kidiq$z_mom_hs <- (kidiq$mom_hs - mean(kidiq$mom_hs))/(2*sd(kidiq$mom_hs))
kidiq$z_mom_iq <- (kidiq$mom_iq - mean(kidiq$mom_iq))/(2*sd(kidiq$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):

```
                   Median MAD_SD
(Intercept)         87.6    0.9
z_mom_hs             2.3    2.1
z_mom_iq            17.7    1.8
z_mom_hs:z_mom_iq  -11.9    4.0
```

```
Auxiliary parameter(s):
      Median MAD_SD
sigma 18.0   0.6
```

Why scale by 2 standard deviations?

We divide by 2 standard deviations rather than 1 because this is consistent with what we do with 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 $\sqrt{0.5 * 0.5} = 0.5$, 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, we can get the same inferences for the coefficients other than the intercept by leaving the regression predictors as is and then creating 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, adjusting 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.

12.3 Correlation and "regression to the mean"

Consider a regression with a constant term and one predictor; thus, $y = a + bx +$ error. If both of the variables 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 in absolute value, then 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 component line and the regression line

Some of the confusing aspects of regression can be understood in the simple case of standardized variables. Figure 12.2 shows a simulated-data example of standardized variables with correlation (and thus regression slope) 0.5. Figure 12.2a shows the *principal component line*, which goes closest through the cloud of points, in the sense of minimizing the sum of squared distances between the points and the line. The principal component line in this case is simply $y = x$.

Figure 12.2b shows the *regression line*, which minimizes the sum of the squares of the *vertical* distances between the points and the line—it is the familiar least squares line, $y = \hat{a} + \hat{b}x$, with $\hat{a}, \hat{b}$

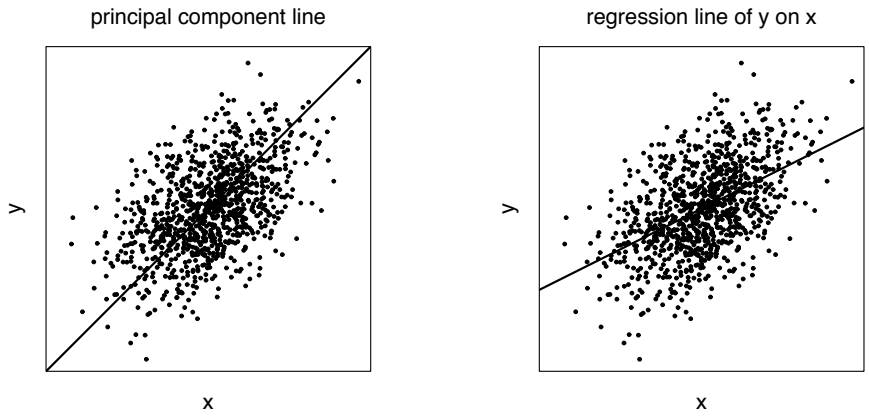

Figure 12.2 *Data simulated from a bivariate normal distribution with correlation 0.5. (a) The principal component line goes closest through the cloud of points. (b) The regression line, which represents the best prediction of y given x, has half the slope of the principal component line.*

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, which is shown in Figure 12.2a. 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 12.2. For example, consider the points at the extreme left of either graph. They all lie above the principal component 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 of y given x.

Regression to the mean

This all connects to our earlier discussion of "regression to the mean" in Section 6.5. When x and y are standardized (that is, placed on a common scale, as in Figure 12.2), the regression line always has slope less than 1. Thus, when x is 1 standard deviation 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 daughters' heights is 0.5, then her daughter's predicted height is 5 inches taller than the average. She 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 variables that are not perfectly correlated. For example, let x_i and y_i be the number of games won by football team i in two successive seasons. They will not be correlated 100%; thus, we 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, we expect their values of y to be closer to the average for all the teams). Similarly, we expect teams with poor records in season 1 to improve on average in season 2, relative to the other teams.

A naive interpretation of regression to the mean is that heights, or football records, or other variable phenomena become more and more "average" over time. As discussed in Section 6.5, 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 observed y_i 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 12.2b.

12.4 Logarithmic transformations

When additivity and linearity are not reasonable assumptions (see Section 11.1), 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 important, 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

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

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.

In Section 3.4, we discussed the connections between logarithmic transformations and exponential and power-law relationships; here we consider these in the context of regression.

Earnings and height example

Example: Earnings and height

We illustrate logarithmic regression by considering models predicting earnings from height.[2] Expression (12.1) shows a linear regression of earnings on height. However, it really makes more sense to model earnings on the logarithmic scale, as long as we exclude 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,

```
logmodel_1 <- stan_glm(log(earn) ~ height, data=earnings, subset=earn>0)
print(logmodel_1)
```

yielding the following estimate:

```
            Median MAD_SD
(Intercept) 5.91   0.38
height      0.06   0.01

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.88   0.02
```

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

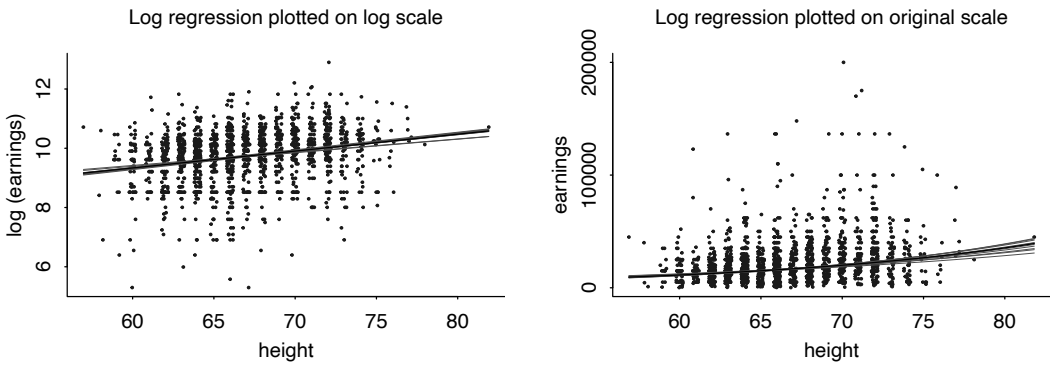

Figure 12.3 *Regression of earnings on log(height), with curves showing uncertainty the model, log(earnings) = $a + b * height$, fit to data with positive earnings. The data and fit are plotted on the logarithmic and original scales. Compare to the linear model, shown in Figure 12.1a. To improve resolution, a data point at earnings of* \$400 000 *has been excluded from the original-scale graph.*

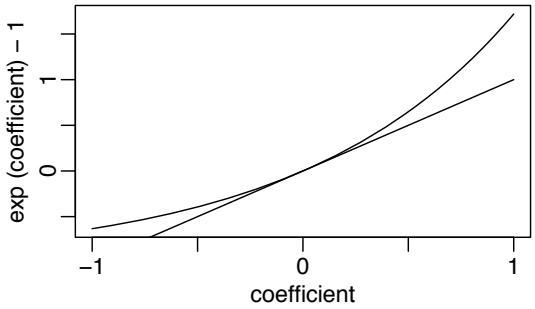

Figure 12.4 *Interpretation of exponentiated coefficients in a logarithmic regression model as relative difference (curved upper line), and the approximation $exp(x) = 1 + x$, which is valid for small coefficients x (straight line).*

Figure 12.3 shows the data and fitted regression on the log and linear scales.

The estimate $\hat{\beta}_1 = 0.06$ implies that a difference of 1 inch in height corresponds to an expected difference of 0.06 in log(earnings), so that earnings are multiplied by exp(0.06). But exp(0.06) $\approx$ 1.06 (more precisely, 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 becomes more nonlinear as the magnitude of the coefficient increases. Figure 12.4 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.

Predictive checking. One way to get a sense of fit is to simulate replicated datasets from the fitted model and compare them to the observed data. We demonstrate for the height and earnings regression.

First we simulate new data:

```
yrep_1 <- posterior_predict(fit_1)
```

The above code returns a matrix in which each row is a replicated dataset from the posterior distribution of the fitted regression of earnings on height. We then plot the density of the observed earnings data, along with 100 draws of the distribution of replicated data:

```
n_sims <- nrow(yrep_1)
```

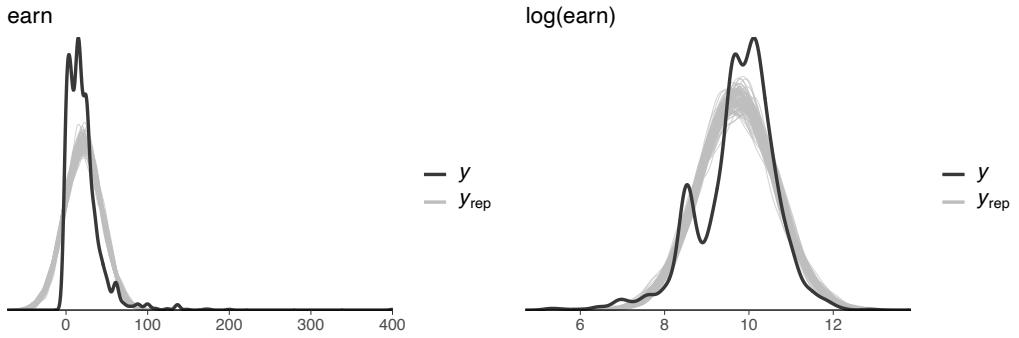

Figure 12.5 *Posterior predictive checks comparing the density plot of earnings data (dark line) to 100 predictive replications (gray lines) of replicated data from fitted models (a) on the original scale, and (b) on the log scale. Both models show some lack of fit. The problem is particularly obvious with the linear-scale regression, as the observed earnings are all positive (with the density function including a small negative tail just as an artifact of the smoothing procedure) and the replicated data include many negative values. The non-smoothed aspects of the observed data arise from discreteness in the survey responses.*

```
subset <- sample(n_sims, 100)
library("bayesplot")
ppc_dens_overlay(earnings$earn, yrep_1[subset,])
```

The result is shown in Figure 12.5a. Unsurprisingly, the fit on the untransformed scale is poor: observed earnings in these data are always positive, while the predictive replications contain many negative values.

We can then do the same predictive checking procedure for the model fit on the log scale, first simulating the predictions:

```
yrep_log_1 <- posterior_predict(logmodel_1)
```

Then we plot 100 simulations along with the observed data:

```
n_sims <- nrow(yrep_log_1)
subset <- sample(n_sims, 100)
ppc_dens_overlay(log(earnings$earn[earnings$earn>0]), yrep_log_1[subset,])
```

The resulting fit on the log scale is not perfect (see Figure 12.5b), which could be of interest, depending on one's goal in fitting the model. The point of this example is that, as a model is altered, we can perform predictive checks to assess different aspects of fit. Here we looked at the marginal distribution of the data, but more generally one can look at other graphical summaries.

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 proportional differences: with a coefficient of 0.05, a difference of 1 in x corresponds to an approximate 5% difference in y, and so forth. Natural log is sometimes written as "ln," but we simply write "log" since this is our default.

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 fit the earnings regression on the $\log_{10}$ scale,

```
logmodel_1a <- stan_glm(log10(earn) ~ height, data=earnings, subset=earn>0)
```

we get,

```
            Median MAD_SD
(Intercept) 2.57   0.16
height      0.02   0.00

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.38   0.01
```

The coefficient of 0.02 tells us that a difference of 1 inch in height corresponds to a difference of 0.02 in $\log_{10}$(earnings), that is, a multiplicative difference of $10^{0.02} = 1.06$ (after fixing roundoff error). 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. A difference of an inch of height corresponds to a difference of 6% 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 understood by differences between the sexes. Do taller people earn more, on average, than shorter people of the same sex?

```
logmodel_2 <- stan_glm(log(earn) ~ height + male, data=earnings, subset=earn>0)

            Median MAD_SD
(Intercept) 7.97   0.51
height      0.02   0.01
male        0.37   0.06

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.87   0.02
```

After adjusting for sex, each inch of height corresponds to an estimated predictive difference of 2%: under this model, two people 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 $\exp(0.37) = 1.45$ times the woman's, that is, 45% more. (We cannot simply convert the 0.37 to 45% because this coefficient is not so close to zero; see Figure 12.4.) This coefficient also is not easily interpretable. Does it mean that "being a man" *causes* one to earn nearly 50% more than a woman? We will explore this sort of troubling question in the causal inference chapters in Part 5 of the book.

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. Another approach would be to consider `sex` as a factor with two named levels, `male` and `female`; see page 198. 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.

Residual standard deviation and R^2. Finally, the regression model has a residual standard deviation σ of 0.87, implying that approximately 68% of log earnings will be within 0.87 of the predicted value. On the original scale, approximately 68% of earnings will be within a factor of $\exp(0.87) = 2.4$ of the prediction. For example, a 70-inch person has predicted earnings of $7.97 + 0.02 * 70 = 9.37$, with a predictive standard deviation of approximately 0.87. Thus, there is an approximate 68% chance that this person has log earnings in the range $[9.37 \pm 0.87] = [8.50, 10.24]$, which corresponds to

earnings in the range $[\exp(8.50), \exp(10.24)]] = [5000, 28\,000]$. This 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 R^2, which is only 0.08, indicating that only 8% of the variance in the log transformed data is explained by the regression model. This low R^2 manifests itself graphically in Figure 12.3, 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:

```
logmodel_3 <- stan_glm(log(earn) ~ height + male + height:male,
  data=earnings, subset=earn>0)
```

which yields,

```
             Median MAD_SD
(Intercept)  8.48   0.66
height       0.02   0.01
male        -0.76   0.94
height:male  0.02   0.01

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.87   0.02
```

That is,

$$\log(\text{earnings}) = 8.48 + 0.02 * \text{height} - 0.76 * \text{male} + 0.02 * \text{height} * \text{male}. \tag{12.2}$$

We shall try to 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`, 0.02, 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 2% for women, with some uncertainty as indicated by the standard error of 0.01.
- 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. If you want to interpret it, you can move to a more relevant value for `height`; as discussed in Section 12.2, it makes sense to use a centered parameterization.
- The coefficient for `height:male` is the difference in slopes of the lines predicting log earnings on height, comparing men to women. Thus, a difference of an inch of height corresponds to 2% more of a difference in earnings among men than among women, and the estimated predictive difference per inch of height among men is 2% + 2% = 4%.

The interaction coefficient has a large standard error, which tells us that the point estimate is uncertain and could change in sign and magnitude if additional data were fed into the analysis.

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

```
earnings$z_height <- (earnings$height - mean(earnings$height))/sd(earnings$height)
```

The mean and standard deviation of heights in these data are 66.6 inches and 3.8 inches, respectively. Fitting the model to `z_height`, `male`, and their interaction yields,

```
             Median MAD_SD
(Intercept)  9.55   0.04
z_height     0.06   0.04
```

```
male          0.35   0.06
z_height:male 0.08   0.06

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.87   0.02
```

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 66.6-inch-tall woman is predicted to have log earnings of 9.55, or earnings of exp(9.55) = 14 000.
- 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 6% for women (but with a standard error indicating much uncertainty in this coefficient).
- The coefficient for `male` is the predicted difference in log earnings between women and men, if `z_height` equals 0. Thus, a 66.6-inch-tall man is predicted to have log earnings that are 0.35 higher than that of a 66.6-inch-tall woman. This corresponds to a ratio of exp(0.35) = 1.42, so the man is predicted to have 42% 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, comparing two men who differ by 3.8 inches in height, the model predicts a difference of 0.06 + 0.08 = 0.14 in log earnings, thus a ratio of exp(0.14) = 1.15, a difference of 15%.

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 challenge 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 192. This seems to be a reasonable comparison.

To interpret the coefficient for `male`, we would say that comparing two adults of the same height but different sex, the man will be expected to earn 45% more on average. But how clear is it for us to interpret this 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 or relevant 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 discuss causal interpretations in Chapters 18–21.

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 difference in y per proportional difference in x. For example:

```
earnings$log_height <- log(earnings$height)
logmodel_5 <- stan_glm(log(earn) ~ log_height + male, data=earnings, subset=earn>0)
```

yields,

```
             Median MAD_SD
(Intercept) 2.76   2.19
log_height  1.62   0.53
male        0.37   0.06

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.87   0.01
```

For each 1% difference in height, the predicted difference in earnings is approximately 1.62%. To be precise, when comparing two people who differ in height by a factor of 1.01, this corresponds to a difference in log(height) of $\log(1.01) = 0.01$, which corresponds to a difference of 0.0162 in the predicted value of $\log y$, so the expected value of y is larger by a factor of $\exp(0.0162) = 1.0163$. 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 12.11 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 as an improvement in fit too.

12.5 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 on 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 (12.1), 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 and those earning $80 000. On the square root scale, however, the predicted differences between the $0 earnings and $10 000 earnings groups are the same as comparisons between $10 000 and $40 000 or between $40 000 and $90 000, in each case stepping up by 100 in square root of earnings. See Chapter 17 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 within models whose coefficients we want to understand.

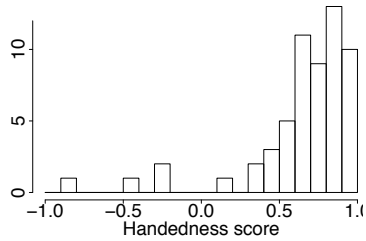

Figure 12.6 *Histogram of handedness scores of a sample of students. Scores range from* −1 *(completely left-handed) to* +1 *(completely right-handed) and are based on the responses to 10 questions such as "Which hand do you write with?" and "Which hand do you use to hold a spoon?" The continuous range of responses shows the limitations of treating handedness as a dichotomous variable.*

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: first model the probability that earnings exceed zero (for example, using a logistic regression; see Chapter 13); then fit a linear regression, conditional on earnings being positive, which is what we did in the example above. One could also model total income, but economists are often interested in modeling earnings alone, excluding so-called unearned income.

In any case, plots and simulations 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 Section 15.8.

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 between −$100 and $100, 1 for assets between $100 and $1000, 2 for assets between $1000 and $10 000, −1 for assets between −$100 and −$1000, 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 10-question handedness scale that gives an essentially continuous scale from −1 to 1 (see Figure 12.6).

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 as 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). 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 10.6.

Using discrete rather than continuous predictors

In some cases, however, it is convenient to discretize a continuous variable if a simple parametric relation does not seem appropriate. For example, in modeling political preferences, it can make sense to include age with four indicator variables: 18–29, 30–44, 45–64, and 65+, to allow for different

sorts of generational patterns. This kind of discretization is convenient, since, conditional on the discretization, the model remains linear. We briefly mention more elaborate nonlinear models for continuous predictors in Section 22.7.

Example: Children's IQ tests

We demonstrate inference with discrete predictors using an example from Chapter 10 of 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,

```
                      Median MAD_SD
(Intercept)           82.0    2.2
as.factor(mom_work)2   3.8    3.0
as.factor(mom_work)3  11.4    3.5
as.factor(mom_work)4   5.1    2.7

Auxiliary parameter(s):
      Median MAD_SD
sigma 20.2   0.7
```

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 + 11.4. 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 unless these variables are included in the model.

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, as discussed in Section 10.4. 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`.

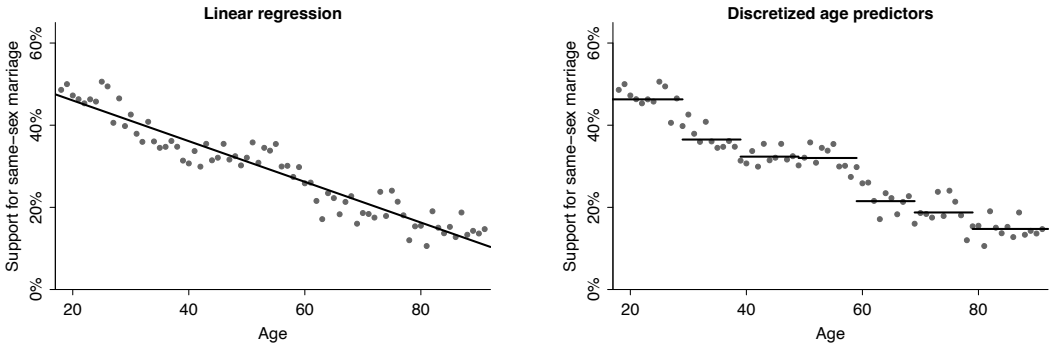

Figure 12.7 *Support for same-sex marriage as a function of age, from a national survey taken in 2004. Fits are shown from two linear regression: (a) using age as a predictor, (b) using indicators for age, discretized into categories.*

Including these variables as regression predictors allows for different means for the populations corresponding to each of the categories delineated by the variable.

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` could have the levels `male` and `female`.

Figure 12.7 demonstrates with a simple example showing support for same-sex marriage as a function of age (and with the few respondents reporting ages greater than 90 all assigned the age of 91 for the purpose of this analysis). Here is the result of a linear regression using an indicator for each decade of age, with age under 30 as the reference category:

```
                             Median MAD_SD
(Intercept)                   0.46   0.01
factor(age_discrete)(29,39]  -0.10   0.01
factor(age_discrete)(39,49]  -0.14   0.01
factor(age_discrete)(49,59]  -0.14   0.01
factor(age_discrete)(59,69]  -0.25   0.01
factor(age_discrete)(69,79]  -0.28   0.01
factor(age_discrete)(79,100] -0.32   0.01

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.03   0.00
```

Figure 12.7a shows the fitted linear regression on age, and Figure 12.7b shows the fit from the linear regression using age indicators: the first bar is at $y = 0.46$, the second is at $0.46 - 0.10$, the third is at $0.46 - 0.14$, and so on. Neither of the two fits in Figure 12.7 is perfect; indeed Figure 12.7b gives a somewhat misleading picture, with the eye being drawn too strongly to the horizontal lines. One reason why we show both graphs is to give two perspectives on the data. For example, the dots in Figure 12.7a show a steady downward trend between the ages of 25 and 40, but in Figure 12.7b, that pattern is obscured by the fitted lines.

Indicator variables, identifiability, and the baseline condition

As discussed in Section 10.7, a regression model is nonidentifiable if its predictors are collinear, that is, if there is a linear combination of them that equals 0 for all the data. This can arise with indicator variables. If a factor takes on J levels, 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 indicators 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. As discussed in Section 10.4, the default in R is to set the alphabetically 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. An option that we often prefer is to embed the varying coefficients in a multilevel model, but this goes beyond the scope of this book.

12.6 Building and comparing 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 or transformed 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 regularized regression (as we discuss in our follow-up book on advanced regression and multilevel models) 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 18–21.

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, and that can result in more stable predictions when coefficients are estimated using maximum likelihood or least squares.
3. For inputs that have large effects, consider including their interactions as well.
4. Use standard errors to get a sense of uncertainties in parameter estimates. Recognize that if new data are added to the model, the estimate can change.
5. Make decisions about including or excluding predictors based on a combination of contextual understanding (prior knowledge), data, and the uses to which the regression will be put:
   (a) If the coefficient of a predictor is estimated precisely (that is, if it has a small standard error), it generally makes sense to keep it in the model as it should improve predictions.
   (b) If the standard error of a coefficient is large and there seems to be no good substantive reason for the variable to be included, it can make sense to remove it, as this can allow the other coefficients in the model to be estimated more stably and can even reduce prediction errors.
   (c) If a predictor is important for the problem at hand (for example, indicators for groups that we are interested in comparing or adjusting for), then we generally recommend keeping it in, even

if the estimate has a large standard error and is not "statistically significant." In such settings one must acknowledge the resulting uncertainty and perhaps try to reduce it, either by gathering more data points for the regression or by adding a Bayesian prior (see Section 9.5).

(d) If a coefficient seems not to make sense (for example, a negative coefficient for years of education in an income regression), try to understand how this could happen. If the standard error is large, the estimate could be explainable from random variation. If the standard error is small, it can make sense to put more effort into understanding the coefficient. In the education and income example, for example, the data could be coming from a subpopulation in which the more educated people are younger and have been in their jobs for a shorter period of time and have lower average incomes.

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 once we have seen it 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.

It is important to record and describe the choices made in modeling, as these choices represent degrees of freedom that, if not understood, can lead to a garden of forking paths and overconfident conclusions. Model performance estimates such as LOO log score can alleviate the problem if there are not too many models.