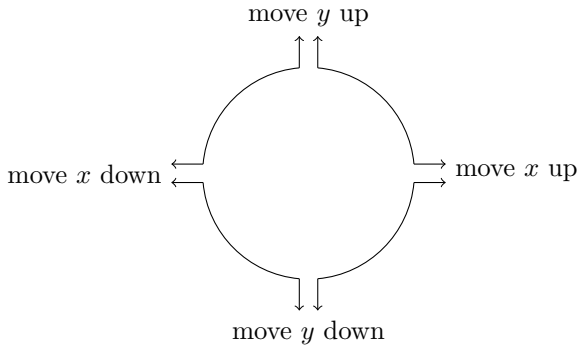

ladder of re-expression

power	transformation
⋮	⋮
3	$x \mapsto x^3$
2	$x \mapsto x^2$
1	$x \mapsto x$
$\frac{1}{2}$	$x \mapsto \sqrt{x}$
0	$x \mapsto \log(x)$
-1	$x \mapsto \frac{1}{x}$
-2	$x \mapsto \frac{1}{x^2}$
⋮	⋮

Figure 6.12. Bulge rules and ladder of re-expression.

In practice, all three of these issues are intertwined. A transformation that improves the fit, for example, may or may not have a good theoretical interpretation. Similarly, a transformation performed to achieve **homoskedasticity** (equal variance; the opposite is called **heteroskedasticity**) may result in a fit that does not match the overall shape of the data very well. Despite these potential problems, there are many situations where a relatively simple transformation is all that is needed to greatly improve the model.

6.5.1. The Ladder of Re-expression

In the 1970s, Mosteller and Tukey introduced what they called the **ladder of re-expression** and **bulge rules** [**Tuk77, MT77**] that can be used to suggest an appropriate transformation to improve the fit when the relationship between two variables (x and y in our examples) is monotonic and has a single bend. Their idea was to apply a power transformation to x or y or both – that is, to work with x^a and y^b for an appropriate choice of a and b. Tukey called this ordered list of transformations the ladder of re-expression. The identity transformation has power 1. The logarithmic transformation is a special case and is included in the list associated with a power of 0. The direction of the required transformation can be obtained from Figure 6.12, which shows four bulge types, represented by the curves in each of the four quadrants. A bulge can potentially be straightened by applying a transformation to one or both variables, moving up or down the ladder as indicated by the arrows. More severe bulges require moving farther up or down the ladder. A curve bulging in the same direction as the one in the first quadrant of Figure 6.12, for example, might be straightened by moving up the ladder of transformations for x or y (or both), while a curve like the one in the second quadrant, might be straightened by moving up the ladder for y or down the ladder for x.

This method focuses primarily on transformations designed to improve the overall fit. The resulting models may or may not have a natural interpretation. These transformations also affect the shape of the distributions of the explanatory

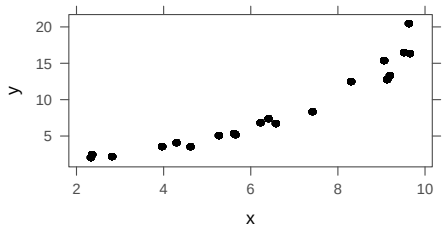

Figure 6.13. A scatterplot illustrating a non-linear relationship between x and y.

and response variables and, more importantly, of the residuals from the linear model (see Exercise 6.18). When several transformations lead to reasonable linear fits, these other factors may lead us to prefer one over another.

Example 6.5.1.

Q. The scatterplot in Figure 6.13 shows a curved relationship between x and y. What transformations of x and y improve the linear fit?

A. This type of bulge appears in quadrant IV of Figure 6.12, so we can hope to improve the fit by moving up the ladder for x or down the ladder for y. As we see in Figure 6.14, the fit generally improves as we move down and to the right – but not too far, lest we over-correct. A log-transformation of the response ($a = 1$, $b = 0$) seems to be especially good in this case. Not only is the resulting relationship quite linear, but the residuals appear to have a better distribution as well. ◁

Example 6.5.2. Some physics students conducted an experiment in which they dropped steel balls from various heights and recorded the time until the ball hit the floor. We begin by fitting a linear model to this data.

balldrop

```
> ball.model <- lm(time~height,balldrop)
> summary(ball.model)
< 8 lines removed >
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19024    0.00430    44.2   <2e-16
height       0.25184    0.00552    45.7   <2e-16

Residual standard error: 0.0101 on 28 degrees of freedom
Multiple R-squared: 0.987,        Adjusted R-squared: 0.986
F-statistic: 2.08e+03 on 1 and 28 DF,  p-value: <2e-16

> ball.plot <- xyplot(time~height,balldrop,type=c('p','r'))
> ball.residplot <- xplot(ball.model,w=1)
```

At first glance, the large value of r^2 and the reasonably good fit in the scatterplot might leave us satisfied that we have found a good model. But a look at the residual plot (Figure 6.15) reveals a clear curvilinear pattern in this data. A knowledgeable physics student knows that (ignoring air resistance) the time should

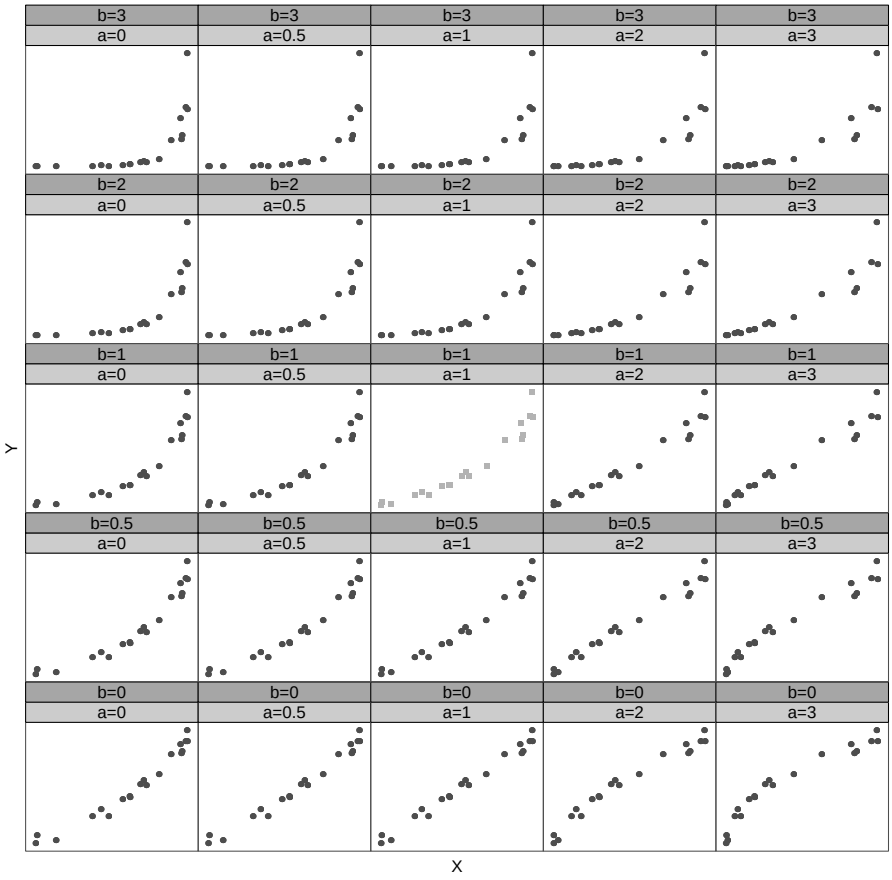

Figure 6.14. Using the ladder of re-expression to find a better fit.

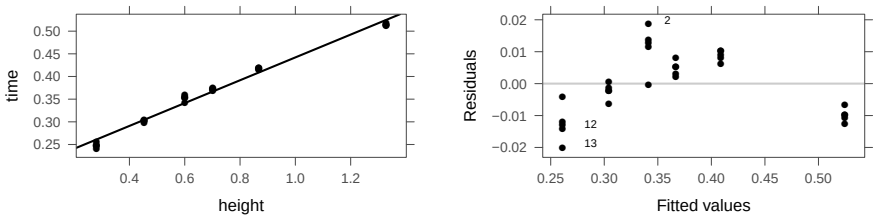

Figure 6.15. A scatterplot and a residual plot for the `balldrop` data set.

be proportional to the *square root* of the height. This transformation agrees with Tukey's ladder of re-expression, which suggests moving down the ladder for `height` or up the ladder for `time`.

balldrop-trans

```
> ball.modelT <- lm(time ~ sqrt(height),balldrop)
> summary(ball.modelT)
```