Chapter 8

Fitting regression models

Most of this book is devoted to examples and tools for the practical use and understanding of regression models, starting with linear regression with a single predictor and moving to multiple predictors, nonlinear models, and applications in prediction and causal inference. In this chapter we lay out some of the mathematical structure of inference for regression models and some algebra to help understand estimation for linear regression. We also explain the rationale for the use of the Bayesian fitting routine `stan_glm` and its connection to classical linear regression. This chapter thus provides background and motivation for the mathematical and computational tools used in the rest of the book.

Note: We won't be covering Bayesian Inference in 201b! As such, I've blanked some of the chapter sections throughout

8.1 Least squares, maximum likelihood, and ~~Bayesian inference~~

We now step back and consider *inference*: the steps of estimating the regression model and assessing uncertainty in the fit. We start with *least squares*, which is the most direct approach to estimation, based on finding the values of the coefficients a and b that best fit the data. We then discuss *maximum likelihood*, a more general framework that includes least squares as a special case and to which we return in later chapters when we get to logistic regression and generalized linear models. Then we proceed to *Bayesian inference*, an even more general approach that allows the probabilistic expression of prior information and posterior uncertainty.

Least squares

In the classical linear regression model, $y_i = a + bx_i + \epsilon_i$, the coefficients a and b are estimated so as to minimize the errors ϵ_i. If the number of data points n is greater than 2, it is not generally possible to find a line that gives a perfect fit (that would be $y_i = a + bx_i$, with no error, for all data points $i = 1, \ldots, n$), and the usual estimation goal is to choose the estimate $(\hat{a}, \hat{b})$ that minimizes the sum of the squares of the residuals,

$$r_i = y_i - (\hat{a} + \hat{b}x_i).$$

We distinguish between the *residuals* $r_i = y_i - (\hat{a} + \hat{b}x_i)$ and the *errors* $\epsilon_i = y_i - (a + bx_i)$. The model is written in terms of the errors, but it is the residuals that we can work with: we cannot calculate the errors as to do so would require knowing a and b.

The residual sum of squares is

$$\text{RSS} = \sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b}x_i))^2. \tag{8.1}$$

The $(\hat{a}, \hat{b})$ that minimizes RSS is called the least squares or ordinary least squares or OLS estimate and can be written in matrix notation as,

$$\hat{\beta} = (X^t X)^{-1} X^t y, \tag{8.2}$$

where $\beta = (a, b)$ is the vector of coefficients and $X = (\mathbf{1}, x)$ is the matrix of predictors in the regression. In this notation, $\mathbf{1}$ represents a column of ones—the constant term in the regression—and must be included because we are fitting a model with an intercept as well as a slope. We show more general notation for linear regression with multiple predictors in Figure 10.8.

Expression (8.2) applies to least squares regression with any number of predictors. In the case of regression with just one predictor, we can write the solution as,

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{8.3}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}. \tag{8.4}$$

We can then can write the least squares line as,

$$y_i = \bar{y} + \hat{b}\,(x_i - \bar{x}) + r_i;$$

thus, the line goes through the mean of the data, $(\bar{x}, \bar{y})$, as illustrated in Figure 6.3.

Formula (8.2) and the special case (8.3)–(8.4) can be directly derived using calculus as the solution to the problem of minimizing the residual sum of squares (8.1). In practice, these computations are done using efficient matrix solution algorithms in R or other software.

Estimation of residual standard deviation σ

In the regression model, the errors ϵ_i come from a distribution with mean 0 and standard deviation σ: the mean is zero by definition (any nonzero mean is absorbed into the intercept, a), and the standard deviation of the errors can be estimated from the data. A natural way to estimate σ would be to simply take the standard deviation of the residuals, $\sqrt{\frac{1}{n}\sum_{i=1}^{n} r_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2}$, but this would slightly underestimate σ because of *overfitting*, as the coefficients $\hat{a}$ and $\hat{b}$ have been set based on the data to minimize the sum of squared residuals. The standard correction for this overfitting is to replace n by $n - 2$ in the denominator (with the subtraction of 2 coming from the estimation of two coefficients in the model, the intercept and the slope); thus,

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2}. \tag{8.5}$$

When $n = 1$ or 2 this expression is meaningless, which makes sense: with only two data points you can fit a line exactly and so there is no way of estimating error from the data alone.

More generally, in a regression with k predictors (that is, $y = X\beta + \epsilon$, with an $n \times k$ predictor matrix X), expression (8.5) becomes $\hat{\sigma} = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - (X_i\hat{\beta}))^2}$, with $n - k$ in the denominator rather than n, adjusting for the k coefficients fit by least squares.

Maximum likelihood

If the errors from the linear model are independent and normally distributed, so that $y_i \sim \text{normal}(a + bx_i, \sigma)$ for each i, then the least squares estimate of (a, b) is also the maximum likelihood estimate. The *likelihood function* in a regression model is defined as the probability density of the data given the parameters and predictors; thus, in this example,

$$p(y|a, b, \sigma, X) = \prod_{i=1}^{n} \text{normal}(y_i|a + bx_i, \sigma), \tag{8.6}$$

where $\text{normal}(\cdot|\cdot, \cdot)$ is the normal probability density function,

$$\text{normal}(y \mid m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-m}{\sigma}\right)^2\right). \tag{8.7}$$

A careful study of (8.6) reveals that maximizing the likelihood requires minimizing the sum of squared residuals; hence the least squares estimate $\hat{\beta} = (\hat{a}, \hat{b})$ can be viewed as a maximum likelihood estimate under the normal model.

There is a small twist in fitting regression models, in that the maximum likelihood estimate of σ is $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2}$, without the $\frac{1}{n-2}$ adjustment given in (8.5).

Where do the standard errors come from? Using the likelihood surface to assess uncertainty in the parameter estimates

In maximum likelihood estimation, the likelihood function can be viewed as a hill with its peak at the maximum likelihood estimate.

Figure 8.1a displays the likelihood for a simple example as a function of the coefficients a and b. Strictly speaking, this model has three parameters—a, b, and σ—but for simplicity we display the likelihood of a and b conditional on the estimated $\hat{\sigma}$.

Figure 8.1b shows the maximum likelihood estimate $(\hat{a}, \hat{b}) = (46.2, 3.1)$. This is the value of the parameters where the likelihood function—the hill in Figure 8.1a—has its peak. Figure 8.1b also includes uncertainty bars showing ±1 standard error for each parameter. For example, the data are consistent with a being roughly in the range 46.2 ± 1.6 and with b being in the range 3.1 ± 0.7.

The likelihood function does not just have a maximum and a range; it also has a correlation. The area with highest likelihood surrounding the peak can be represented by an ellipse as is shown in Figure 8.1c. The shape of the uncertainty ellipse tells us something about the information in the data and model about the two parameters jointly. In this case the correlation is negative.

To understand this inferential correlation, see the scatterplot of the data from Figure 7.2 which we have reproduced in Figure 8.2a: the regression line goes through the cloud of points, most of which have positive values for x. Figure 8.2b shows a range of lines that are consistent with the data, with

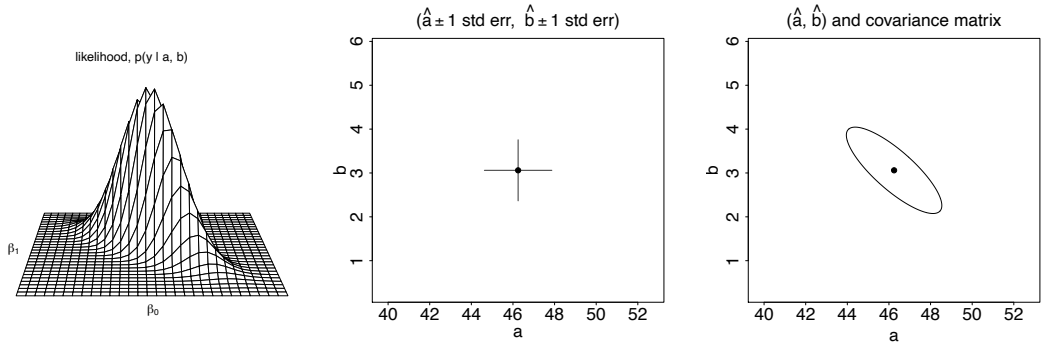

Figure 8.1 *(a) Likelihood function for the parameters a and b in the linear regression $y = a + bx +$ error, of election outcomes, y_i, on economic growth, x_i. (b) Mode of the likelihood function (that is, the maximum likelihood estimate $(\hat{a}, \hat{b})$) with ± 1 standard error bars shown for each parameter. (c) Mode of the likelihood function with an ellipse summarizing the inverse-second-derivative-matrix of the log likelihood at the mode.*

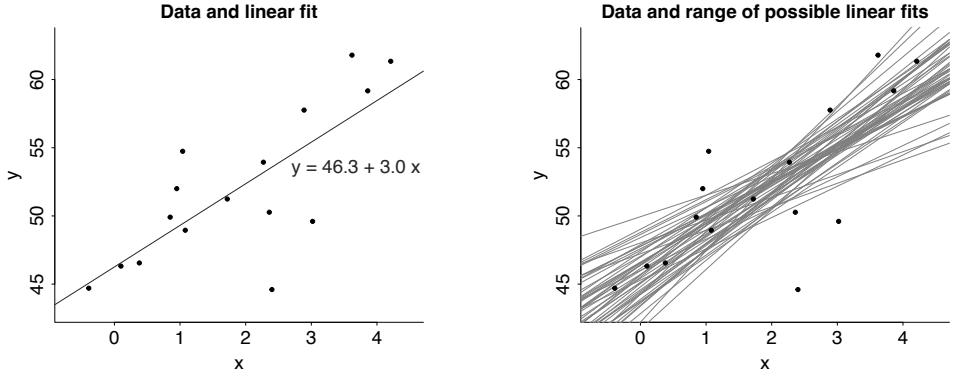

Figure 8.2 *(a) Election data with the linear fit, $y = 46.3 + 3.0x$, repeated from Figure 7.2b. (b) Several lines that are are roughly consistent with the data. Where the slope is higher, the intercept (the value of the line when $x = 0$) is lower; hence there is a negative correlation between a and b in the likelihood.*

the lines representing 50 draws from the Bayesian posterior distribution (see below). Lines of higher slope (for which b is higher) intersect the y-axis at lower values (and thus have lower values of a), and vice versa, hence the negative correlation in Figure 8.1c.

8.2 Influence of individual points in a fitted regression

From expressions (8.3) and (8.4), we can see that the least squares estimated regression coefficients $\hat{a}$ and $\hat{b}$ are linear functions of the data, y. We can use these linear expressions to understand the *influence* of each data point by looking at how much a change in each y_i would change $\hat{b}$. We could also work out the influence on the intercept—the predicted value when $x = 0$—or any other prediction under the model, but typically it is the slope that is most of interest.

From equation (8.3), we see that an increase of 1 in y_i corresponds to a change in $\hat{b}$ that is proportional to $(x_i - \bar{x})$:

- If $x_i = \bar{x}$, the influence of point i on the regression slope is 0. This makes sense: taking a point in the center and moving it up or down will affect the height of the fitted line but not its slope.
- If $x_i > \bar{x}$, the influence of point i is positive, with greater influence the further x_i is from the mean.
- If $x_i < \bar{x}$, the influence of point i is negative, with greater absolute influence the further x_i is from the mean.

One way to understand influence is to consider the fitted regression line as a rod attached to the data by rubber bands; then imagine how the position and orientation of the rod changes as individual data points are moved up and down. Figure 8.3 illustrates.

Influence can also be computed for multiple regression, using the matrix expression (equation

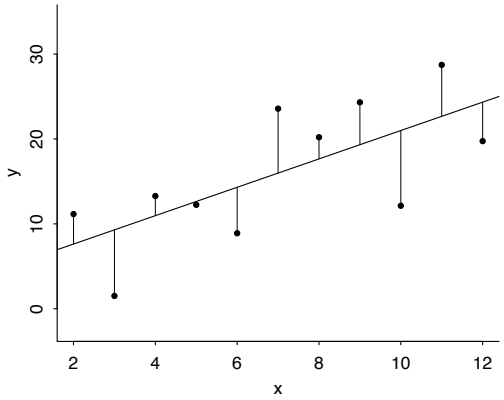

Figure 8.3 *Understanding the influence of individual data points on the fitted regression line. Picture the vertical lines as rubber bands connecting each data point to the least squares line. Take one of the points on the left side of the graph and move it up, and the slope of the line will* decrease*. Take one of the points on the right side and move it up, and the slope will* increase*. Moving the point in the center of the graph up or down will not change the slope of the fitted line.*

(8.2)), which reveals how the estimated vector of regression coefficients $\hat{\beta}$ is a linear function of the data vector y, and for generalized linear models, by re-fitting the regression after altering data points one at a time.

8.3 Least squares slope as a weighted average of slopes of pairs

In Section 7.3, we discussed that, when a regression $y = a + bx +$ error is fit with just an indicator variable (that is, where x just takes on the values 0 and 1), the least squares estimate of its coefficient b is simply the average difference in the outcome between the two groups; that is, $\bar{y}_1 - \bar{y}_0$.

There is a similar identity when the predictor x is continuous; in this case, we can express the estimated slope $\hat{b}$ from (8.3) as a weighted average of slopes.

The basic idea goes as follows. With n data points (x, y) there are n^2 pairs (including the possibility of taking the same data point twice). For each pair i, j we can compute the slope of the line connecting them:

$$\text{slope}_{ij} = \frac{y_j - y_i}{x_j - x_i}.$$

This expression is not defined when the two points have the same value of the predictor—that is, when $x_i = x_j$—but don't worry about that now; it will turn out that these cases drop out of our equation.

We would like to define the best-fit regression slope as an average of the individual slopes, but it makes sense to use a weighted average, in which slope_{ij} counts more if the two points are further apart in x. We might, then, weight each slope by the difference between the two values, $|x_j - x_i|$. For mathematical reasons that we do not discuss here but which relate to the use of the normal distribution for errors (which is in turn motivated by the Central Limit Theorem, as discussed in Section 3.5), it makes sense to weight each slope by the squared separation, $(x_j - x_i)^2$.

We can then compute the weighted average:

$$\begin{aligned}\text{weighted average of slopes} &= \frac{\sum_{i,j}(x_j - x_i)^2 \frac{y_j - y_i}{x_j - x_i}}{\sum_{i,j}(x_j - x_i)^2} \\ &= \frac{\sum_{i,j}(x_j - x_i)(y_j - y_i)}{\sum_{i,j}(x_j - x_i)^2}. \qquad (8.8)\end{aligned}$$

If you collect the terms carefully, you can show that this expression is the same as $\hat{b}$ in (8.3), so we can interpret the estimated coefficient $\hat{b}$ as the weighted average slope in the data, and we can interpret the underlying parameter b as the weighted average slope in the population.