$\mathbf{E}[Y_0 \mid X_0]$ minimizes the mean squared estimation error $\mathbf{E}\big[\big(Y_0 - g(X_0)\big)^2\big]$, over all functions g. Under our assumptions, $\mathbf{E}[Y_0 \mid X_0] = \theta_0 + \theta_1 X_0$. Thus, the true parameters θ_0 and θ_1 minimize

$$\mathbf{E}\big[(Y_0 - \theta_0' - \theta_1' X_0)^2\big],$$

over all θ_0' and θ_1'. By the weak law of large numbers, this expression is the limit as $n \to \infty$ of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_0' - \theta_1' X_i)^2.$$

This indicates that we will obtain a good approximation of the minimizers of $\mathbf{E}\big[(Y_0 - \theta_0' - \theta_1' X_0)^2\big]$ (the true parameters), by minimizing the above expression (with X_i and Y_i replaced by their observed values x_i and y_i, respectively). But minimizing this expression is the same as minimizing the sum of the squared residuals.

Bayesian Linear Regression†

Linear models and regression are not exclusively tied to classical inference methods. They can also be studied within a Bayesian framework, as we now explain. In particular, we may model $x_1, \ldots, x_n$ as given numbers, and $y_1, \ldots, y_n$ as the observed values of a vector $Y = (Y_1, \ldots, Y_n)$ of random variables that obey a linear relation

$$Y_i = \Theta_0 + \Theta_1 x_i + W_i.$$

Here, $\Theta = (\Theta_0, \Theta_1)$ is the parameter to be estimated, and $W_1, \ldots, W_n$ are i.i.d. random variables with mean zero and known variance σ^2. Consistent with the Bayesian philosophy, we model Θ_0 and Θ_1 as random variables. We assume that $\Theta_0, \Theta_1, W_1, \ldots, W_n$ are independent, and that Θ_0, Θ_1 have mean zero and variances σ_0^2, σ_1^2, respectively.

We may now derive a Bayesian estimator based on the MAP approach and the assumption that Θ_0, Θ_1, and $W_1, \ldots, W_n$ are normal random variables. We maximize over θ_0, θ_1 the posterior PDF $f_{\Theta|Y}(\theta_0, \theta_1 \mid y_1, \ldots, y_n)$. By Bayes' rule, the posterior PDF is‡

$$f_\Theta(\theta_0, \theta_1) f_{Y|\Theta}(y_1, \ldots, y_n \mid \theta_0, \theta_1),$$

divided by a positive normalizing constant that does not depend on (θ_0, θ_1). Under our normality assumptions, this expression can be written as

$$c \cdot \exp\left\{-\frac{\theta_0^2}{2\sigma_0^2}\right\} \cdot \exp\left\{-\frac{\theta_1^2}{2\sigma_1^2}\right\} \cdot \prod_{i=1}^{n} \exp\left\{-\frac{(y_i - \theta_0 - x_i\theta_i)^2}{2\sigma^2}\right\},$$

† This subsection can be skipped without loss of continuity.

‡ Note that in this paragraph, we use conditional probability notation since we are dealing with a Bayesian framework.

where c is a normalizing constant that does not depend on (θ_0, θ_1). Equivalently, we minimize over θ_0 and θ_1 the expression

$$\frac{\theta_0^2}{2\sigma_0^2} + \frac{\theta_1^2}{2\sigma_1^2} + \sum_{i=1}^{n} \frac{(y_i - \theta_0 - x_i\theta_1)^2}{2\sigma^2}.$$

Note the similarity with the expression $\sum_{i=1}^{n}(y_i-\theta_0-x_i\theta_1)^2$, which is minimized in the earlier classical linear regression formulation. (The two minimizations would be identical if σ_0 and σ_1 were so large that the terms $\theta_0^2/2\sigma_0^2$ and $\theta_1^2/2\sigma_1^2$ could be neglected.) The minimization is carried out by setting to zero the partial derivatives with respect to θ_0 and θ_1. After some algebra, we obtain the following solution.

Bayesian Linear Regression

- **Model:**
  (a) We assume a linear relation $Y_i = \Theta_0 + \Theta_1 x_i + W_i$.
  (b) The x_i are modeled as known constants.
  (c) The random variables $\Theta_0, \Theta_1, W_1, \ldots, W_n$ are normal and independent.
  (d) The random variables Θ_0 and Θ_1 have mean zero and variances σ_0^2, σ_1^2, respectively.
  (e) The random variables W_i have mean zero and variance σ^2.
- **Estimation Formulas:**
  Given the data pairs (x_i, y_i), the MAP estimates of Θ_0 and Θ_1 are

$$\hat{\theta}_1 = \frac{\sigma_1^2}{\sigma^2 + \sigma_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sum_{i=1}^{n}(x_i - \overline{x})\big(y_i - \overline{y}\big),$$

$$\hat{\theta}_0 = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\,(\overline{y} - \hat{\theta}_1\overline{x}),$$

  where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

We make a few remarks:

(a) If σ^2 is very large compared to σ_0^2 and σ_1^2, we obtain $\hat{\theta}_0 \approx 0$ and $\hat{\theta}_1 \approx 0$. What is happening here is that the observations are too noisy and are essentially ignored, so that the estimates become the same as the prior means, which we assumed to be zero.

(b) If we let the prior variances σ_0^2 and σ_1^2 increase to infinity, we are indicating the absence of any useful prior information on Θ_0 and Θ_1. In this case, the MAP estimates become independent of σ^2, and they agree with the classical linear regression formulas that we derived earlier.

(c) Suppose, for simplicity, that $\overline{x} = 0$. When estimating Θ_1, the values y_i of the observations Y_i are weighted in proportion to the associated values x_i. This is intuitive: when x_i is large, the contribution of $\Theta_1 x_i$ to Y_i is relatively large, and therefore Y_i contains useful information on Θ_1. Conversely, if x_i is zero, the observation Y_i is independent of Θ_1 and can be ignored.

(d) The estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ are linear functions of the y_i, but not of the x_i. Recall, however, that the x_i are treated as exogenous, non-random quantities, whereas the y_i are observed values of the random variables Y_i. Thus the MAP estimators $\hat{\Theta}_0$, $\hat{\Theta}_1$ are linear estimators, in the sense defined in Section 8.4. It follows, in view of our normality assumptions, that the estimators are also Bayesian linear LMS estimators as well as LMS estimators (cf. the discussion near the end of Section 8.4).

Multiple Linear Regression

Our discussion of linear regression so far involved a single **explanatory variable**, namely x, a special case known as **simple regression**. The objective was to build a model that explains the observed values y_i on the basis of the values x_i. Many phenomena, however, involve multiple underlying or explanatory variables. (For example, we may consider a model that tries to explain annual income as a function of both age and years of education.) Models of this type are called **multiple regression** models.

For instance, suppose that our data consist of triples of the form (x_i, y_i, z_i) and that we wish to estimate the parameters θ_j of a model of the form

$$y \approx \theta_0 + \theta_1 x + \theta_2 z.$$

As an example, y_i may be the income, x_i the age, and z_i the years of education of the ith person in a random sample. We then seek to minimize the sum of the squared residuals

$$\sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i - \theta_2 z_i)^2,$$

over all θ_0, θ_1, and θ_2. More generally, there is no limit on the number of explanatory variables to be employed. The calculation of the regression estimates