Approach 1: Minimizing Loss

1. Simple Linear Regression

Model Structure

Simple linear regression models the target variable, y, as a linear function of just one predictor variable, x, plus an error term, ϵ. We can write the entire model for the n^{th} observation as

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n.$$

Fitting the model then consists of estimating two parameters: β_0 and β_1. We call our estimates of these parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. Once we've made these estimates, we can form our prediction for any given x_n with

$$\hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n.$$

One way to find these estimates is by minimizing a loss function. Typically, this loss function is the *residual sum of squares* (RSS). The RSS is calculated with

$$\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2.$$

We divide the sum of squared errors by 2 in order to simplify the math, as shown below. Note that doing this does not affect our estimates because it does not affect which $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the RSS.

Parameter Estimation

Having chosen a loss function, we are ready to derive our estimates. First, let's rewrite the RSS in terms of the estimates:

$$\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{2}\sum_{n=1}^{N}\left(y_n - \left(\hat{\beta}_0 + \hat{\beta}_1 x_n\right)\right)^2.$$

To find the intercept estimate, start by taking the derivative of the RSS with respect to $\hat{\beta}_0$:

$$\begin{aligned}\frac{\partial \mathcal{L}(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} &= -\sum_{n=1}^{N}\left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right) \\ &= -N(\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}),\end{aligned}$$

where $\bar{y}$ and $\bar{x}$ are the sample means. Then set that derivative equal to 0 and solve for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

This gives our intercept estimate, β_0, in terms of the slope estimate, β_1. To find the slope estimate, again start by taking the derivative of the RSS:

$$\frac{\partial \mathcal{L}(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = -\sum_{n=1}^{N}\left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right) x_n.$$

Setting this equal to 0 and substituting for $\hat{\beta}_0$, we get

$$\begin{aligned}\sum_{n=1}^{N}\left(y_n - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_n\right) x_n &= 0 \\ \hat{\beta}_1 \sum_{n=1}^{N}(x_n - \bar{x}) x_n &= \sum_{n=1}^{N}(y_n - \bar{y})\, x_n \\ \hat{\beta}_1 &= \frac{\sum_{n=1}^{N} x_n (y_n - \bar{y})}{\sum_{n=1}^{N} x_n (x_n - \bar{x})}.\end{aligned}$$

To put this in a more standard form, we use a slight algebra trick. Note that

$$\sum_{n=1}^{N} c(z_n - \bar{z}) = 0$$

for any constant c and any collection $z_1, \ldots, z_N$ with sample mean $\bar{z}$ (this can easily be verified by expanding the sum). Since $\bar{x}$ is a constant, we can then subtract $\sum_{n=1}^{N} \bar{x}(y_n - \bar{y})$ from the numerator and $\sum_{n=1}^{N} \bar{x}(x_n - \bar{x})$ from the denominator without affecting our slope estimate. Finally, we get

$$\hat{\beta}_1 = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N}(x_n - \bar{x})^2}.$$

2. Multiple Regression

Model Structure

In multiple regression, we assume our target variable to be a linear combination of *multiple* predictor variables. Letting x_{nj} be the j^{th} predictor for observation n, we can write the model as

$$y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_D x_{nD} + \epsilon_n.$$

Using the vectors $\mathbf{x}_n$ and $\boldsymbol{\beta}$ defined in the previous section, this can be written more compactly as

$$y_n = \boldsymbol{\beta}^\top \mathbf{x}_n + \epsilon_n.$$

Then define $\hat{\boldsymbol{\beta}}$ the same way as $\boldsymbol{\beta}$ except replace the parameters with their estimates. We again want to find the vector $\hat{\boldsymbol{\beta}}$ that minimizes the RSS:

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) = \frac{1}{2}\sum_{n=1}^{N}\left(y_n - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_n\right)^2 = \frac{1}{2}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2,$$

Minimizing this loss function is easier when working with matrices rather than sums. Define $\mathbf{y}$ and $\mathbf{X}$ with

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \dots \\ y_N \end{bmatrix} \in \mathbb{R}^N, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \dots \\ \mathbf{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times (D+1)},$$

which gives $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \mathbb{R}^N$. Then, we can equivalently write the loss function as

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Parameter Estimation

We can estimate the parameters in the same way as we did for simple linear regression, only this time calculating the derivative of the RSS with respect to the entire parameter vector. First, note the commonly-used matrix derivative below [1].

> **Math Note**
>
> For a symmetric matrix $\mathbf{W}$,
>
> $$\frac{\partial}{\partial \mathbf{s}}(\mathbf{q} - \mathbf{A}\mathbf{s})^\top \mathbf{W}\,(\mathbf{q} - \mathbf{A}\mathbf{s}) = -2\mathbf{A}^\top \mathbf{W}\,(\mathbf{q} - \mathbf{A}\mathbf{s})$$

Applying the result of the Math Note, we get the derivative of the RSS with respect to $\hat{\boldsymbol{\beta}}$ (note that the identity matrix takes the place of $\mathbf{W}$):

$$\begin{aligned} \mathcal{L}(\hat{\boldsymbol{\beta}}) &= \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \frac{\partial \mathcal{L}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} &= -\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \end{aligned}$$

We get our parameter estimates by setting this derivative equal to 0 and solving for $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} (\mathbf{X}^\top \boldsymbol{\Phi})\hat{\boldsymbol{\beta}} &= \mathbf{X}^\top \mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^\top \mathbf{X}^\top \mathbf{y} \end{aligned}$$

---

[1] A helpful guide for matrix calculus is The Matrix Cookbook

---

By Danny Friedman