Approach 2: Maximizing Likelihood

Contents

1. Simple Linear Regression

Model Structure

Using the maximum likelihood approach, we set up the regression model probabilistically. Since we are treating the target as a random variable, we will capitalize it. As before, we assume

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n,$$

only now we give ϵ_n a distribution (we don't do the same for x_n since its value is known). Typically, we assume the ϵ_n are independently Normally distributed with mean 0 and an unknown variance. That is,

$$\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

The assumption that the variance is identical across observations is called *homoskedasticity*. This is required for the following derivations, though there are *heteroskedasticity-robust* estimates that do not make this assumption.

Since β_0 and β_1 are fixed parameters and x_n is known, the only source of randomness in Y_n is ϵ_n. Therefore,

$$Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_n, \sigma^2),$$

since a Normal random variable plus a constant is another Normal random variable with a shifted mean.

Parameter Estimation

The task of fitting the linear regression model then consists of estimating the parameters with maximum likelihood. The joint likelihood and log-likelihood across observations are as follows.

$$\begin{aligned}
L(\beta_0, \beta_1; Y_1, \ldots, Y_N) &= \prod_{n=1}^{N} L(\beta_0, \beta_1; Y_n) \\
&= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_n - (\beta_0 + \beta_1 x_n))^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\sum_{n=1}^{N} \frac{(Y_n - (\beta_0 + \beta_1 x_n))^2}{2\sigma^2}\right) \\
\log L(\beta_0, \beta_1; Y_1, \ldots, Y_N) &= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (Y_n - (\beta_0 + \beta_1 x_n))^2.
\end{aligned}$$

Our $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates are the values that maximize the log-likelihood given above. Notice that this is equivalent to finding the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the RSS, our loss function from the previous section:

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

In other words, we are solving the same optimization problem we did in the last section. Since it's the same problem, it has the same solution! (This can also of course be checked by differentiating and optimizing for $\hat{\beta}_0$ and $\hat{\beta}_1$). Therefore, as with the loss minimization approach, the parameter estimates from the likelihood maximization approach are

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{\sum_{n=1}^{N} (x_n - \bar{x})(Y_n - \bar{Y})}{\sum_{n=1}^{N} (x_n - \bar{x})^2}.
\end{aligned}$$

Print to PDF

2. Multiple Regression

Still assuming Normally-distributed errors but adding more than one predictor, we have

$$Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\beta}^\top \mathbf{x}_n, \sigma^2).$$

We can then solve the same maximum likelihood problem. Calculating the log-likelihood as we did above for simple linear regression, we have

$$
\begin{aligned}
\log L(\beta_0, \beta_1; Y_1, \ldots, Y_N) &= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(Y_n - \boldsymbol{\beta}^\top \mathbf{x}_n\right)^2 \\
&= -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).
\end{aligned}
$$

Again, maximizing this quantity is the same as minimizing the RSS, as we did under the loss minimization approach. We therefore obtain the same solution:

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

By Danny Friedman