Simple Linear Regression

Prof. Dennis Sun

Data 401
Simple Linear Regression

In **simple linear regression**, we observe data \((x_i, Y_i), i = 1, ..., n\). We assume that the data are generated according to:

\[
Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,
\]

where \(\epsilon_i \sim \text{Normal}(0, \sigma^2)\).

How many parameters are there in this model? 3 \((\beta_0, \beta_1, \sigma^2)\)
MLE for Simple Linear Regression

In simple linear regression, the $x_i$'s are assumed fixed. Only the $Y_i$'s are random.

Are $Y_1, ..., Y_n$ i.i.d.? No! They're independent, but not identically distributed. Each one has a different mean.

Another way to write the model: $Y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$.

To avoid unnecessary algebra, we'll work with a simpler version of this model: $Y_i \sim \text{Normal}(\beta x_i, \sigma^2)$, with just 2 parameters.

Let's first write down the likelihood of the data.

$$p_{\beta, \sigma^2}(Y_1, ..., Y_n) = \prod_{i=1}^{n} p_{\beta, \sigma^2}(Y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} (Y_i - \beta x_i)^2}$$
The likelihood is \( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta x_i)^2} \) and now we take logs:

\[
\log L(\beta, \sigma^2) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - \beta x_i)^2
\]

\[
= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(Y_i - \beta x_i)^2
\]

How do we find the optimal \( \beta \) and \( \sigma^2 \)?

Taking partial derivatives, we obtain:

\[
\frac{\partial \log L}{\partial \beta} = -\frac{1}{\sigma^2} \sum_{i=1}^{n}(Y_i - \beta x_i)(-x_i) = 0 \quad \hat{\beta}_{ML} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}
\]

\[
\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n}(Y_i - \beta x_i)^2 = 0 \quad \hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n}(Y_i - \hat{\beta}_{ML} x_i)^2
\]
Extension to Model with Intercept

In the above calculation, we assumed \( Y_i \sim \text{Normal}(\beta x_i, \sigma^2) \). In other words, we assumed a linear model where the intercept was zero.

More realistically, we would like to assume a model with intercept: \( Y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2) \). Now we have 3 parameters.

With some more algebra, you can show:

\[
(\hat{\beta}_1)_{\text{ML}} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}
\]

\[
(\hat{\beta}_0)_{\text{ML}} = \overline{Y} - (\hat{\beta}_1)_{\text{ML}} \overline{x}
\]

\[
(\hat{\sigma}^2)_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0)_{\text{ML}} - (\hat{\beta}_1)_{\text{ML}} x_i)^2
\]
A Look Ahead

Simple linear regression is not very interesting. We want to be able to fit linear regression models with lots of variables:

\[ Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \]

where \( \epsilon_i \sim \text{Normal}(0, \sigma^2) \).

How do we estimate all of these parameters \((\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)\)? To answer this question, we’ll need another tool in our mathematical toolbelt: linear algebra.