2021F_Week04

September 20, 2021

0.1 Week04

Regression is a fundemental tool for studying relationships between random variables. We will first study Simple Linear Regression because this model can serve as a foundation for most future regression techniques.

Our goal this week is to: 1. Understand regression as a probabilistic model 2. Estimate the parameters of this model using a Residual Sum of Squares (RSS) strategy 3. Estimate the parameters of this model using a maximum likelihood strategy 4. Discuss geometric properties of these estimators

0.1.1 Data Setup

Suppose we are given a dataset that contains pairs of points $D = [(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_N, y_N)]$ and further we assume that the data point x_i was generated by a random variable X_i and that the data point y_i was generated from a random variable Y_i.

0.1.2 Probabilistic and model form

Simple Linear Regression supposes the following conditional probability between the above random variables

$$Y_i | x_i \sim N(\beta_0 + x_i\beta_1, \sigma^2)$$

The conditional probability of Y_i is linearly related to x_i with two parameters: an intercept (β_0) and a slope (β_1). A third parameter σ^2 is ued to express the variability around the conditional mean. To note, this setup assumes every Y_i has a similar normal distribution, using the same parameter values but because x_i is not necessarily the same for each Y_i, the mean of this normal distribution may differ.

Often, the equation

$$Y_i | x_i \sim N(\beta_0 + x_i\beta_1, \sigma^2)$$

is written

$$Y_i | x_i \sim N(\mu(x_i), \sigma^2)$$

where $\mu(x_i) = \beta_0 + x_i\beta_1$ to emphasize that the Normal distribution is governed by two parameters and that our focus is on μ as a function of data points we collected.

When we write a regression model in terms of a single, or in more complex case many, probability distributions, it is called **probabilistic form**. Probabilsitic form highlights the distribution of our variable of interest (Y).

Another common way to write this relationship is

$$y_i = \beta_0 + x_i * \beta_1 + \epsilon_i \tag{1}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{2}$$

This is called **model form** for SLR. Model form highlights the relationship between Y and X, focusing less on the distribution of Y.

0.2 Expected value and Variance

0.2.1 Expected value

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ then the expected value of X is

$$E(X) = \mu \tag{3}$$

We can apply the above to compute the expected value of $Y|x$. Because the conditional probability of Y given x has a normal distribution with $\mu(x_i)\beta_0 + x_i\beta_1$ then the expected value is

$$E(Y_i|x_i) = \mu(x_i) = \beta_0 + x_i\beta_1 \tag{4}$$

0.2.2 Variance

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ then the variance of X is

$$Var(X) = \sigma^2 \tag{5}$$

We can apply the above to compute the variance of $Y|x$. Because the conditional probability of Y given x has a normal distribution with $\mu(x_i)\beta_0 + x_i\beta_1$ and variance σ^2 then the variance is

$$Var(Y_i|x_i) = \sigma^2 \tag{6}$$

0.2.3 Maximum likelihood

To compute the likelihod function, we will assume the random variables Y_i and Y_j for any i and any j are independent. This means then $p(Y_j|Y_i) = p(Y_j)$.

The likelihood is

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{N} f(y_i|x_i, \beta_0, \beta_1, \sigma^2) && (7) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} \right\} && (8) \\
\mathcal{L}(\beta_0, \beta_1, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp\left\{ -\sum_{i=1}^{N} \frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} \right\} && (9) \\
& && (10)
\end{aligned}
$$

and the loglikelihood is

$$
\begin{aligned}
\ell\ell = \log\left[\mathcal{L}(\beta_0, \beta_1, \sigma^2)\right] &= \log\left[\left(\frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp\left\{ -\sum_{i=1}^{N} \frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} \right\} \right] && (11) \\
&= N \log\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \sum_{i=1}^{N} \frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} && (12) \\
\ell\ell(\beta_0, \beta_1, \sigma^2) &= -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - [\beta_0 + \beta_1 x_i])^2 && (13) \\
& && (14)
\end{aligned}
$$

The above loglieklihood, when maximized for β_0,β_1, and σ will return the maximum likelihod estimates for these parameters $\hat{\beta}_0$,$\hat{\beta}_1$, and $\hat{\sigma}$.

With the above maximum likelihood estimates we could find the maximum likelihood esitmate of the expected value of Y_i, in other words we can compute the mle of $E(Y_i|x_i)$ as $E(Y_i|x_i) = \hat{\beta}_0 + x_i\hat{\beta}_1$.

0.2.4 Residual Sum Squares

In BSTA001 we talked about one intuitive way to find "optimal" β_0 and β_1 parameters. That setup went as follows:

We would like to find parameters β_0 (the intercept) and β_1 (the slope) so that they are, in some sense, optimal. There are many different ways to define optimal. The most common method to define an optimal β_0 and β_1 for linear regression is least squares.

Given N pairs (x_i, y_i), a solution to the least squares equation is the pair (β_0, β_1) such that

$$L(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - [\beta_0 + \beta_1 x_i])^2 \tag{15}$$

We want to find β_0 and β_1 so that the squared **vertical** distance between any pair (x_i, y_i) and our line is on average minimized.

That is, we wish to find the pair (β_0^*, β_1^*) such that

$$L(\beta_0^*, \beta_1^*) \leq L(\beta_0, \beta_1) \tag{16}$$

for all pairs (β_0, β_1).

Recall our equation for the loglikelihood

$$\ell\ell(\beta_0, \beta_1, \sigma^2) = -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - [\beta_0 + \beta_1 x_i])^2 \tag{17}$$

We see the same term $L(\beta_0, \beta_1)$ appear in the loglikelihood. However, in the loglikelihood that term has a negative sign infront of it. But to maximize $-L(\beta_0, \beta_1)$ is equivalent to minimizing $L(\beta_0, \beta_1)$, and so the least squares solution and the maximum likeliehood solution are the same.