Maximum Likelihood Estimation For Regression

Ashan Priyadarshana
Feb 15, 2018 · 9 min read

Maximum likelihood estimation or otherwise noted as MLE is a popular mechanism which is used to estimate the model parameters of a regression model. Other than regression, it is very often used in statics to estimate the parameters of various distribution models.

photo courtesy Andrew Gook

A section wise summary of the artical is as follows. Although post is written with assumption of reader being started from begining, feel free to jump to any section at your desire.
-Normal / Gaussian distribution
-Binomial distribution
-What is MLE

-How to Calculate Likelihood

-How to use MLE for linear regression

Normal / Gaussian distribution

In probability the normal or gaussian distribution is a very famous continous probability distribution. It basically depends on two factors — the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph and the standard deviation determines the height and width of the graph. So notation of normal distribution becomes:

$$N(\mu, \sigma^2) \quad ; \ \mu \text{ - mean } , \sigma^2 \text{ - variance}$$

PDF of Normal distribution is:

$$f(x \mid \mu, \ \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Note: from a pdf function, we can get the probability related to a data point x. So by substituting values for x, the relevant probabilities can be obtained. We can think of PDFs as models and that they are defined by their parameters. So parameters of normal distribution are : *μ (mean)* and *σ² (variance)*

Binomial distribution:

When a trail of two outcomes (as success and fail) is repeated for *n* times and when the probabilities of "*number of success event*" is logged, the resultant distribution is called a binomial distribution. For an example lets toss a coin for 10 times (*n = 10*) and the success is getting head. So if we log the probabilities of getting head only one times, two times, three times, … then that distribution of probabilities is in a binomial distribution.

Hence the binomial distribution depends mainly upon "number of trials" and "probability of success in an individual trail". Thus the notation of binomial distribution is:

$$b(x;\, n,\, P) \ \text{;n - number of trials, P - success probability of an individual trial}$$

PDf of Binomial Distribution:

$$b(x; n, P) = {}^{n}C_{x} . P^{x} . (1 - P)^{n - x}$$

So *n* and *P* are the parameters of a Binomial distribution.

Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a technique used for estimating the parameters of a given distribution, using some observed data. For example, if a population is known to follow a "normal distribution" but the "mean" and "variance" are unknown, MLE can be used to estimate them using a limited sample of the population. MLE does that by finding particular values for the parameters (mean and variance) so that the resultant model with those parameters (mean and variance) would have generated the data.

So generally, likelihood expression is in the form of: ***L(parameters | data).*** Meaning of this is, "*likelihood of having these parameters, once the data are these*".

Likelihood and Probability are two different things although they look and behaves same. We talk about probability when we know the model parameters and when predicting a value from that model. So there we talk about how probable is the resultant value to be come out from that model. So probability is: ***P(data | parameters)***

Now we can see that Likelihood is other side of probability. That is we are going to guess the model parameters from the data. So there we know the results well and we know for sure that they have occured (probability = 1).

A simple example

Suppose we have 3 data points as 2, 2.5 and 3. And let's suppose that these values are from a normal/gaussian distribution. So now we have some data from a model (here model being gaussian), but we don't know it's parameters. Let's see how to estimate model parameters from MLE.

First let's calculate some Likelihoods after assuming some values for parameters. For instance let's think mean and variance as 2 and 1 for one instance and for another instance let's assume they are as 4 and 2.

Calculating Likelihood

It's very important to undestand that likelihood is also calculated from PDF functions but by calculating the joint probabilities of data points from a particular PDF function. That is, we can write likelihood calculation as:

$$\text{L(parameters | data)} = \prod_{i=1}^{n} \text{f(data}_i \text{ | parameters)}$$

So likelihood when model is N(μ=2, σ^2=1):

$$\text{L}(\mu=2,\ \sigma^2=1 \mid x = 2,\ 2.5,\ 3) = \text{PDF}(x = 2) \times \text{PDF}(x = 2.5) \times \text{PDF}(x = 3)$$
$$\text{L}(2,\ 1 \mid x = 2,\ 2.5,\ 3) = f(x = 2 \mid 2,\ 1)\ .\ f(x = 2.5 \mid 2,\ 1)\ .\ f(x = 3 \mid 2,\ 1)$$

In above we have calculated the joint probability of 3 points assuming that they are independent. Recall that we calculate the probability from the PDF of that particular distribution.

$$\text{L}(2,\ 1 \mid x = 2,\ 2.5,\ 3) = \frac{1}{\sqrt{2\pi 1}} e^{-\frac{(2-2)^2}{2.1}}\ .\ \frac{1}{\sqrt{2\pi 1}} e^{-\frac{(2.5-2)^2}{2.1}}\ .\ \frac{1}{\sqrt{2\pi 1}} e^{-\frac{(3-2)^2}{2.1}}$$
$$\text{L}(2,\ 1 \mid x = 2,\ 2.5,\ 3) = 0.39 * 0.35 * 0.24 = 0.03276$$

Likelihood when model is N(4,2):

$$\text{L}(\mu=4,\ \sigma^2=2 \mid x = 2,\ 2.5,\ 3) = \text{PDF}(x = 2) \times \text{PDF}(x = 2.5) \times \text{PDF}(x = 3)$$
$$\text{L}(4,\ 2 \mid x = 2,\ 2.5,\ 3) = \frac{1}{\sqrt{2\pi.2}} e^{-\frac{(2-4)^2}{2.2}}\ .\ \frac{1}{\sqrt{2\pi.2}} e^{-\frac{(2.5-4)^2}{2.2}}\ .|\ \frac{1}{\sqrt{2\pi.2}} e^{-\frac{(3-4)^2}{2.2}}$$
$$\text{L}(4,\ 2 \mid x = 2,\ 2.5,\ 3) = 0.10 * 0.16 * 0.21 = 0.00336$$

So now we can see that likelihood of parameter values being μ=2, σ^2=1 is greater than them being μ=4, σ^2=2. But how can we calculate the precise values for μ and σ^2 ? So in mathematics, to find values regarding optimizations, derivation is used.

So what we do is we write the likelihood calculation function for all the known/selected data points as follows.

$$\text{L}(\mu,\ \sigma^2 \mid x = 2,\ 2.5,\ 3) = \prod_{i=1}^{3} \text{f}(\ x_i \mid \mu,\ \sigma^2\ )$$

$$L(\mu, \sigma^2 \mid x = 2, 2.5, 3) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{}{2\sigma^2}} \cdot| \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{}{2\sigma^2}}$$

$$L(\mu, \sigma^2 \mid x = 2, 2.5, 3) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^3 e^{-\frac{(2-\mu)^2}{2\sigma^2}} \cdot e^{-\frac{(2.5-\mu)^2}{2\sigma^2}} \cdot e^{-\frac{(3-\mu)^2}{2\sigma^2}}$$

Let’s take the natural logarithms in both sides,

$$\begin{aligned} \ln(L(\mu, \sigma^2 \mid x = 2, 2.5, 3)) &= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^3 + \ln\left(e^{-\frac{(2-\mu)^2}{2\sigma^2}}\right) + \ln\left(e^{-\frac{(2.5-\mu)^2}{2\sigma^2}}\right) + \ln\left(e^{-\frac{(3-\mu)^2}{2\sigma^2}}\right) \\ &= |3\ln(1) - 3\ln(\sqrt{2\pi\sigma^2}) - \frac{(2-\mu)^2}{2\sigma^2} - \frac{(2.5-\mu)^2}{2\sigma^2} - \frac{(3-\mu)^2}{2\sigma^2} \\ &= -3\ln(\sqrt{2\pi\sigma^2}) - \frac{(2-\mu)^2}{2\sigma^2} - \frac{(2.5-\mu)^2}{2\sigma^2} - \frac{(3-\mu)^2}{2\sigma^2} \\ &= -3\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(4\text{-}4\mu + \mu^2 + 6.25\text{-}5\mu + \mu^2 + 9\text{-}6\mu + \mu^2) \end{aligned}$$

$$\ln(L(\mu, \sigma^2 \mid x = 2, 2.5, 3)) = -3\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(19.25 - 15\mu + 3\mu^2)$$

Now let’s find Maximum likelihood estimators for μ. For that lets partially differentiate the above equation with respect to μ.

$$\frac{\partial(\ln(L(\mu,\sigma^2 \mid x))}{\partial\mu} = \frac{1}{2\sigma^2}(0 - 15 + 6\mu)$$

In order to find maximum(optimal) likelihood estimators let’s equal above value to 0.

$$\frac{1}{2\sigma^2}(0 - 15 + 6\mu) = 0$$

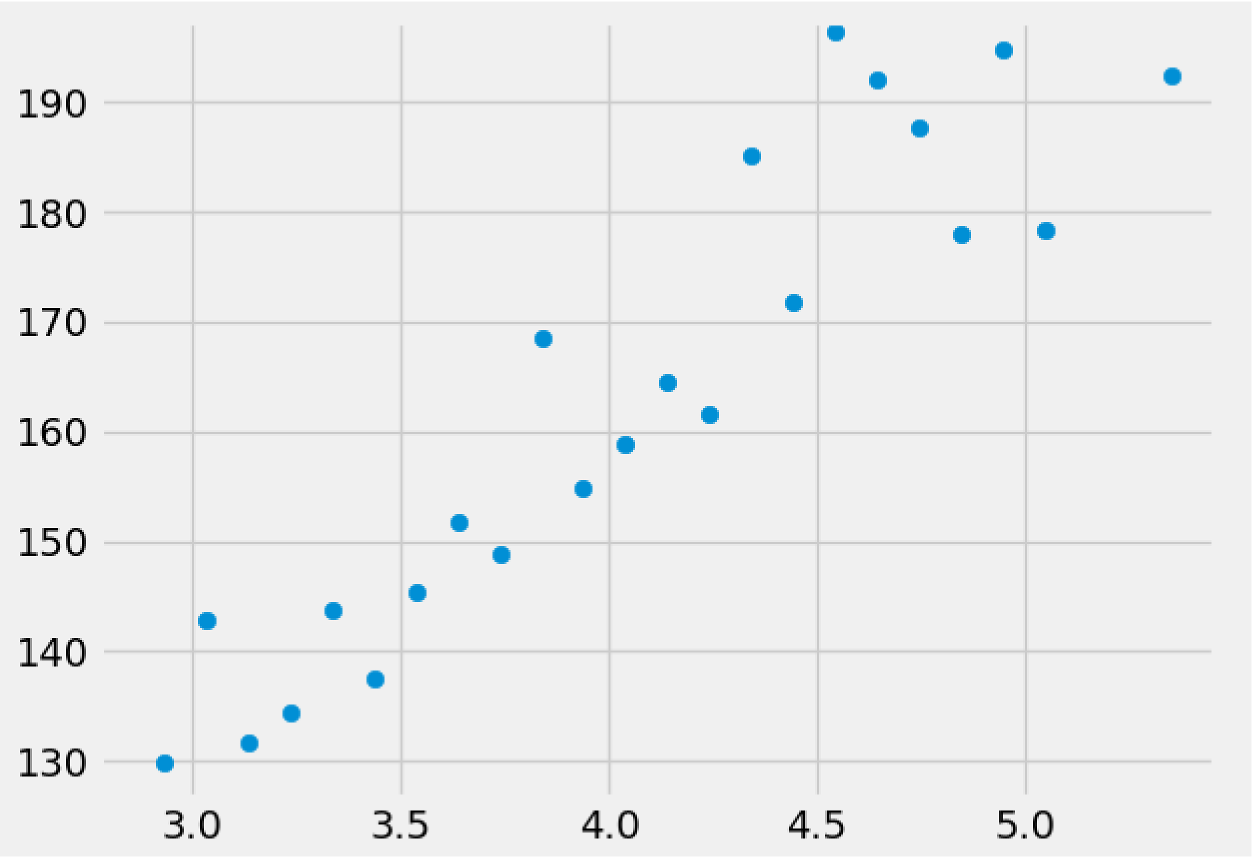
$$\mu = 15/6 = 2.5$$

So now we know what is the MLE of μ. Like this we can get the MLE of σ^2 also by derivative w.r.t σ^2.

MLE for Linear Regression

As we have used likelihood calculation to find the best parameter values for various distribution models in statistics, MLE method can also be used to find the best model parameters of a linear regression model. But when calculating parameters values for those statistical distribution models, we knew what kind of distributions was it and the relevant PDF function. What kind of distribution are we going to use in linear regression? With that question, we can talk about the main assumption in linear regression:

> “The independent variable (y values) is assumed be in a normal distribution”



An example data set. Blue dots are labels (y values)

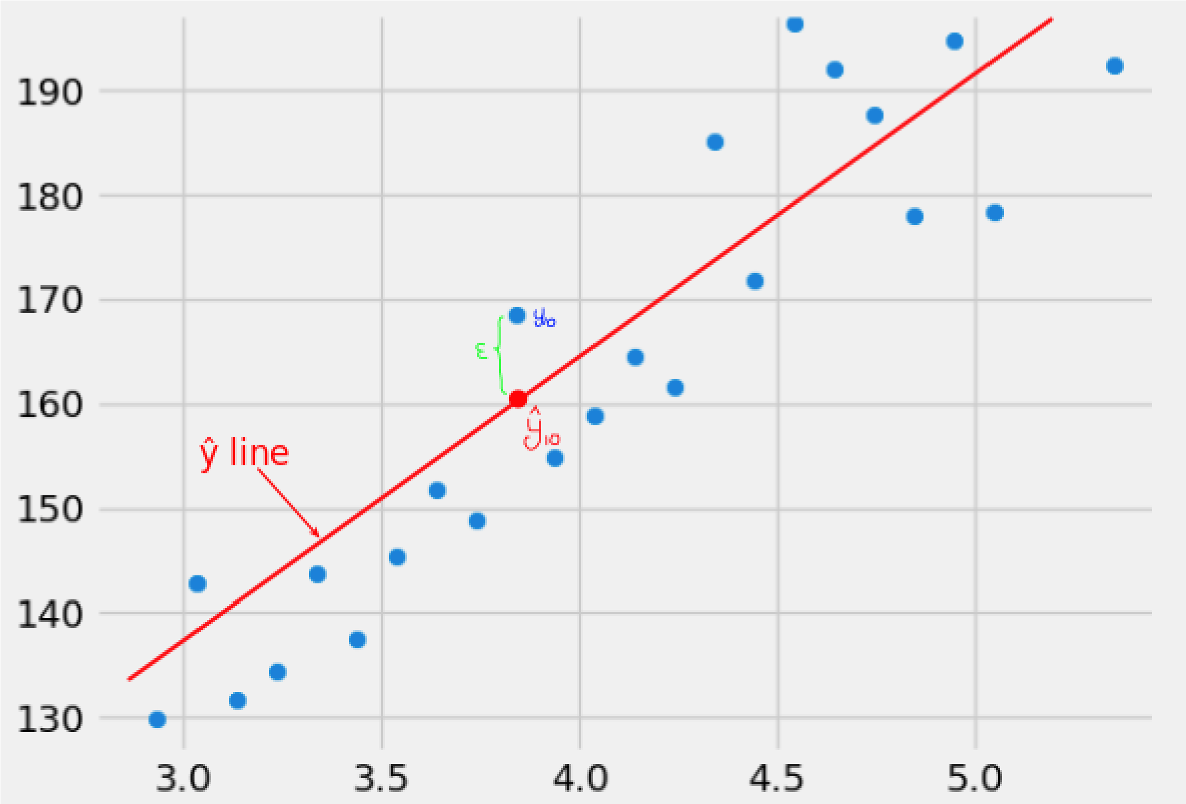
But that’s not a full description. When we talk about some values being in a normal distribution, we need to describe more about that normal distribution ; like what kind of normal distribution? More precisely what are the mean and variance of that normal distribution?

In linear regression the trick that we do is, **we take the model that we need to find, as the mean** of the above stated normal distribution. Because we know how to find MLE values of a mean in a normal distribution.

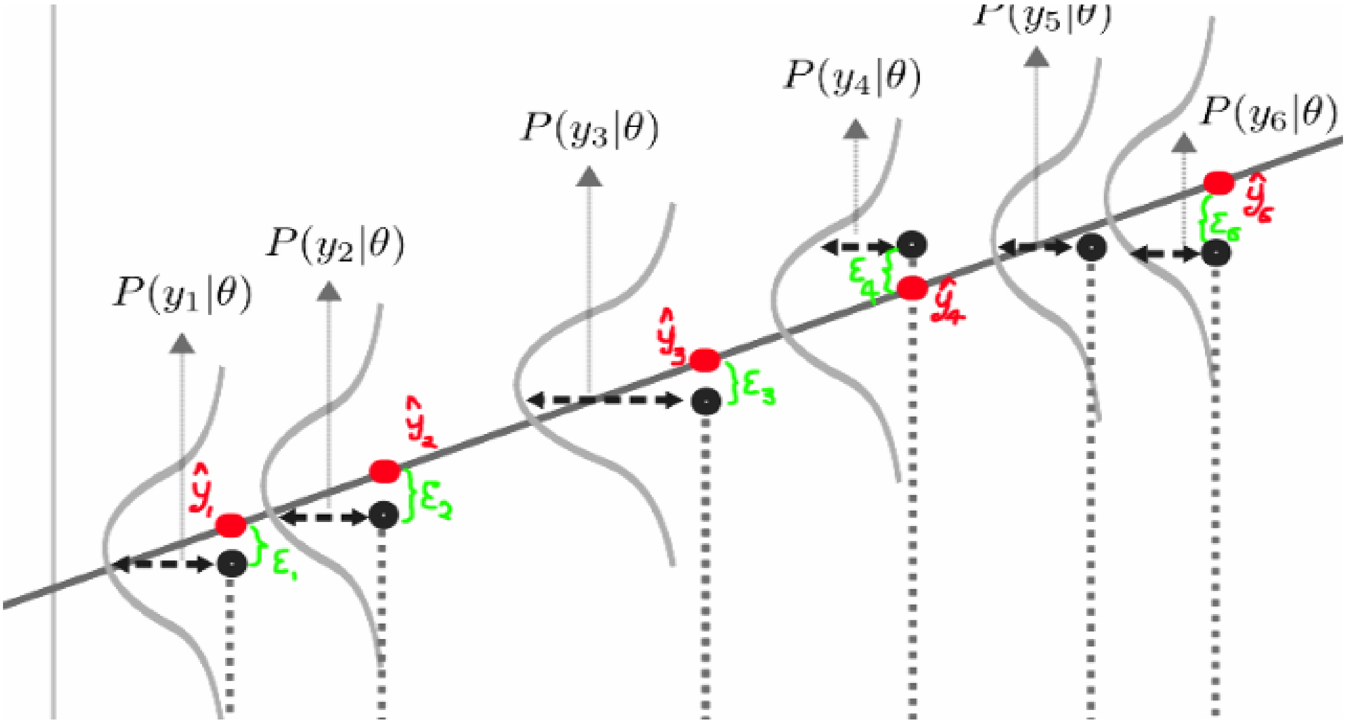
So let’s define our linear model that needed to be estimated as *ŷ*.

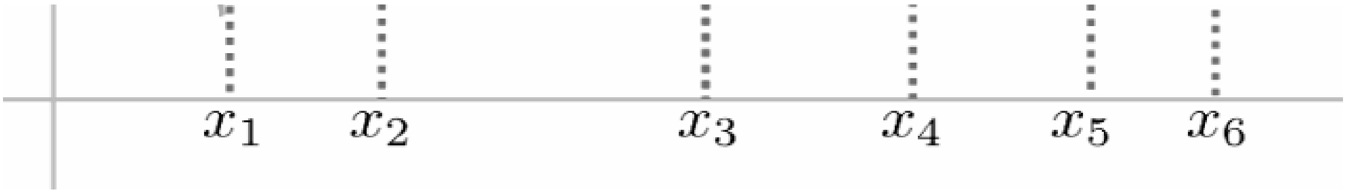
$$\hat{y} = w_0 + w_1x_1 + \dots + w_dx_d \quad \text{- Model to be predicted}$$

But there is a special thing about this mean. That is, it’s not a fixed value. In earlier discussed normal distribution models, the mean was a fixed value (a number). But in linear regression, the mean is a function (ŷ). So you need to understand that for every x value (input) , there will be a number generated by function ŷ as the mean. So from ŷ function, we get a set of values as means. And the important thing to understand from this is that mean for each y value (each result / label) is a different. That is, mean for each y value (label) is the value predicted by our model.



As an example, in the above figure, we know that blue dots are labels (y) and they are assumed to be from a normal distribution. And mean for each is not a fixed number, while it's the corresponding value from the ŷ function. So the mean for label y-10 is ŷ-10 and we should understand the graph as follows. That is each labels in the data set have their own mean and variance in a normal distribution.





Note: ŷ is called a linear model as there are only variables with degree one and lesser. If there were variables with degree 2, then it's not a linear model

As stated above In linear regression, we treat above line ŷ as the "mean" of the normal distribution.

$$\hat{y} = w_0 + w_1x_1 + \ldots + w_dx_d$$
$$\hat{Y}_{n*1} = X_{n*d}\, W_{d*1} \quad \text{; in vector form, n - number of records}$$

We can consider this ŷ data as also in a normal distribution. But this time, their mean values will be them self since they fall along on top of the ŷ line perfectly. And so the variance of these ŷ data (predicted labels) will be 0. So, ŷ~N(XW , 0)

We have an error term called ε (residual) which is the distance between predicted value (ŷ) and actual value(y). And there are some important assumptions that we do in linear regression regarding these residuals, and they are:

> *" Residuals are normally distributed"*
> *" Residuals have an equal variance"*
> *" Means of residuals are 0"*

So:

$$\varepsilon \sim N(0\ ,\ \sigma^2)$$

And we know that y = ŷ + ε and y labels are normally distributed. Our aim is to estimate the best values for mean and variance of normal distribution y. Let's get the mean and variance of y in terms of ŷ and ε normal distributions. We know the mean is termed as expectation. So let's get the expectation of y=ŷ + ε equation in order to find the mean (expectation) of y.

E(y) = E(ŷ + ε)
E(y) = E(ŷ) + E(ε)
E(y) = XW + 0 = XW

And,

Variance(y) = Variance($\hat{y}$ + ε)
Variance(y) = Variance($\hat{y}$)+ Variance(ε)
Variance(y) = 0 + σ^2

So we can say that y is a normal distribution with mean XW and variance σ^2.

$$y \sim N(XW\ ,\ \sigma^2)$$

Now let's calculate the MLEs for XW and σ^2 as we did in previous example. But note that here we hava n data points (in earlier example we had only 3), and each of those data point are of dimension d.

$$L(XW,\ \sigma^2 |\ x\ ) = \prod_{i=1}^{n} f_y(y_i,\ x_i w,\ \sigma^2)$$

As y is a normal distribution,

$$L(XW,\ \sigma^2 |\ x\ ) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i w)^2}{2\sigma^2}}$$

$$L(XW,\ \sigma^2 |\ x\ ) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n . e^{-\frac{\sum_{i=1}^{n} (y_i - x_i w)^2}{2\sigma^2}}$$

$$L(XW,\ \sigma^2 |\ x\ ) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n . e^{-\frac{(Y - XW)^T (Y - XW)}{2\sigma^2}}$$

Let's get natural logarithms in both sides so that we get the log likelihood,

$$ln(\ L(XW,\ \sigma^2 |\ x\ )\ ) = -\frac{n}{2}\ ln(2\pi) - \frac{n}{2}\ ln(\sigma^2) - \frac{(Y - XW)^T (Y - XW)}{2\sigma^2}$$

In order to estimate the best set of weights (weight matrix), let's partially differentiate the above equation from w.

$$\frac{\partial ln(\ L(XW, \sigma^2 |\ x\ )\ )}{} = \frac{1}{} \frac{\partial (Y - XW)^T (Y - XW)}{}$$

$$\partial w \qquad 2\sigma^2 \qquad \partial w$$

$$\frac{\partial ln(L(XW, \sigma^2 | x))}{\partial w} = \frac{1}{2\sigma^2} \frac{\partial (Y^2 - 2X^T WY + X^T XW^2)}{\partial w}$$

$$\frac{\partial ln(L(XW, \sigma^2 | x))}{\partial w} = \frac{1}{2\sigma^2} (0 - 2X^T Y + 2X^T XW)$$

Optimal values for W is when

$$\frac{\partial ln(L(XW, \sigma^2 | x))}{\partial w} = 0$$

Then,

$$\frac{1}{2\sigma^2} (0 - 2X^T Y + 2X^T XW) = 0$$

$$W = \frac{X^T Y}{X^T X}$$

$$W = (X^T X)^{-1} X^T Y$$

So now we have found the optimal values for Ws in our model. And that is the main aim of linear regression since once found the w matrix, we can predict.

Machine Learning Data Science Mle Regression Linear Regression