Analytic Maximum Likelihood for the parameters of the Gaussian and Bernoulli distributions, as well as the parameters of a linear model using Ordinary Least Squares

Joel S Steele

In this demonstration we will be looking at how to solve for the parameters of a given model using Maximum Likelihood. The models to be included are the Gaussian or Normal distribution the Bernoulli distribution and the Ordinary Least Squares loss function.

1 Gaussian Likelihood

When data are drawn from a Normal distribution, $\sim \mathcal{N}(\mu, \sigma^2)$, we can use the Gaussian distribution function to describe the probability of the data.

$$f(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)} \tag{1}$$

This specifications represents how to compute the probability for a single value x_i. That means, we can get the value of the function for any particular input, x_i, if we supply the parameters μ and σ^2.

A quick aside

You may be wondering why we have discussed probabilities but we are interested in likelihoods? Well, the terms are often used interchangeably, which is a shame. In our application however, we will say that, if we know the parameter values for a distribution, we can compute a probability of any observation we obtain. The result tells us the probability (or how likely we are to see) a value like that given the distribution that we have at hand.

If, however, we don't know the exact parameters of our distribution, but instead we have a set of observations, we must figure out which values of the parameters would result in the largest probability. In essence we are going *backwards* and using the data along with the hypothesized shape of the probability distribution, in order to find the parameters that we believe produced our observations. In this latter case, we are interested in finding the parameters which maximize the likelihood that our observations are distributed a particular way.

1.1 Likelihood of a set of values

The specification in equation 1 works for a single observation. However, the specification changes when we are dealing with an entire set of observations. From probability theory, we know that, if observations are independent, their *joint* probability is the product of their individual probabilities. So, for our set of observations, we compute the probability value of each point, and then multiply them all together to get the probability of the entire sample. What does this mean? Well, we literally multiply each obtained value from the function. The result is,

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\mu, \sigma^2) &= \prod_i^n f(x_i|\mu, \sigma^2) \\ &= f(x_1|\mu, \sigma^2) \times f(x_2|\mu, \sigma^2) \times \ldots \times f(x_n|\mu, \sigma^2) \end{aligned} \tag{2}$$

Using our Gaussian function this translates to,

$$\begin{aligned} &= \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right)} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x_2-\mu)^2}{2\sigma^2}\right)} \times \ldots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)}. \end{aligned} \tag{3}$$

This product can be simplified somewhat. To help illustrate we will take advantage of the fact that the product operator, $\prod_i^n$, can be distributed algebraically.

$$\prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)} = \prod_i^n \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right] \times \prod_i^n \left[e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}\right] \tag{4}$$

Thus, we can deal with each portion one at a time.

1.1.1 First portion.

First, we see that the $\frac{1}{\sqrt{2\pi\sigma^2}}$ term does not involve the observation x_i, which makes it a constant. We also know that taking the product of a constant, is equivalent to having the constant multiplied by itself a number of times. In this case n times. So, we can express the first portion of the joint probability as,

$$\prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n. \tag{5}$$

We may remember that a faction can be expressed as some term raised to a negative power, and that the square root is equal to the raising a term to the $\frac{1}{2}$ power. Thus, we can alternatively express the fraction in our first term as follows,

$$\frac{1}{\sqrt{2\pi\sigma^2}} = (2\pi\sigma^2)^{-\frac{1}{2}}. \tag{6}$$

This is helpful, since raising a power to another power is the same as multiplying the two powers together, $(a^b)^c = a^{bc}$. This means that the product of the first term can be simplified as the fraction to the power n. Again, this is the same as multiplying the powers together. The result is,

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n = \left((2\pi\sigma^2)^{-\frac{1}{2}}\right)^n = (2\pi\sigma^2)^{(-\frac{1}{2})\times(n)} = (2\pi\sigma^2)^{(-\frac{n}{2})}. \tag{7}$$

1.1.2 Second portion

Okay, on to the $e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$ part. First, we can re-express the entire power portion as $(-\frac{1}{2\sigma^2}) \times (x_i - \mu)^2$, so this can be rewritten as $e^{\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)}$.

It is important to recognize that if we have a base number, raised to a power, multiplied by the same base number, raised to a different power, this is equal to the base raised to the sum of the two powers. For example

$$2^2 \times 2^3 = (2 \times 2) \times (2 \times 2 \times 2) = 2^{2+3} = 2^5 = 32. \tag{8}$$

What this allows us to do is to express the product of our second portion as e raised to a summation over $-\frac{1}{2\sigma^2}(x_i - \mu)^2$ as seen below,

$$\begin{aligned} \textstyle\prod_i^n e^{\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)} &= e^{\left(-\frac{1}{2\sigma^2}(x_1-\mu)^2\right)} \times e^{\left(-\frac{1}{2\sigma^2}(x_2-\mu)^2\right)} \times \ldots \times e^{\left(-\frac{1}{2\sigma^2}(x_n-\mu)^2\right)} \\ &= e^{\left[-\frac{1}{2\sigma^2}(x_1-\mu)^2 + -\frac{1}{2\sigma^2}(x_2-\mu)^2 + \ldots + -\frac{1}{2\sigma^2}(x_n-\mu)^2\right]} \\ &= e^{\left(\sum_i^n -\frac{1}{2\sigma^2}(x_i-\mu)^2\right)} \end{aligned} \tag{9}$$

Notice that in the exponent the term $-\frac{1}{2\sigma^2}$ is a constant relative to the portion involving x_i. This means that it can be moved outside of the summation. This allows us to simplify the expression as,

$$e^{\left(-\frac{1}{2\sigma^2}\sum_i^n (x_i-\mu)^2\right)} \tag{10}$$

Knowing all of this, we can express the *joint probability* of all our observations using the *Gaussian* distribution function as,

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\mu, \sigma^2) &= \textstyle\prod_i^n (2\pi\sigma^2)^{-\frac{1}{2}} e^{\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} e^{\left(-\frac{1}{2\sigma^2}\sum_i^n (x_i-\mu)^2\right)}. \end{aligned} \tag{11}$$

But as you can imagine, if the probabilities are less than 1, then the product of a bunch of these is going to be **SUPER** small. It's not that big of a deal for the math, at least symbolically, but dealing with repeated multiplication of small things is *tedious* and *error prone*, for both humans and computers alike. Practically speaking, a computer has a limit on how small it can represent things and still be accurate.

Fortunately, we may, or may not, remember a special property of *logs*, that the *log* function can turn a product into sum—this will be illustrated below. So,

by taking the *log* of the probability function we can make the computation much easier while still keeping the same functional relations among the parameter in our original probability function.

1.2 Quick and dirty logs

Just a refresher, *logs* are meant to show the number of times a number, the *base*, is to be multiplied by itself to get a particular value. As the YouTuber ***Vihart*** put it, if we were counting in a "times the base sort of way",[1] how many steps would we need to go to get to the answer. So, the answer of the *log* function represents what power of the *base* is needed to get the input value. For example, if the base is 10, and input value is 10, then the answer of the *log* function is 1, because $10^1 = 10$, and so $log_{10}(10) = 1$. Additionally, counting in a "times ten" sort of way, how many steps to get to 100? The answer is 2.

In order to show some of the other properties of *logs* we will work with an easy example. We will use 100, which can be expressed the following **equivalent** ways.

$$\begin{aligned} 100 &= 10^2 \\ &= 10 \times 10 \\ &= 1000/10 \end{aligned} \tag{12}$$

So, let's work with *log* with a base of 10, this means we are interested in what power to raise 10 to in order to produce the result of 100.

$$\begin{aligned} \text{if} \quad & 10^2 = 100 \\ \text{then} \quad & log_{10}(100) = 2 \end{aligned} \tag{13}$$

As we can see, 2 is the answer for base 10. Below we present 3 of the basic properties of *logs*. These are not all of the properties, just the ones that are important for our illustration.

We assume base 10 for the following rules:

power rule

$log(A^n) = n \times log(A)$

- $log(10^2) = 2 \times log(10) = 2 \times 1 = 2$

product rule

$log(A \times B) = log(A) + log(B)$

- $log(10 \times 10) = log(10) + log(10) = 1 + 1 = 2$

quotient rule

$log(\frac{A}{B}) = log(A) - log(B)$

- $log(\frac{1000}{10}) = log(1000) - log(10) = 3 - 1 = 2$

[1] Check out Vihart's video "How I Feel About Logarithms" for a great explanation.

1.3 Log likelihood derivation

So, why does this matter? Well, because we are interested in fitting our previous function of the likelihood of a set of data, but we don't want to cause our computer to start to smoke computing very small numbers. If we take the *log* of the likelihood function we get another function that preserves our main properties, but that will also turn our product into a sum.

We will take the *log* of our joint probability specification in equation 11 above. In this case it is easiest to use a base of e for the log of the likelihood, or *natural log*, ln which equals log_e—so remember, this means $ln(e) = 1$. This makes the exponential part much easier to understand. Here are the steps for expressing the new log-likelihood function,

$$\begin{aligned} ln(f(x_1, x_2, ..., x_n|\mu, \sigma^2)) &= ln\left[(2\pi\sigma^2)^{-\frac{n}{2}} e^{\left(-\frac{1}{2\sigma^2}\sum_i^n (x_i-\mu)^2\right)}\right] \\ \text{by the } \textbf{product rule} &= ln\left[(2\pi\sigma^2)^{-\frac{n}{2}}\right] + ln\left[e^{\left(-\frac{1}{2\sigma^2}\sum_i^n (x_i-\mu)^2\right)}\right] \\ \text{by the } \textbf{power rule} &= \left[\left(-\tfrac{n}{2}\right) ln(2\pi\sigma^2)\right] + \left[\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_i^n (x_i-\mu)^2\right) ln(e)\right] \end{aligned}$$

simplify and we get

$$\mathcal{L}(X|\mu, \sigma^2) = -\left(\tfrac{n}{2}\right) ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\textstyle\sum_i^n (x_i-\mu)^2 \tag{14}$$

Minus 2 of the log of the likelihood

$$-2\mathcal{L}(X|\mu, \sigma^2) = n(ln(2\pi\sigma^2)) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{15}$$

2 Maximum Likelihood

2.1 Analytic solution

In this section we will work to solve for the specific parameters that will maximize our observed data. Again we are basing this on the distribution that we believe generated our data, in this case the Gaussian probability function. Below we need to solve for the parameters μ and σ^2 in terms of the observed data $X \in \{x_i, x_2, \ldots, x_n\}$.

To do this we will use calculus to find the maximum of the above function with regard to each parameter. First we will express the function in terms of the specific parameter, then take the derivative of the function with respect to the parameter to isolate its influence on the function overall. This step helps us understand how the function changes with respect to the parameter of interest. We set the partial derivative equal to zero and solve for the parameter to get where the changes in the function reach a maximum.

2.1.1 Partial derivative wrt μ

$$\begin{aligned}
\mathcal{L} \text{ wrt } \mu &= \tfrac{1}{\sigma^2} \textstyle\sum_{i=1}^{n} (x_i - \mu)^2 \\
&= \tfrac{1}{\sigma^2} \left[\textstyle\sum_{i=1}^{n} x_i^2 - 2x_i\mu + \mu^2 \right] \\
&= \tfrac{1}{\sigma^2} \left[\textstyle\sum_{i=1}^{n} x_i^2 - 2 \textstyle\sum_{i=1}^{n} x_i\mu + \textstyle\sum_{i=1}^{n} \mu^2 \right] \\
&= \tfrac{1}{\sigma^2} \left[\textstyle\sum_{i=1}^{n} x_i^2 - 2\mu \textstyle\sum_{i=1}^{n} x_i + n\mu^2 \right] \\
\tfrac{\partial \mathcal{L}}{\partial \mu} &= \tfrac{1}{\sigma^2} \left[-2 \textstyle\sum_{i=1}^{n} x_i + 2n\mu \right]
\end{aligned} \tag{16}$$

Set the result equal to zero and solve

$$\begin{aligned}
0 &= \tfrac{1}{\sigma^2} \left[-2 \textstyle\sum_{i=1}^{n} x_i + 2n\mu \right] \\
&= -2 \textstyle\sum_{i=1}^{n} x_i + 2n\mu \\
2 \textstyle\sum_{i=1}^{n} x_i &= 2n\mu \\
\tfrac{\sum_{i=1}^{n} x_i}{n} &= \mu
\end{aligned} \tag{17}$$

2.2 Partial derivative wrt σ^2

Substitute $\sigma^2 = u$

$$\begin{aligned}
\mathcal{L} \text{ wrt } u &= n(ln(2\pi u)) + \tfrac{1}{u} \textstyle\sum_{i=1}^{n} (x_i - \mu)^2 \\
&= n \left[ln(2\pi) + ln(u) \right] + \tfrac{1}{u} \textstyle\sum_{i=1}^{n} (x_i - \mu)^2 \\
&= nln(2\pi) + nln(u) + \tfrac{1}{u} \textstyle\sum_{i=1}^{n} (x_i - \mu)^2 \\
\tfrac{\partial \mathcal{L}}{\partial u} &= n \left(\tfrac{1}{u} \right) - \tfrac{1}{u^2} \textstyle\sum_{i=1}^{n} (x_i - \mu)^2
\end{aligned} \tag{18}$$

Set the result equal to zero and solve

$$\begin{aligned}
0 &= n\left(\tfrac{1}{u}\right) - \tfrac{1}{u^2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2 \\
\tfrac{1}{u^2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2 &= n\left(\tfrac{1}{u}\right) \\
\tfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{u^2} &= \tfrac{n}{u} \\
\textstyle\sum_{i=1}^{n}(x_i-\mu)^2 &= \tfrac{nu^2}{u} \\
\textstyle\sum_{i=1}^{n}(x_i-\mu)^2 &= nu \\
\tfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} &= u \\
&\text{back substitute for } \sigma^2 \\
\tfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} &= \sigma^2
\end{aligned} \tag{19}$$

3 Bernoulli Likelihood

Here we will see how to do the same process for a discrete variable. In this case we assume that this variable follows the Bernoulli distribution in which x can only take on two values $x_i \in \{0, 1\}$ specified as,

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{20}$$

Joint Bernoulli Likelihood

For a given set of such observations, $X = \{x_1, x_2, \ldots, x_n\}$, we get the joint distribution specified as,

$$p(X|\mu) = \prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{(1-x_i)}, \tag{21}$$

which directly translates to,

$$\begin{aligned}
p(X|\mu) &= \textstyle\prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{(1-x_i)} \\
&= \mu^{x_1}(1-\mu)^{(1-x_1)} + \mu^{x_2}(1-\mu)^{(1-x_2)} + \ldots + \mu^{x_n}(1-\mu)^{(1-x_n)} \quad .
\end{aligned} \tag{22}$$

First portion

We can see here, as with the Gaussian likelihood (see equation 10), that we have a based raised to a variable exponent. Thus, by the rule of exponents in

equation 8 we can translate this into,

$$\begin{aligned} \prod_{i=1}^{n} \mu^{x_i} &= \mu^{x_1} \times \mu^{x_2} \times \ldots \times \mu^{x_n} \\ &= \mu^{x_1+x_2+\ldots+x_n} \\ &= \mu^{\sum_{i=1}^{n} x_i} \end{aligned} \tag{23}$$

Second portion

Following the same rules as with the first portion, the second portion becomes,

$$\prod_{i=1}^{n} (1-\mu)^{(1-x_i)} = (1-\mu)^{\sum_{i=1}^{n}(1-x_i)}. \tag{24}$$

Recombining the two portions into the full joint likelihood for Bernoulli distributed data can be expressed as,

$$p(X|\mu) = \mu^{\sum_{i=1}^{n} x_i} (1-\mu)^{\sum_{i=1}^{n}(1-x_i)} \tag{25}$$

Loglikelihood of a Bernoulli distribution

As before, we can take the log of this joint distribution to simplify things.

$$\begin{aligned} ln(p(X|\mu)) &= ln[\mu^{\sum_{i=1}^{n} x_i} (1-\mu)^{\sum_{i=1}^{n}(1-x_i)}] \\ \text{by the } \textbf{product rule} &= ln\left[\mu^{\sum_{i=1}^{n} x_i}\right] + ln\left[(1-\mu)^{\sum_{i=1}^{n}(1-x_i)}\right] \\ \text{by the } \textbf{power rule} &= [\textstyle\sum_{i=1}^{n} x_i ln(\mu)] + [\sum_{i=1}^{n}(1-x_i) ln(1-\mu)] \end{aligned} \tag{26}$$

Thus the loglikelihood of the joint Bernoulli distribution is expressed as,

$$\mathcal{L}(X|\mu) = ln(\mu) \sum_{i=1}^{n} x_i + ln(1-\mu) \sum_{i=1}^{n} (1-x_i) \tag{27}$$

Partial derivative

From the equation in 27, we can take the partial derivative in order to solve for our parameter of interest μ.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu} &= \tfrac{1}{\mu} \textstyle\sum_{i=1}^{n} x_i + \frac{1}{(1-\mu)}(-1) \sum_{i=1}^{n}(1-x_i) \\ &= \tfrac{1}{\mu} \textstyle\sum_{i=1}^{n} x_i - \frac{1}{(1-\mu)} \sum_{i=1}^{n}(1-x_i) \end{aligned} \tag{28}$$

Next we set equation 28 to zero and solve for μ,

$$\begin{aligned}
0 &= \frac{1}{\mu}\sum_{i=1}^{n} x_i - \frac{1}{(1-\mu)}\sum_{i=1}^{n}(1-x_i) \\
\frac{1}{\mu}\sum_{i=1}^{n} x_i &= \frac{1}{(1-\mu)}\sum_{i=1}^{n}(1-x_i) \\
\frac{(1-\mu)}{1} \times \frac{1}{\mu} &= \sum_{i=1}^{n}(1-x_i) \times \left(\frac{1}{\sum_{i=1}^{n} x_i}\right) \\
\frac{(1-\mu)}{\mu} &= \frac{\sum_{i=1}^{n}(1-x_i)}{\sum_{i=1}^{n} x_i} \\
\frac{1}{\mu} - 1 &= \frac{\sum_{i=1}^{n} 1}{\sum_{i=1}^{n} x_i} - 1 \\
\frac{1}{\mu} &= \frac{\sum_{i=1}^{n} 1}{\sum_{i=1}^{n} x_i} \\
\frac{1}{\mu} &= \frac{n}{\sum_{i=1}^{n} x_i}
\end{aligned} \tag{29}$$

Thus,

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{30}$$

4 Minimizing a loss function

In this section we will explore the use of this same basic approach to solve for the parameters in a different equation. In this case we will look to solve for the parameters of a linear model expressed as,

$$y = b_0 + b_1 x, \tag{31}$$

where b_0 is the y intercept and b_1 is the slope.

Our goal here is to solve for each of these parameters and express these solutions in terms of our observed data $y_i \in \{y_1, y_2, \ldots, y_n\}$ and $x_i \in \{x_1, x_2, \ldots, x_n\}$.

4.1 Loss function

Now, if we are interested in solving for the parameters in equation 31, we need observations. If we have as many observations as we have parameters to solve, in our models there are 2, b_0 and b_1 respectively, there is assumed to be one unique solution that will satisfy the equation. This is exactly the same task as solving simultaneous equation from Algebra, keep in mind that each observation is thought to follow the same model in equation 31. For example if we had two observations $(x_1, y_1) = (0, 3)$ and $(x_2, y_2) = (1, 3.5)$ we could construct the following lines,

$$\begin{aligned}
3 &= b_0 + b_1 0 \\
3.5 &= b_0 + b_1 1
\end{aligned}. \tag{32}$$

Solving for each parameter above we would find that $b_0 = 3$ and $b_1 = .5$.

The situation becomes more difficult when we have more observations than we have parameters. In practice this is most common, but what we need to do is find a solution that is "close enough" to the data that we are satisfied. We will call the values that are computed from this "close enough" solution our estimates and designate them as $\hat{y}_i$. These values represent a direct application of our model in equations 31,

$$\hat{y}_i = b_0 + b_1 x_{1i} \tag{33}$$

Our estimates however are not perfect, and most times will be off by some amount. This amount will be different for each estimate we generate, so it's helpful to think about capturing this error. We will define our error term as,

$$\begin{aligned} y_i &= \hat{y}_i + \epsilon_{y_i} \\ \epsilon_{y_i} &= y_i - \hat{y}_i \end{aligned} \quad . \tag{34}$$

Since our goal is to find the parameters, b_0 and b_1 that are "close enough" we will try to minimize the errors expressed in equation 34. Specifically, we will look to minimize the variance of these errors.

4.1.1 Error variance

To compute the variance of the errors we need to start with the sum of squared errors,

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{35}$$

Substituting equation 31 for our $\hat{y}_i$ term we get,

$$\begin{aligned} \textstyle\sum_{i=1}^{n} \epsilon_i^2 &= \textstyle\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \\ &= \textstyle\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(y_i - b_0 - b_1 x_i) \\ &= \textstyle\sum_{i=1}^{n} (y_i^2 - b_0 y_i - b_1 x_i y_i) \\ &\quad + (-y_i b_0 + b_0^2 + b_0 b_1 x_i) \\ &\quad + (-y_i b_1 x_i + b_0 b_1 x_i + b_1^2 x_i^2) \\ &= \textstyle\sum_{i=1}^{n} (y_i^2 + b_0^2 + b_1^2 x_i^2) \\ &\quad + (-y_i b_0 - y_i b_0) \\ &\quad + (-b_1 x_i y_i - b_1 x_i y_i) \\ &\quad + (b_0 b_1 x_i + b_0 b_1 x_i) \\ &= \textstyle\sum_{i=1}^{n} (y_i^2 + b_0^2 + b_1^2 x_i^2 - 2 y_i b_0 - 2 b_1 x_i y_i + 2 b_0 b_1 x_i) \end{aligned} \tag{36}$$

We distribute the summation operator to get,

$$\sum_{i=1}^{n}(y_i^2 + b_0^2 + b_1^2 x_i^2 - 2y_i b_0 - 2b_1 x_i y_i + 2b_0 b_1 x_i)$$

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} b_0^2 + \sum_{i=1}^{n} b_1^2 x_i^2 \\ - \sum_{i=1}^{n} 2y_i b_0 - \sum_{i=1}^{n} 2b_1 x_i y_i + \sum_{i=1}^{n} 2b_0 b_1 x_i\end{aligned} \tag{37}$$

$$\begin{aligned}\sum_{i=1}^{n}(y_i^2) + nb_0^2 + b_1^2 \sum_{i=1}^{n}(x_i^2) \\ - 2b_0 \sum_{i=1}^{n}(y_i) - 2b_1 \sum_{i=1}^{n}(x_i y_i) + 2b_0 b_1 \sum_{i=1}^{n}(x_i)\end{aligned}$$

We will simplify the above expression by making the following substitutions,

- $SS_y = \sum_{i=1}^{n}(y_i^2)$
- $SS_x = \sum_{i=1}^{n}(x_i^2)$
- $SCP_{xy} = \sum_{i=1}^{n}(x_i y_i)$
- $SS_\epsilon = \sum_{i=1}^{n} \epsilon_i^2$.

Using these definitions our expression in equation 37 becomes,

$$SS_\epsilon = SS_y + nb_0^2 + b_1^2 SS_x - 2b_0 \sum_{i=1}^{n}(y_i) - 2b_1 SCP_{xy} + 2b_0 b_1 \sum_{i=1}^{n}(x_i). \tag{38}$$

4.2 Solving for the parameters

From here we will perform our partial differentiation in order to solve for the parameters b_0 and b_1.

4.2.1 The Intercept b_0

$$SS_e \text{ w.r.t. } b_0 = nb_0^2 - 2b_0 \sum_{i=1}^{n}(y_i) + 2b_0 b_1 \sum_{i=1}^{n}(x_i) \tag{39}$$

Next we take the partial derivative of the above with respect to b_0, set it to zero and solve for b_0,

$$\begin{aligned}\frac{\partial SS_e}{\partial b_0} &= 2nb_0 - 2\sum_{i=1}^{n}(y_i) + 2b_1 \sum_{i=1}^{n}(x_i) \\ 0 &= 2nb_0 - 2\sum_{i=1}^{n}(y_i) + 2b_1 \sum_{i=1}^{n}(x_i) \\ 2nb_0 &= 2\sum_{i=1}^{n}(y_i) - 2b_1 \sum_{i=1}^{n}(x_i) \\ b_0 &= \frac{2\sum_{i=1}^{n}(y_i) - 2b_1 \sum_{i=1}^{n}(x_i)}{2n} \\ b_0 &= \frac{\sum_{i=1}^{n}(y_i)}{n} - b_1 \frac{\sum_{i=1}^{n}(x_i)}{n}\end{aligned} \tag{40}$$

We can use the fact that $\bar{x} = \frac{\sum_{i=1}^{n}(x_i)}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n}(y_i)}{n}$, thus turning the result of equation 40 into,

$$b_0 = \bar{y} - b_1 \bar{x} \tag{41}$$

4.2.2 The Slope b_1

$$SS_e \text{ w.r.t. } b_1 = b_1^2 SS_x - 2b_1 SCP_{xy} + 2b_0 b_1 \sum_{i=1}^{n}(x_i). \tag{42}$$

Next we take the partial derivative of the above with respect to b_1, set it to zero and solve for b_1.

$$\frac{\partial SS_e}{\partial b_1} = 2b_1 SS_x - 2SCP_{xy} + 2b_0 \sum_{i=1}^{n}(x_i) \tag{43}$$

Before we do the algebra to find the solution, we will take advantage of our solution for b_0 in equation 41 and use this to transform all terms in our expression in equation 43 to be expressed as functions of the b_1 parameter,

$$\begin{aligned} 0 &= 2b_1 SS_x - 2SCP_{xy} + 2(\tfrac{\sum_{i=1}^{n}(y_i)}{n} - b_1 \tfrac{\sum_{i=1}^{n}(x_i)}{n}) \textstyle\sum_{i=1}^{n}(x_i) \\ &= 2b_1 SS_x - 2SCP_{xy} + \tfrac{2}{n} \textstyle\sum_{i=1}^{n}(y_i x_i) - b_1 \tfrac{2}{n} \textstyle\sum_{i=1}^{n}(x_i x_i) \quad . \\ &= 2b_1 SS_x - 2SCP_{xy} + \tfrac{2}{n} SCP_{xy} - b_1 \tfrac{2}{n} SS_x \end{aligned} \tag{44}$$

Move all common terms to one side

$$-2b_1 SS_x + b_1 \tfrac{2}{n} SS_x = -2SCP_{xy} + \tfrac{2}{n} SCP_{xy} \quad . \tag{45}$$

Next, multiply through by $-\frac{1}{2}$

$$\begin{aligned} b_1 SS_x - b_1 \tfrac{1}{n} SS_x &= SCP_{xy} - \tfrac{1}{n} SCP_{xy} \\ b_1 (SS_x - \tfrac{1}{n} SS_x) &= (SCP_{xy} - \tfrac{1}{n} SCP_{xy}) \\ b_1 [SS_x (1 - \tfrac{1}{n})] &= [SCP_{xy} (1 - \tfrac{1}{n})] \\ b_1 &= \tfrac{SCP_{xy}(1-\frac{1}{n})}{SS_x(1-\frac{1}{n})} \\ b_1 &= \tfrac{SCP_{xy}}{SS_x} \times \tfrac{(1-\frac{1}{n})}{(1-\frac{1}{n})} \\ b_1 &= \tfrac{SCP_{xy}}{SS_x} \end{aligned} \tag{46}$$

Thus our maximum likelihood estimates for the parameters b_0 and b_1 from equation 31 are,

$$b_0 = \bar{y} - b_1 \bar{x} \tag{47}$$

and

$$b_1 = \frac{SCP_{xy}}{SS_x} \tag{48}$$