1 The Simple Linear Regression Model

1.1 Introduction

Suppose we have a sample set (X, Y) of size m, that is $(X, Y) = \{(x_1, y_1), (x_2, y_2) \cdots (x_m, y_m)\}$. Then a simple linear model assumes a linear relationship between variables (x_i, y_i), and tries to estimate that. For example, observe a sample scatter plot of sample set in Figure 1. By looking at the figure, one could intuitively guess a linear relation between x and y variables as y increasing roughly with x. It is this we will try to find, and in that, find the best possible one.

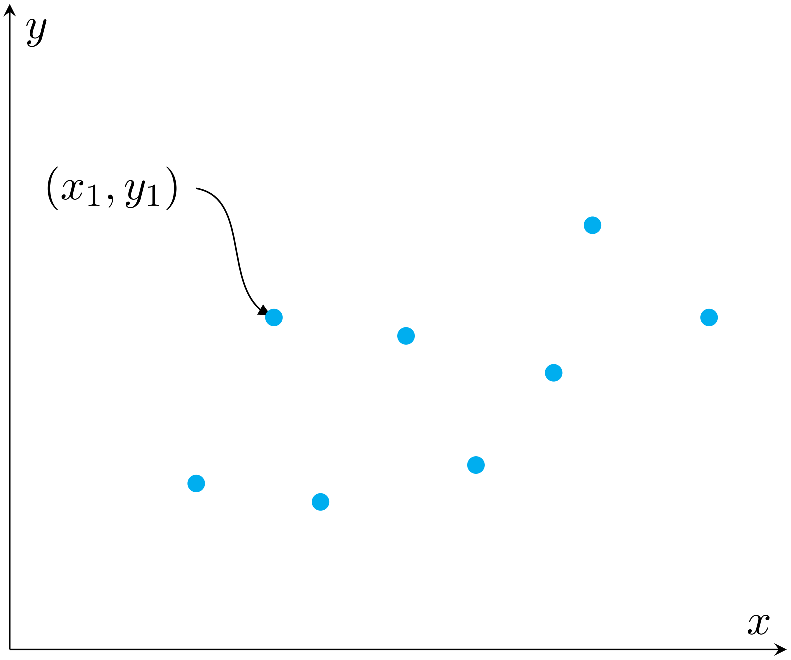

Fig 1: Given Sample Set

We will find a line that passes through these points, there by being the best line, that has minimum vertical or Δy distance from all the sample points. Typically such a line would be unique to given any sample set and it is the **best fit** line possible. Figure 2 shows such a *potential* line. The vertical difference Δy_1 as shown in figure, is the distance between the point (x_1, y_1) and the line.

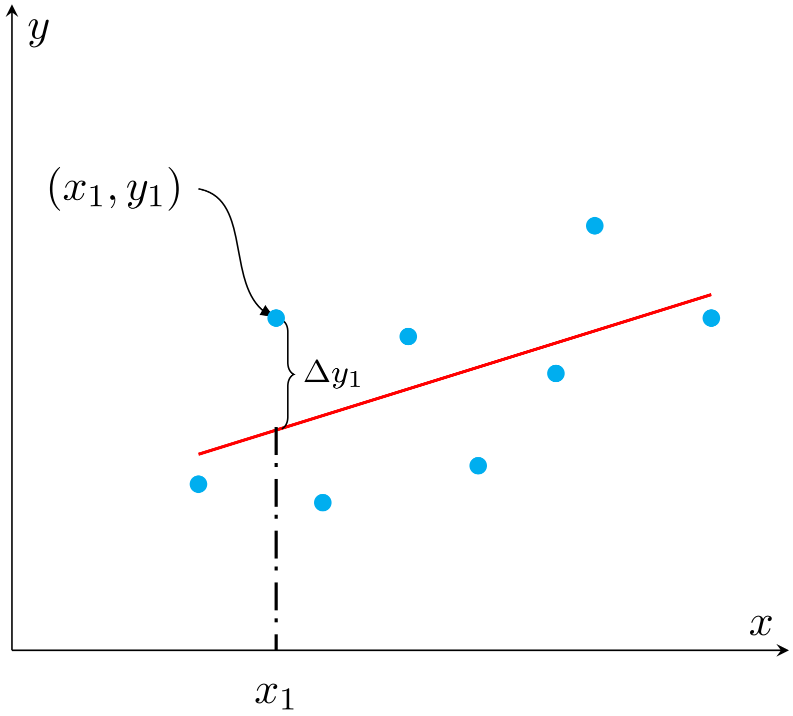

Fig 2: Finding a best-fit line is the goal

When a sample set is given, we will assume such a line exists and that ideally, all sample points should have fallen on that line, implying a perfect linear relationship between x and y. However, because of an **underlying error** ε, the sample points have fallen apart, around the line, giving us the sample set. Suppose, such a perfect linear relationship exists ideally, let us say, it could be defined as below by using a regular line equation with slope β_1 and y-intercept β_0, as

$$y = \beta_0 + \beta_1 x$$

Thus in this ideal world, y is completely deterministic from x. However, when we introduce randomness in the form or error ε, the y value also becomes a random variable associated with the randomness from ε. That is, if we describe such a RV as Y, then

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

We do not know ε. Naturally, the *expectation* of the error is to be zero, or in other words we assume, though there is room for error, but the zero error has maximum probability. Thus assuming a normal distribution of $N(0, \sigma^2)$,

$$E(\varepsilon) = 0 \quad Var(\varepsilon) = \sigma^2 \tag{2}$$

Assumptions and Approach

- Given the sample set value, we will imagine there to be an *ideal* linear relationship, and try to find that hypothetical line which will have *least error* for all observed sample points.
- Also assume that error has maximum probability to be 0, and normally distributed, formulating as a cause for observing sample values as observed instead of, on the line where error would have been zero.

This line of thought is important and fundamental to our model. Because of this assumption, we could now say, the points should have ideally sat on the line, but resulted in their places in reality as we find them, because of the error. Thus the observed y value is the result of the error ε, while its **expected y value** $E(Y|x)$ **or** $\mu_{Y.x_1}$, should sit on the line. This is illustrated in Figure 3. Similarly the only randomness comes from error ε, so its variance directly transfers to the Y random variable due to 2. That is, $\sigma_{Y.x_1} \to \sigma$.

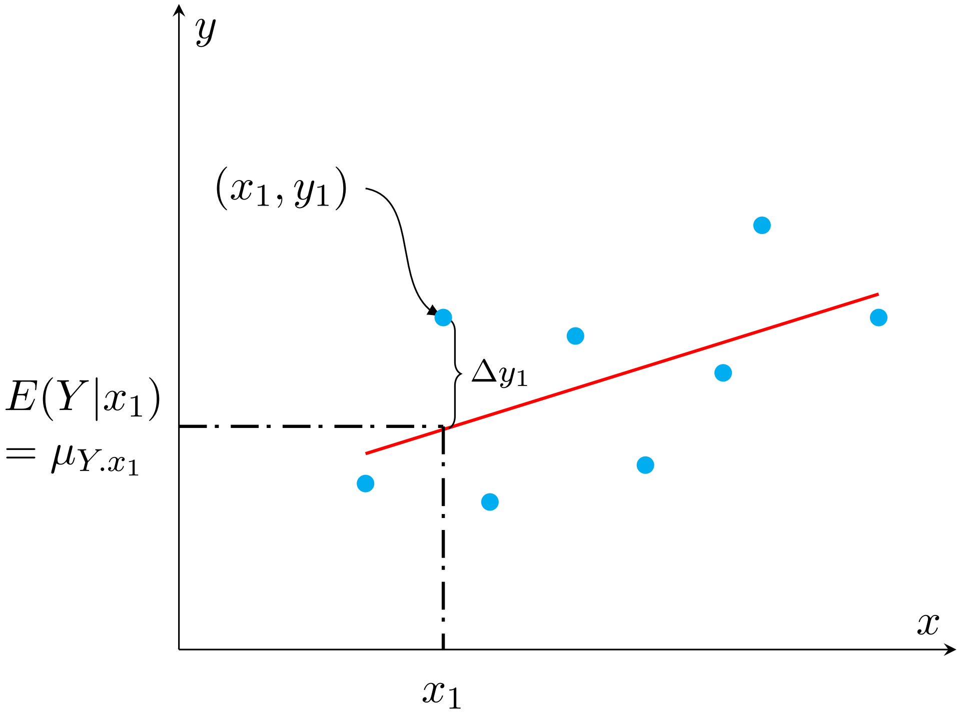

Fig 3: $y1$ and $E(Y|x_1)$

We could also prove them mathematically as below. For any point (x_1, y_1)

$$\mu_{Y.x_1} = E(Y|x_1) = E(\beta_0 + \beta_1 x_1 + \varepsilon) = E(\beta_0) + E(\beta_1 x_1) + E(\varepsilon)$$

If a is a constant observed, then $E(a) = a$ only as its the only value and already observed. And since $\varepsilon = N(0, \sigma^2)$ we could write,

$$\mu_{Y.x_1} = E(Y|x_1) = \beta_0 + \beta_1 x_1$$

Similary,

$$\sigma^2_{Y.x_1} = Var(Y|x_1) = Var(\beta_0 + \beta_1 x_1 + \varepsilon)$$

If a is a constant observed, then $Var(a) = 0$ only as its already observed and there is no uncertainty. And since $\varepsilon = N(0, \sigma^2)$ we could write,

$$\sigma^2_{Y.x_1} = Var(Y|x_1) = 0 + 0 + \sigma^2$$

Thus, in general for any x, in continuous scale, we could say,

$$\mu_{Y.x} = E(Y|x) = \beta_0 + \beta_1 x$$
$$\sigma^2_{Y.x} = Var(Y|x) = \sigma^2$$

Note, though our sample values are discrete, we are able to get a line at continuous scale, because its the ideal situation, where all the expected values should lie on that hypothetical line $y = \beta_0 + \beta_1 x$. So this line should stay true for any value of x. It is a hypothetical line of expected or mean values $E(Y|x)$, so understandably, its called **line of mean values**. It should also have been the ideal line, where all sample points should have rested, provided there were no errors. So this line is also called **True regression line**.

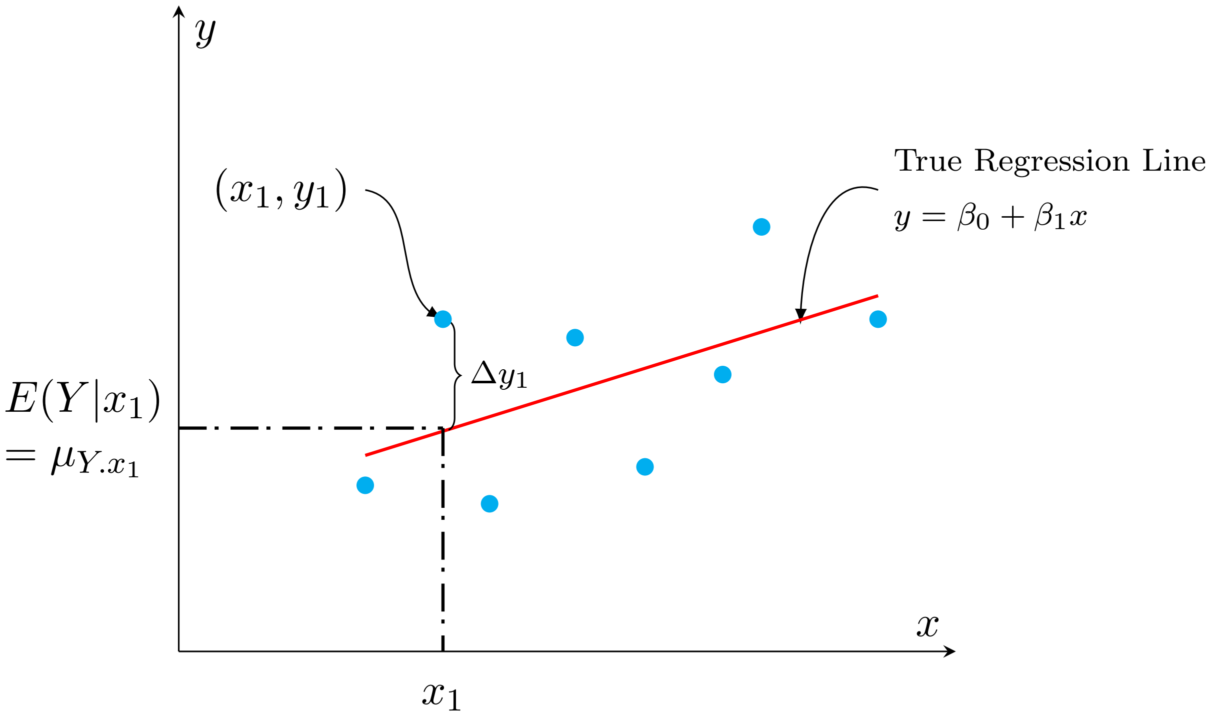

Fig 4: $\beta_0 + \beta_1 x$ is the ideal hypothetical line with no error

Expected value and Variance of Y given a sample x^*

For any observed value (x^*, y^*),

$$\begin{aligned} \mu_{Y.x^*} &= E(Y|x^*) = \beta_0 + \beta_1 x^* \\ \sigma^2_{Y.x^*} &= Var(Y|x^*) = \sigma^2 \end{aligned} \tag{3}$$

In continuous scale, for any (x, y),

$$\begin{aligned} \mu_{Y.x} &= E(Y|x) = \beta_0 + \beta_1 x \\ \sigma^2_{Y.x} &= Var(Y|x) = \sigma^2 \end{aligned} \tag{4}$$

It is difficult to visualize the error randomness (say, its pdf) in the x, y graph as ε is another 3rd variable hidden underneath. However we just saw, how that distribution transfers to the random variable Y. If ε has $N(0, \sigma^2)$, then Y has distribution $N(\beta_0 + \beta_1 x, \sigma^2)$. This facilitates us to view the randomness on the face of random variable Y as shown in 5. Observe that, for a point, say (x_1, y_1), for the given x_1, ideally, y should have been the mean value $E(Y|x_1) = \beta_0 + \beta_1 x$, that has the highest probability of the normal distribution. That is our assumption and then we say, because there exists an error, we got y at y_1. Note for the sample location y_1, the error is low, but still had a chance.

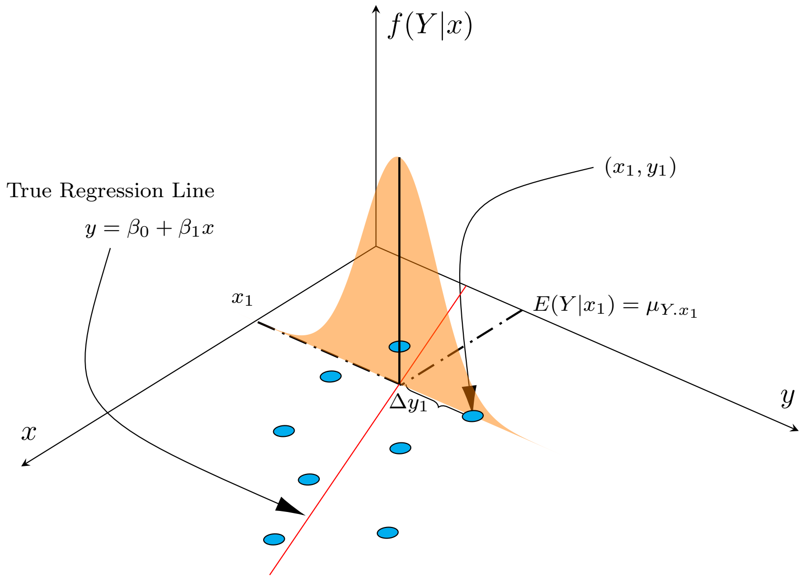

Fig 5: The Probability Distribution $f(Y|x_1)$

The distance how much the *erroneous* locations of sample points spread out from the mean value is determined by variance σ of the error. Note that, we assume this error is constant for all sample values. This means, any point x_m, y_m has same probability distribution of committing an error, as any other point in the sample set. This assumed property is called **Homoscedasticity**. If this is not the case, then the characteristic is called **Heteroscedasticity**. One could fairly assume from given a sample set, if the underlying error could be Homoscedastic or Heteroscedastic, by eyeballing at the spread from the regression line. We will focus and assume Homoscedasticity and for any one interested, **?**] has written an interesting article about dealing with the same. Given that Homoscedasticity is assumed, the probability distribution would be uniform across the regression line. This is illustrated in 6. That is, for any x value, the equivalent $f(Y|x)$ could be picked up like a card from a stack. This distribution across the regression line could be continuous or discrete, depending on x is continuous or discrete.

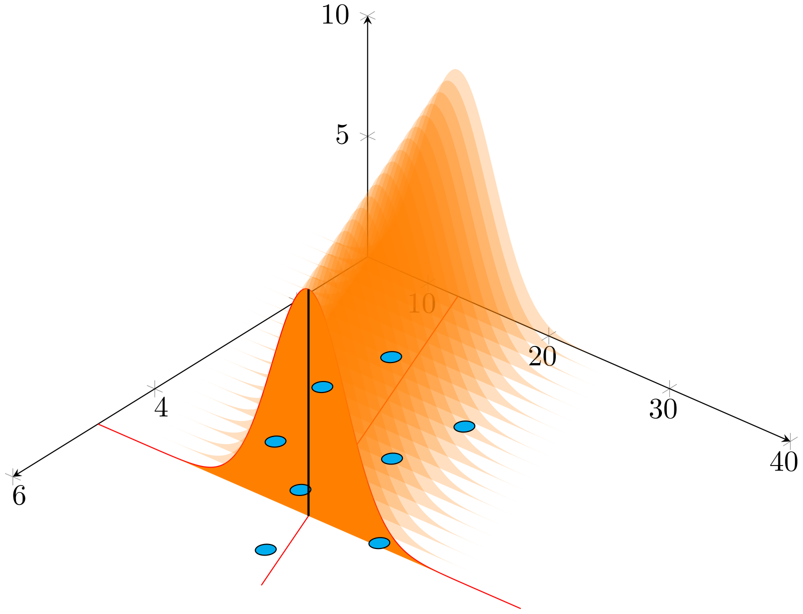

Fig 6: The *pdf* $f(Y|x)$ is continuous or discrete along the regression line depending on x is continuous or discrete

Now that our sample set is discrete, let us focus on that. We need to find out, for given sample set, what would be the optimal values of β_0 and β_1.

1.2 Estimating Model Parameters

The goal is to find (β_0, β_1) such that, the resulting line is some how "best-fit" among all possible lines of $E(Y|x)$. You see, our sample set could be a part of a bigger population, and thus the hypothetical line for entire population could be anything. However, we have only a sample set, so our best bet is always what is the best representative of the sample. That is, **given the samples**, what would be the best representative regression line is what our goal is. Imagine, if all sample lines, line up in a certain way, then our best bet would be just a line cutting across all those points. This suggests, all sample points have zero error, or have fallen at their respective highest probability mean locations, thus one could expect any more new sample to take a similar place on that line. Note that in this case, all lines are at *zeroth distance* from the mean line. This is illustrated in 7 where the vertical red dotted line represents maximum probability.

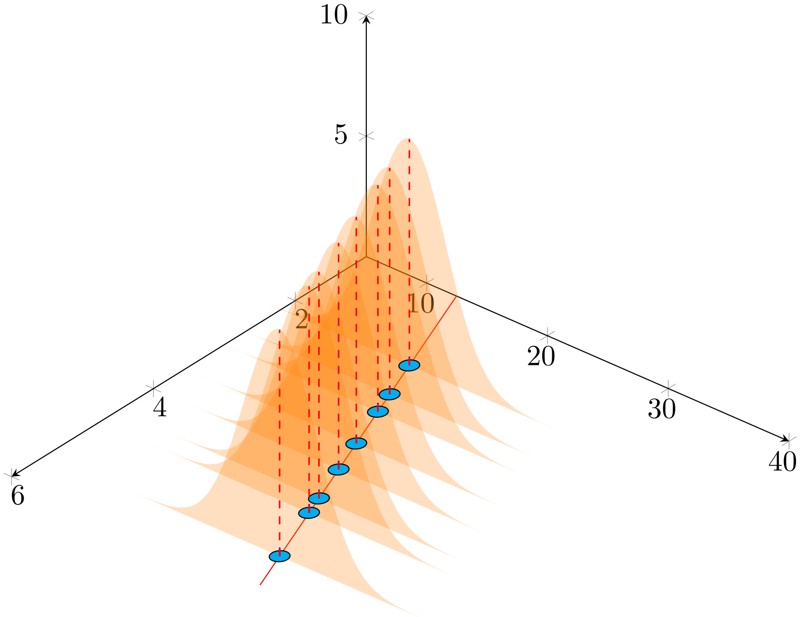

Fig 7: An ideal case

Now when the samples deviate from such a hypothetical mean line, best bet then to find the mean line is to find one, that has *least distance* from all the sample points. The sum of all the distances from all sample points to that line would be minimal compared to any other lines' similar sum of distances. The distances are illustrated in 8, where blue lines indicate the actual distance from the true regression line. Now, naturally, since the points could lie on either side of the line, would give rise to relatively positive or negative distances, and thus cancelling each others' distances out partly here and there. To avoid that, one could take absolute distances from the point to the line.

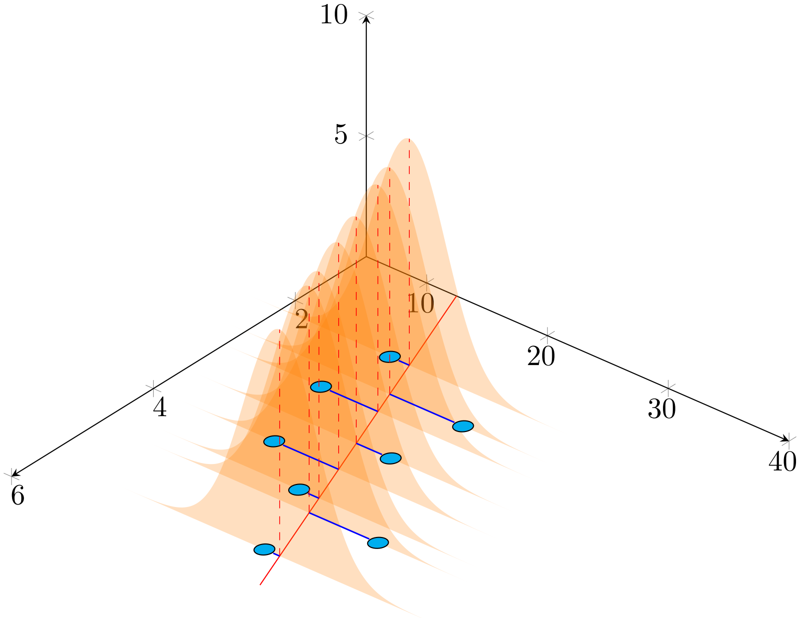

Fig 8: A practical case

Principle of Least Squares

However, instead of taking the absolute distances, we now, out of nowhere(?!) choose to take the square of the calculated distance and sum up to find the total distance. As per my current understanding, this was mearly a choice for algebraic convenience [1]. We also have other ways of measuring approaches (angled distance instead of vertical etc) but we shall not get in to it as this is only Simple Regression Model.

Now that we have fixated on finding the least sum of squares of the distances (note because we squared, there was no absoluteness to be considered in equation), let us look in to the mathematical form of it. This principle which can be traced back to famous mathematician Guass, says that, a line provides a good fit to the data if the vertical distances (deviations) from the observed points to the line are small. The measure of the goodness of fit is the sum of the squares of these deviations. The best-fit line is then the one having the smallest possible sum of squared deviations.

Principle of Least Squares (from **?**])

The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1 x$ is

$$\text{height of point - height of line} = y_i - (b_0 + b_1 x_i)$$

The sum of squared vertical deviations from the points $(x_1, y_1), \cdots, (x_m, y_m)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \tag{5}$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least square estimates**, are those values that minimize $f(b_0, b_1)$. That is $(\hat{\beta}_0, \hat{\beta}_1)$ are such that, $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ for any $(b_0.b_1)$. The **estimated regression line or least squares line** is then the line whose equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6}$$

Note that 6 is same as expected mean line or true regression line as expressed in 4. Here we just devised a way to find those optimal (β_0, β_1).

Using Maximum Likelihood Estimation

We could also arrive at 5 via Maximum Likelihood Estimation (which was the reason we had entire chapter on MLE before regression in first place). Recall each sample point as shown on figure 8, has the pdf $f(Y|x) = N(\beta_0 + \beta_1 x, \sigma^2)$. Then, as per MLE, we would like to know what is the joint probability of all these samples points to be at their observed locations. It will be useful to recall MLE derivation for Normal distribution as we saw in **??**. In similar fashion, for each sample point, the pdf could be written as,

$$f(Y|x_i; \beta_0, \beta_1) = N(\beta_0 + \beta_1 x_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \right\}$$

And as usual, assuming all these sample points are *independent and identically distributed*, we could arrive at their likelihood function as

[1] http://www.bradthiessen.com/html5/docs/ols.pdf

$$
\begin{aligned}
L(\beta_0, \beta_1) &= f(Y|x_1; \beta_0, \beta_1, Y|x_2; \beta_0, \beta_1, \cdots Y|x_m; \beta_0, \beta_1) \\
&= f(Y|x_1; \beta_0, \beta_1) f(Y|x_2; \beta_0, \beta_1) \cdots f(Y|x_m; \beta_0, \beta_1) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \Big\} \\
&= \Bigg\{ \Big(\frac{1}{2\pi\sigma^2} \Big)^{\frac{m}{2}} \Bigg\} \Bigg\{ \prod_{i=1}^{m} \exp\Big\{ - \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \Big\} \Bigg\} \\
&= \big(2\pi\sigma^2 \big)^{\frac{-m}{2}} \Bigg\{ \exp\Big\{ - \frac{\sum_{i=1}^{m} [y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \Big\} \Bigg\}
\end{aligned}
$$

Note, using product rule of logarithms, for any function $f = p^a e^b$,

$$ln(p^a e^b) = aln(p) + b$$

Thus, taking natural logarithm on both sides of likelihood function,

$$ln(L(\beta_0, \beta_1)) = -\frac{m}{2} \big(ln(2\pi\sigma^2) \big) - \frac{\sum_{i=1}^{m} [y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \tag{7}$$

The function 7 is a function of two variables (β_0, β_1), thus graphically represents a 3D surface plot as shown in figure 9, with height of the surface at any point is the function value evaluated at that point. We need to find out a point on this surface, where the function reaches maximum. The value of (β_0, β_1) at that point represents optimal values $(\hat{\beta}_0, \hat{\beta}_1)$. Why? Because, associated with those points, is the probability density function that yields maximum probability of getting all those sample sets in the places they are observed.

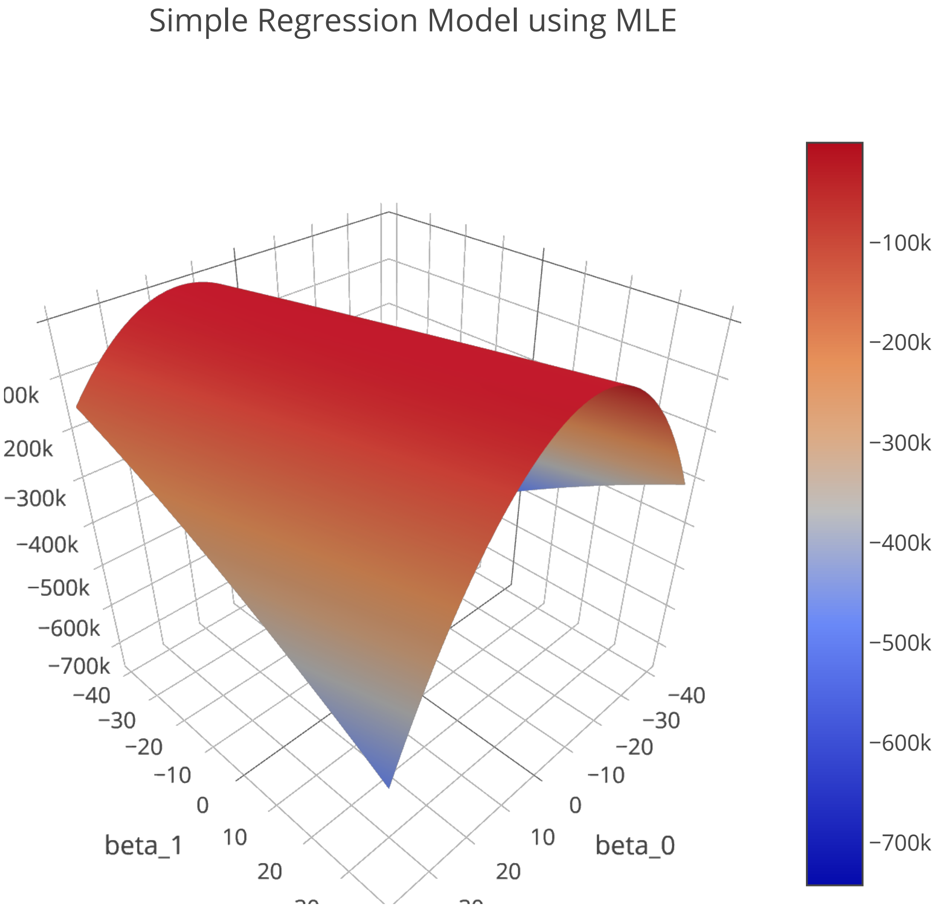

Fig 9: Log Likelihood function of given sample set

MLE leads to OLS

Before we find the optimal points, note that equation 7 has the variables (β_0, β_1) in the second term of RHS, and thus it is on that term we would be operating upon to find the optimal value. That is, when we derive w.r.t. (β_0, β_1) , the first term on RHS is a constant so goes away and constants in 2nd term too, would not offer any information, which we will see shortly, due to which we would just be equating the numerator of 2nd term RHS, to find the optimal value. That is, let

$$H(\beta_0, \beta_1) = \sum_{i=1}^{m} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{8}$$

then, by attempting to find the critical points of log likelihood $lnL(\beta_0, \beta_1)$ of given sample set, we would essentially operate upon $H(\beta_0, \beta_1)$. Note that this $H(\beta_0, \beta_1)$ is exactly equivalent to the ordinary least squares equation we saw in 5.

Derivation

To find the critical points on the surface (which could be maximum or minimum or saddle point), let us take first order partial derivatives and equate to 0. For details on why we do this, refer appendix **??** where we have shortly explained the concept behind using derivatives for finding critical points.

Keeping β_0 as constant and taking partial derivative with respect to β_1, we get,

$$\begin{aligned}\frac{\partial ln(\beta_0, \beta_1)}{\partial \beta_1}\bigg|_{\beta_0 = k} &= 0 - 2\Big\{\frac{\sum_{i=1}^{m}[y_i - (\beta_0 + \beta_1 x_i)(-x_i)]}{2\sigma^2}\Big\} \\ &= \frac{1}{\sigma^2}\Big\{\sum_{i=1}^{m}[y_i - \beta_0 - \beta_1 x_i](x_i)\Big\}\end{aligned}$$

Keeping β_1 as constant and taking partial derivative with respect to β_0, we get,

$$\begin{aligned}\frac{\partial ln(\beta_0, \beta_1)}{\partial \beta_0}\bigg|_{\beta_1 = k} &= 0 - 2\Big\{\frac{\sum_{i=1}^{m}[y_i - (\beta_0 + \beta_1 x_i)]}{2\sigma^2}\Big\}(-1) \\ &= \frac{1}{\sigma^2}\Big\{\sum_{i=1}^{m}[y_i - \beta_0 - \beta_1 x_i]\Big\}\end{aligned}$$

Equating both to 0, we get, (note, now the paramters are $(\hat{\beta}_0, \hat{\beta}_1)$) because they are the optimal *values* we are going to find out by equating to 0.

$$\sum_{i=1}^{m}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] = 0 \tag{9}$$

$$\sum_{i=1}^{m}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]x_i = 0 \tag{10}$$

Due to repeated use, for a while, let $\sum\limits_{i=1}^{m} \implies \sum_i$.

We know $\overline{x} = \dfrac{1}{m}\sum_i x_i$, and $\overline{y} = \dfrac{1}{m}\sum_i y_i$. Thus,

$$\sum_i x_i = m\overline{x} \tag{11}$$

$$\sum_i y_i = m\overline{y} \tag{12}$$

Substituting in 9,

$$\begin{aligned}\sum_i[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] &= 0 \\ \sum_i y_i - m\hat{\beta}_0 - \hat{\beta}_1\sum_i x_i &= 0 \\ m\overline{y} - m\hat{\beta}_0 - m\hat{\beta}_1\overline{x} &= 0 \\ \overline{y} - \hat{\beta}_0 - \hat{\beta}_1\overline{x} &= 0 \\ \overline{y} &= \hat{\beta}_0 + \hat{\beta}_1\overline{x} \qquad (13)\end{aligned}$$

For any x_i, let

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{14}$$

Substituting 14 in 10,

$$\begin{aligned}\sum_i [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] x_i &= 0 \\ \sum_i [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] x_i &= 0 \\ \sum_i (y_i - \hat{y}_i) x_i &= 0 \qquad (15)\end{aligned}$$

Solving for β_1

Subtract 13 from 14,

$$\begin{aligned}\hat{y}_i - \overline{y} &= (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \\ &= \hat{\beta}_1 (x_i - \overline{x}) \qquad (16)\end{aligned}$$

Adding and cancelling y_i on LHS,

$$\begin{aligned}(\hat{y}_i - \overline{y}) + (y_i - y_i) &= \hat{\beta}_1 (x_i - \overline{x}) \\ (\hat{y}_i - y_i) + (y_i - \overline{y}) &= \hat{\beta}_1 (x_i - \overline{x})\end{aligned}$$

Multipying both sides by $(x_i - \overline{x})$ and summing up

$$\begin{aligned}(\hat{y}_i - y_i)(x_i - \overline{x}) + (y_i - \overline{y})(x_i - \overline{x}) &= \hat{\beta}_1 (x_i - \overline{x})(x_i - \overline{x}) \\ \sum_i (\hat{y}_i - y_i)(x_i - \overline{x}) + \sum_i (y_i - \overline{y})(x_i - \overline{x}) &= \hat{\beta}_1 \sum_i (x_i - \overline{x})^2 \qquad (17)\end{aligned}$$

Focussing on $\sum_i (\hat{y}_i - y_i)(x_i - \overline{x})$

$$\sum_i (\hat{y}_i - y_i)(x_i - \overline{x}) = \sum_i (\hat{y}_i - y_i) x_i - \overline{x} \sum_i (\hat{y}_i - y_i)$$

Note from 15, $\sum_i (\hat{y}_i - y_i) x_i$ is 0. Thus,

$$\sum_i (\hat{y}_i - y_i)(x_i - \overline{x}) = -\overline{x} \sum_i (\hat{y}_i - y_i)$$

Let us calculate $\sum_i (\hat{y}_i - y_i)$ separately,..

$$\begin{aligned}\sum_i(\hat{y}_i - y_i) &= \sum_i \hat{y}_i - \sum_i y_i \\ &= \sum_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) - m\overline{y} \\ &= \sum_i \hat{\beta}_0 + \sum_i \hat{\beta}_1 x_i - m\overline{y} \\ &= m\hat{\beta}_0 + m\hat{\beta}_1\overline{x} - m\overline{y} \\ &= m(\hat{\beta}_0 + \hat{\beta}_1\overline{x}) - m\overline{y} \\ &= m\overline{y} - m\overline{y} \\ &= 0\end{aligned} \tag{18}$$

Thus,

$$\sum_i(\hat{y}_i - y_i)(x_i - \overline{x}) = 0 \tag{19}$$

Substituting 19 in 17,

$$\begin{aligned}\sum_i(y_i - \overline{y})(x_i - \overline{x}) &= \hat{\beta}_1 \sum_i(x_i - \overline{x})^2 \\ \implies \hat{\beta}_1 &= \frac{\sum_i(y_i - \overline{y})(x_i - \overline{x})}{\sum_i(x_i - \overline{x})^2}\end{aligned}$$

From 13,

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

Regression Parameters using MLE

For the true line of regression $E(Y|x) = \hat{\beta}_0 + \hat{\beta}_1 x$,

$$\hat{\beta}_1 = \frac{\sum_i(y_i - \overline{y})(x_i - \overline{x})}{\sum_i(x_i - \overline{x})^2} \tag{20}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{21}$$

It is strongly advised to check out our interactive example [2] where we have shown visually and also proven how close the results are, between direct formula we just derived and also if directly picking up point of maximum value from the log likelihood graph itself.

[2] http://nbviewer.jupyter.org/gist/parthi2929/e092970b94ee6aeb99519457df41921a