# OLS Baics

Prof. Mary Kaltenberg

Last updated Spring 2022

**Statistical modeling** is the art of:

1. specifying a realistic set of assumptions for the issue and data under study,
2. using estimators and test statistics that have good properties under the chosen assumptions.

## 1 Estimation

Estimation is the "best guess" for an unknown characteristic (moments) of a population distribution.

We rarely know the true characteristic of a population - we can't get every single person's height in the world, for example. We use our knowledge of probability to help us estimate some characteristic.

For example, the mean $\mu$ is an estimator. It is in fact, a least squares estimator.
There is a difference between our model and estimation technique.

### 1.1 Sample Distribution

We can take a sample of height from our class. This sample has its own mean and its own variance - and we can plot its histogram to see the sample distribution.

We know that the sample variance is :

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1}$$

Unsurprisingly, the sample standard deviation is:

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{2}$$

Note that this is **different** from your **population** standard deviation:

$$\sigma = \frac{1}{N}\sum_{i=1}^{n}(y_i - \bar{\mu})^2 \tag{3}$$

For a population standard deviation, you must know the population mean ($\mu$), which we rarely actually know.
AND this is **different** from a standard error.

$$SE(\bar{y}) = \frac{\sqrt{\bar{y}(1-\bar{y})}}{n} \tag{4}$$

which reduces to:

$$SE(\bar{y}) = \sqrt{\frac{\bar{y} - \bar{y}^2}{n}} \tag{5}$$

and thus:

$$SE(\bar{y}) = \frac{s_y}{\sqrt{n}} \tag{6}$$

That's just the standard deviation divided by the square root of the sample size

The standard error is the mean of means. Our sample is one sample of possible millions of samples. If we take the mean we collected in class and compared it to the distribution of all means of samples - the standard deviation of that distribution is the standard error.

We want to know, is the mean that we collected reasonable compared to the distribution of means?

But, wait, won't we have to know something about the function of the distribution of means? Yes!

But, thanks to the power or the central limit theorem, we know that the distribution of means is Normally distributed! This makes inference easier.

For a random variable

$$Z_n = \frac{\bar{y} - E(\bar{y})}{\sqrt{n}\sigma} \tag{7}$$

converges in distribution to the standard normal

$$\lim_{n\to\infty} P(y_n \leq x) = \Phi(x), \forall x \in R \tag{8}$$

Recall that the Normal distribution:

$$\Phi = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{z-\mu}{\sigma})^2} \tag{9}$$

Another important theorem in probability is the Law of Large Number (LLN). This says something about what happens when we repeat experiments - the more times we repeat the experiment, the closer we will get to the expected value. We saw this in our experiment, when everyone rolled the dice over and over again, we got closer and closer to the expected value.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{10}$$

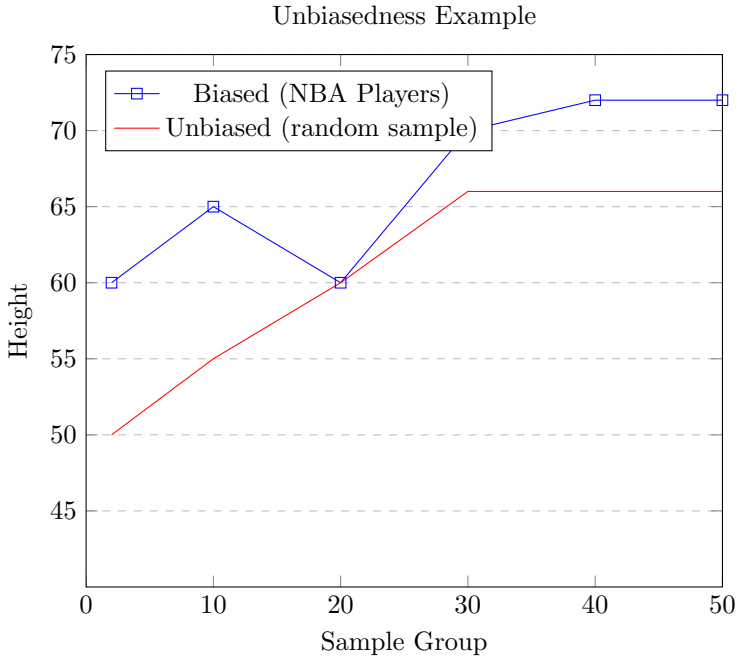
$$LLN = \bar{y} \xrightarrow{p} \mu \tag{11}$$

## 1.2 Properties

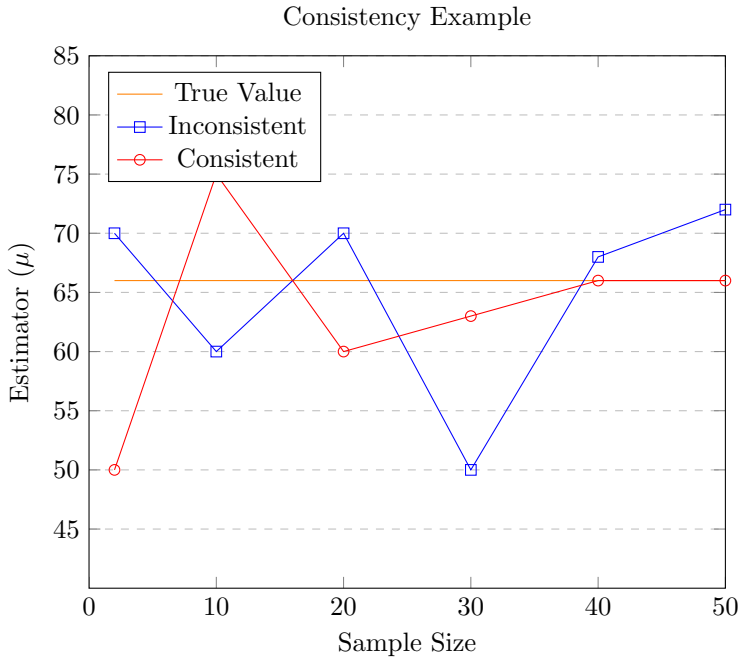
We said something about using the right estimators with properties we like. So, what are the properties that we're really interested in?

1. Unbiasedness
   As we saw from our experiment rolling dice, unbiasedness does not disappear if we continue to roll the dice. Even with CLT, biasedness will remain - as you can see in the numerator, if the $E(\bar{y}) \neq \bar{y}$, then the peak of the distribution will not equal to 0 (the mean of the normal distribution).

   A **very important** property that we want in our estimator is that it is unbiased. We want to ensure that the average variance

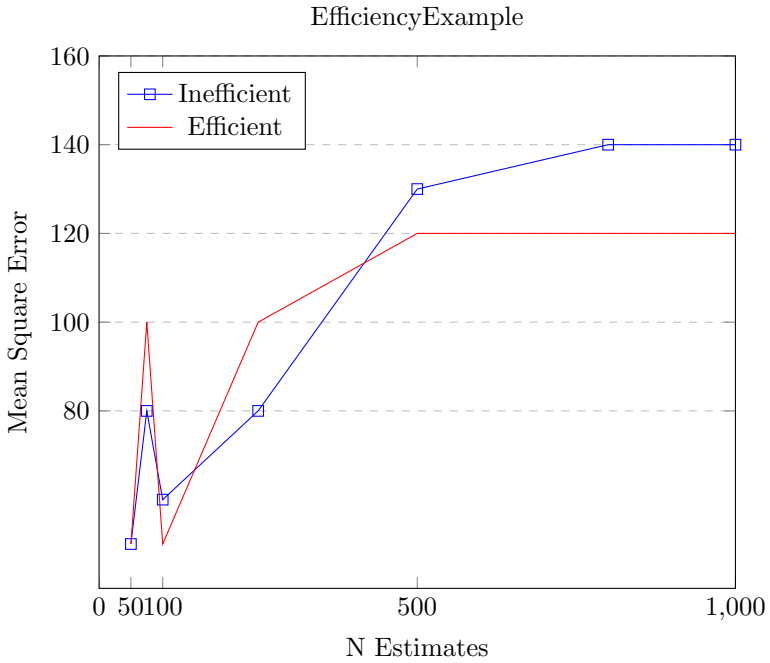



2. Consistency
   As the sample size increases, the estimator converges in probability to the true value of the parameter. Estimators can be consistent **and** biased or inconsistent **and** unbiased. Just because your estimate produces unbiased estimates, that doesn't mean that the estimator is consistent so that as you increase your sample size, your estimator will converge to the true value.




3. Efficiency & Minimum Variance

   Efficiency is a measure of the quality of the estimator that we use. Often this is related to variance because the smallest variance is the most efficient estimator. Imagine you have two unbiased estimators, how do you choose between the two? This decision is based on which estimator is the most efficient. Another way to view this is, how well is my estimator using the information I provided?




Again, we can have an unbiased estimator and poor efficiency, we can have inconsistent estimates and good efficiency - any combination of these properties can happen. However, for most estimators, we want unbiased, consistent, and efficient estimates. When we have these properties and the estimator is linear - we have the **Best Unbiased Linear Estimate (BLUE)**.

**OLS** is one such estimator! That's why it's so frequently used. However, for OLS to be BLUE, it must meet some assumptions. We will get to that a little bit later, but just recall that we are interested in having assumptions that ensure our estimator (OLS) is BLUE.

## 1.3 OLS: The Basics

OLS minimizes the sum of squares residuals. OLS is an estimator for the parameters $(\beta_1, \beta_0)$ that minimizes the sum of square residuals.

We can represent this statistical relationship as:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_i + u_i \tag{12}$$

Or in the conditional expectations function (CEF) :

$$E(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{13}$$

The latter more clearly points out the expectation of Y conditional on some value of X. $E(Y|X)$ represents the mean of a probability distribution shifting with X. Sometimes you hear the CEF also called the *population regression line*.

We are most interested in the slope coefficient of $\beta_1$ - if we are able to carefully craft a causal analysis, $\beta_1$ is the effect or *marginal effect* of X on Y. We can think of this as:

$$\beta_1 = \frac{\Delta E(Y|X)}{\Delta X} \tag{14}$$

- A change in $E(Y|X)$ per unit shift in X
- OR the expected change in Y per unit shift in X (just think about what is a slope in algebra)

The intercept is $\beta_0$, which is $E[y|x = 0]$ (what the expected value of y is given that x is equal to 0). In other words, if you extrapolate the regression relationship until it crosses the vertical axis, you find the crossing at the point (0, ?0).

In some applications, where X = 0 is within the data range, $\beta_0$ has an interesting interpretation. In many other cases, X = 0 is beyond the data range, and it is impossible to give $\beta_0$ a meaningful interpretation. Still, the intercept is necessary for the mathematical representation of any relationship that may not cross the origin.

$u_i$ is the error term (sometimes you see it as $\epsilon_i$ We need the error term, because a random variable generally differs from its expectation: $Y_i$ fluctuates around $E(Y_i|X_i)$.

1. It is unobservable
2. It is a random variable, and we will need assumptions about its probability distribution.
3. It collects all possible reasons why observations deviate from a straight line, and form instead a scatter of sample points:
   - small, unpredictable variations in the effect of Xi on Yi
   - omitted factors, other than the variable Xi, also having an effect on Yi
   - possible lack of precision or errors of measurement in Yi
   - random shocks affecting Yi, like one-shot events or the idiosyncrasies of an individual.

## 1.4 OLS Estimator

We have some data that was randomly sampled where the two random variables, ray blasts and defeated rebels are identically and independently distributed. Our research question is: What's the effectiveness of ray blasts on defeating the rebels? This will inform policy makers if investment in ray blasts is a good strategy to defeat the rebels.

Our data set:

Table 1: Data Set

| X | Y |
|---|---|
| N. Ray Blasts | N. defeated rebels |
| 4 | 1 |
| 7 | 2 |
| 5 | 2 |
| 11 | 5 |
| 9 | 4 |
| 8 | 6 |
| $\bar{x} = 7.3$ | $\bar{y} = 3.3$ |

We will use **Ordinary Least Squares** to estimate the effectiveness of ray blasts on defeating the rebels. How do we apply OLS to data?

OLS is an estimator that minimizes the sum of squares error term. We can solve for $beta_1$ and $beta_0$ with the data that we have collected.

The estimate regression line is a function of $x_i$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{15}$$

where the slope is $\beta_1$, the intercept is $\beta_0$, which is $E[y|x = 0]$ (what the expected value of y is given that x is equal to 0), and $\hat{y}$ is the predicted y or fitted y. y is the dependent variable or regressand, and the x is the independent variable or regressor.

The above equation also uses other notation, but is equivalent:

$$\hat{y} = \alpha + \hat{\beta}_1 x_i \tag{16}$$

The residual is:

$$\hat{u}_i = Y_i - \hat{Y}_i \tag{17}$$

and if you plug in the definition of $\hat{Y}_i$ into the above equation, you get:

$$\hat{u}_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{18}$$

The residuals are our best estimates of the regression errors. However, do not confuse the two!
**The residuals estimate, but do not equal, the regression errors!**
The residual is *estimated* and must sum up to 0.

So, we have the full estimated regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i \tag{19}$$

OLS estimation minimizes the average squared difference between the observed (actual) values of Y and its predicted (fitted) values $\hat{Y}$. It solves the minimization problem:

$$min \sum_{i=1}^{n} (Y_i - \hat{Y})^2 \tag{20}$$

and if you recall, $u_i = Y_i - \hat{Y}$, an thus:

$$min \sum_{i=1}^{n} \hat{u}^2 \tag{21}$$

We can also solve this by replacing our known definition of $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and thus:

$$min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{22}$$

There are two parameters that OLS estimates: $\hat{\beta}_1$ and $\hat{\beta}_0$ and with some calculus we can solve them (more on that later):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \tag{23}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{24}$$

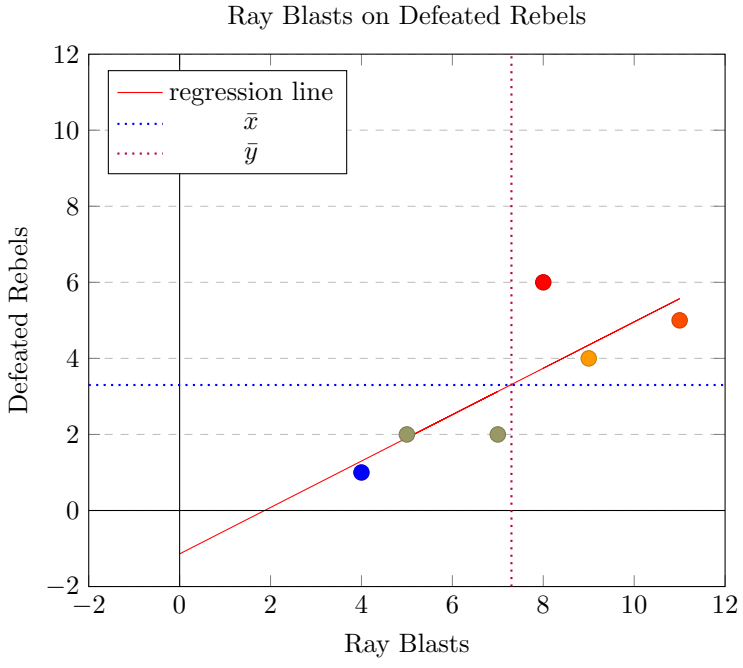
The population regression line *must* go through $\bar{x}$ and $\bar{y}$ - this is an algebraic definition as show above.

Recall that:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{n} Y_i, \bar{X} = \frac{1}{N} \sum_{i=1}^{n} X_i \tag{25}$$

Ok, now we see that every part of these estimates can be solved with our dataset. Utilizing our dataset, we can find the $\beta_0$ and $\beta_1$ estimators:

$$\hat{y} = -1.14 + 0.61x \tag{26}$$



Now, recall the $\beta$ has it's own sampling distribution. If we were to get another random sample and then estimate the same regression model, we would produce another estimate. We can imagine doing this many times so that we have a distribution of $\beta$ parameters. Thus, every $\beta$ has its own sampling distribution AND its own variance. We can estimate the variance of $\beta$ below.

We can calculate the variance with homoskedastic error terms:

$$\sigma^2_{\beta_1} = \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{27}$$

This assumes that :

$$var(u_i|x_i) = 0 \tag{28}$$

However, often we may come across heteroskedasticity (a violation of the above assumption), and one way to deal with this issue is by calculating the variance of our coefficient slightly different. We can incorporate the changing variance of X within our estimation of the variance of $\beta$.

$$\sigma^2_{\beta_1} = \frac{1}{n}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\hat{u}_i^2}{[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2]^2}\right] \tag{29}$$

### 1.4.1 Measures of fit

How do we know how well our regression line "fits" or explains or predicts the data. How well does the regression fit the variation in the data? People usually report two different measures of fit following an OLS regression:

1. $R^2$, the coefficient of determination, often simply called R-squared;
2. The Standard Error of Regression (SER), variously known as Root MSE (Stata), or Residual standard error (R).

Cautionary tale: Often prediction is *not* the goal, it depends on the question. Often, in econometrics we really care about the estimated effect - the causal effect.

**R-Squared**

$R^2$ is a the ratio of what is explained over the total variation.

$$R^2 = \frac{ESS}{TSS} \tag{30}$$

$$R^2 = 1 - \frac{SSR}{TSS} \tag{31}$$

$$TSS = ESS + SSR \tag{32}$$

Sum of squared residuals:

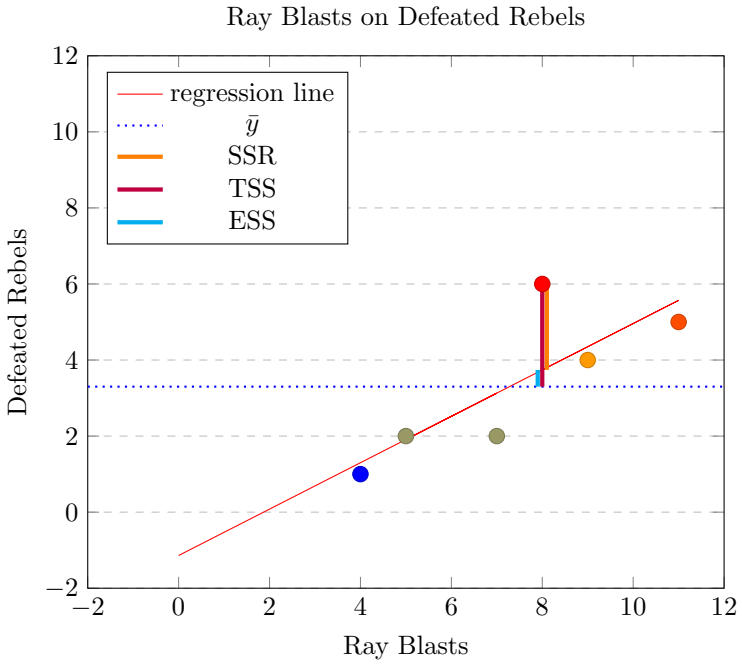
$$SSR = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{33}$$

SSR should look *very* familiar.
Total sum of squares

$$TSS = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{34}$$

Explained sum of squares

$$ESS = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 \tag{35}$$




Pros and Cons $R^2$:

- $R^2$ stays between 0 and 1, and is interpreted as the proportion of the variation of $Y_i$ that can be explained by (and possibly attributed to) variation of $X_i$.
- $R^2$ happens to be also the squared correlation between Y and $\hat{Y}$ , hence its name.
- $R^2$ is popular. Naive users think that $R^2$ is a total quality measure, in the sense that ?high $R^2$ is good?, and that a decent regression should have $R^2$ ? 0.9. Here is a piece of good advice: don?t let yourself be fooled by $R^2$ idolization.
- Values of $R^2$ are strictly comparable across regressions only insofar as the underlying ?Total Sum of Squares? is identical. That means the entire sample of Y -values (Y1,...,YN) must be the same.
- Furthermore, values of $R^2$ are strictly comparable across regressions only insofar as the number of regressors is the same. We will return to this issue in the chapter on multiple regression.
- A low $R^2$ is not necessarily an indication that the model is bad or wrong, just that X (on its own) has low explanatory power.
- A high $R^2$ is not necessarily an indication that the model is good or right, much less that the relationship between Y and X is causal. As you will realize, fit and correlation are not the same as causation!
- Finally, after a good econometrics course, one knows that $R^2$ is easily manipulated.

**Standard Error of Regression**

The Standard Error of Regression (SER) measures the typical size of the residual $\hat{u}$ by means of its quadratic average or "mean square".

I often say, our "average mistake" in our prediction, but that's not technically correct. You cannot measure this typical size with a simple average, because negative and positive residuals will cancel each other out, however big their absolute values. But, intuitively, it says something about how typical the size of our error is with our regression.

$$SER = \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}^2 \tag{36}$$

$$= \frac{SSR}{n-2} \tag{37}$$

I prefer using this metric as a measure of fit, especially when comparing between similar regressions.

- Since SER represents the typical magnitude of $\hat{u}$ in the decomposition $Y_i = \hat{Y}_i + \hat{u}_i$, SER has the same units as the dependent variable Y. This means you can interpret it in terms of those same units.
- SER tells you the typical size, in units of Y, of the ?mistake? you make if you use the fitted values $\hat{Y}$ as predictions of the actual values Y.

- In other words, SER measures the dispersion of the scatter of observed points around the estimated regression line.

Note: Stata outputs the RMSE, which is the square root of SER (also sometimes terms as the mean square error (MSE)). So, the $RMSE = \sqrt{SER}$

## 1.5 Derivation of OLS Estimators

There are 3 steps:

1. Solve partial derivatives w.r.t $\beta_1$ and $\beta_0$
2. Set these equations to 0
3. Solve for $\beta_0$ and plug the definition into $\beta_1$ equation to solve

Objective:

$$min \sum_{i=1}^{n} (Y_i - \hat{Y})^2 \tag{38}$$

Replace our known definition of $\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_i$, and thus:

$$min \sum_{i=1}^{n} (Y_i - (\beta_0 - \beta_1 X_i))^2 \tag{39}$$

### 1.5.1 Partial Derivatives

**w.r.t** $\beta_0$

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{N} (Y_i - (\beta_0 - \beta_1 X_i))^2 \tag{40}$$

Power rule:

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{N} 2(Y_i - (\beta_0 - \beta_1 X_i)) \tag{41}$$

Chain rule:

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{N} 2(Y_i - (\beta_0 - \beta_1 X_i))(-1) \tag{42}$$

Clean it up:

$$\frac{\partial}{\partial \beta_0} = -2 \sum_{i=1}^{N} (Y_i - (\beta_0 - \beta_1 X_i)) \tag{43}$$

**w.r.t** $\beta_1$

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{N} (Y_i - (\beta_0 - \beta_1 X_i))^2 \tag{44}$$

Power rule:

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{N} 2(Y_i - (\beta_0 - \beta_1 X_i)) \tag{45}$$

Chain rule:

$$\frac{\partial}{\partial \beta_0} = \sum_{i=1}^{N} 2(Y_i - (\beta_0 - \beta_1 X_i))(-X_i) \tag{46}$$

Clean it up:

$$\frac{\partial}{\partial \beta_0} = -2 \sum_{i=1}^{N} X_i (Y_i - (\beta_0 - \beta_1 X_i)) \tag{47}$$

### 1.5.2 Set Equal to 0

$\beta_0$

$$0 = -2 \sum_{i=1}^{N} (Y_i - (\beta_0 - \beta_1 X_i)) \tag{48}$$

$\beta_1$

$$0 = -2 \sum_{i=1}^{N} X_i (Y_i - (\beta_0 - \beta_1 X_i)) \tag{49}$$

### 1.5.3 Solve

2 unknowns, solve for $\beta_0$ then plug into equation of $\beta_1$

Solve for $\beta_0$ :

$$0 = -2\sum_{i=1}^{N}(Y_i - (\beta_0 - \beta_1 X_i)) \tag{50}$$

Divide by 2:

$$0 = \sum_{i=1}^{N}(Y_i - (\beta_0 - \beta_1 X_i)) \tag{51}$$

Pull out sums:

$$0 = \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N}\beta_0 - \sum_{i=1}^{N}\beta_1 X_i \tag{52}$$

$\sum_{i=1}^{N}\beta_0$ is a constant, therefore

$$0 = \sum_{i=1}^{N} Y_i - n\beta_0 - \sum_{i=1}^{N}\beta_1 X_i \tag{53}$$

rearrange:

$$n\beta_0 = \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N}\beta_1 X_i \tag{54}$$

Divide by n and recall the definition of $\bar{y}$ and $\bar{x}$,

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{N} Y_i - \frac{1}{n}\sum_{i=1}^{N}\beta_1 X_i \tag{55}$$

And finally:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{56}$$

Solve for $\beta_1$:

$$0 = -2\sum_{i=1}^{N} X_i(Y_i - (\beta_0 - \beta_1 X_i)) \tag{57}$$

Replace $\beta_0$ definition, divide by 2 and put like terms together:

$$0 = \sum_{i=1}^{N} X_i(Y_i - \bar{Y} - \beta_1(X_i - \bar{X})) \tag{58}$$

Carry summation through:

$$0 = \sum_{i=1}^{N} X_i(Y_i - \bar{Y}) - \sum_{i=1}^{N}\beta_1 X_i(X_i - \bar{X})) \tag{59}$$

Pull out the constant ($\beta_1$)

$$0 = \sum_{i=1}^{N} X_i(Y_i - \bar{Y}) - \beta_1\sum_{i=1}^{N} X_i(X_i - \bar{X})) \tag{60}$$

$$\beta_1\sum_{i=1}^{N} X_i(X_i - \bar{X})) = \sum_{i=1}^{N} X_i(Y_i - \bar{Y}) \tag{61}$$

$$\beta_1 = \frac{\sum_{i=1}^{N} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{N} X_i(X_i - \bar{X}))} \tag{62}$$

Note:

$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N} X_i(Y_i - \bar{Y}) - \sum_{i=1}^{N}\bar{X}(Y_i - \bar{Y}) \tag{63}$$

Pull out constants:

$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N} X_i(Y_i - \bar{Y}) - \bar{X}\sum_{i=1}^{N}(Y_i - \bar{Y}) \tag{64}$$

Know that $\sum_{i=1}^{N}(Y_i - \bar{Y}) = \sum_{i=1}^{N} Y_i - n\bar{Y} = 0$, thus:

$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N} X_i(Y_i - \bar{Y}) \tag{65}$$

Also, note:

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X}) \tag{66}$$

Pull out summation:

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{i=1}^{N}(X_i - \bar{X}) - \sum_{i=1}^{N}\bar{X}(X_i - \bar{X}) \tag{67}$$

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{i=1}^{N}(X_i - \bar{X}) - \bar{X}\sum_{i=1}^{N}(X_i - \bar{X}) \tag{68}$$

Know that $\sum_{i=1}^{N}(X_i - \bar{X}) = \sum_{i=1}^{N} X_i - n\bar{X} = 0$, thus:

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{i=1}^{N} X_i(X_i - \bar{X}) \tag{69}$$

Thus,

$$\beta_1 = \frac{\sum_{i=1}^{N} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{N} X_i(X_i - \bar{X}))} \tag{70}$$

is equivalent to:

$$\beta_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{71}$$