## Sample and Population Linear Regression

We can define regression problems for both population and sample data.

Population means we have access to the whole dataset. For example in our predicting hourly wage example, the population will be all the people living in the US. Using this data we can easily get the expected value of the population or the average of the population.

We denote the *expected value of a population by EY.*

Now our goal is to make a model which could estimate the EY given some set of regressor X.

Where X= (x1, x2, ….xp )

We can also denote the X as $(X_{j})^{p}_{j=1}$, this represents the same meaning as X= (x1, x2, ….xp ) where j goes from 1 to p.

We will now make the best linear predictor of Y using X, this means that when we estimate Y using X, while will be the best approximation of given data.

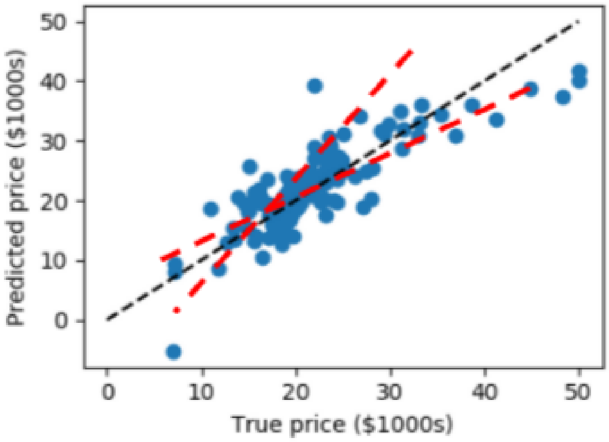


**Figure 1**

Consider Figure1, blue points are the data points, and the red and black lines are different lines that can fit the data. But our goal is to find the black line, which best fits the data, i.e. it gives the least error.

Mathematically, we define our predicted value of Y using a set of regressor (X) as:

β'*X: =  Σ (β'j*xj) where j goes from 1 to p.

We call these parameters βas regression parameters.

For example:

*For our prediction wage example: we can represent the* ***predicted hourly wage*** *as β’X.*

Where X is our job-relevant characteristics like gender, education level, skill set, experience, etc. and β are regression parameters.

If you expand β’X it will look like:

β’X. := β1* gender + β2*( education level) +β3* (skill set) …….

Where β1, β2…. represents the best regression parameters of the linear predictor using which we are going to get the least error.

In Figure 1, the black line is the best linear predictor as it best fits the data and gives the least error. The coefficients of this black line are the beta, i.e. regression parameters.

For the wage example, we can write the equation of the black line as: β’X. := β1* gender + β2*( education level) +β3* (skill set)…

Now let’s define **error** for this problem,

In supervised learning, we know the actual value of the target variable i.e. Y from the data. Now, Using the best linear predictor we estimated the predicted values can be represented as β’X.

Therefore the error we got is defined as:

**Error= E(Y – β’X)**$^{2}$

As we saw at the start, EY is the expected value or the average of the population of Y,

Therefore

$$\mathbf{E(Y - \beta'X)^2 = \frac{1}{N}\Sigma(\ (Y - \beta'X)^2)}$$

Where N is the number of data points.

Consider the image given below:

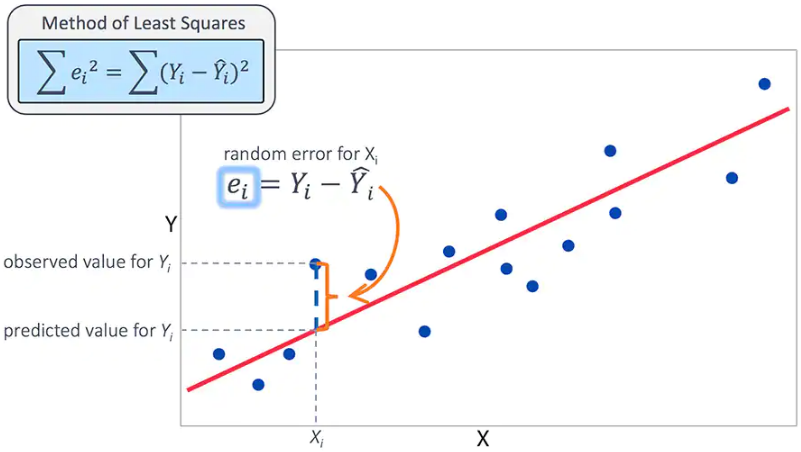


**Figure 2**

In the above image: Yi is the actual value, $\widehat{Yi}$ is the Predicted/estimated value, or EY

The blue dotted line Yi- $\widehat{Yi}$ is the error we got while predicting. So, we take the sum of the square of this error for each data point and minimize that error. As shown in the blue box Figure 2.

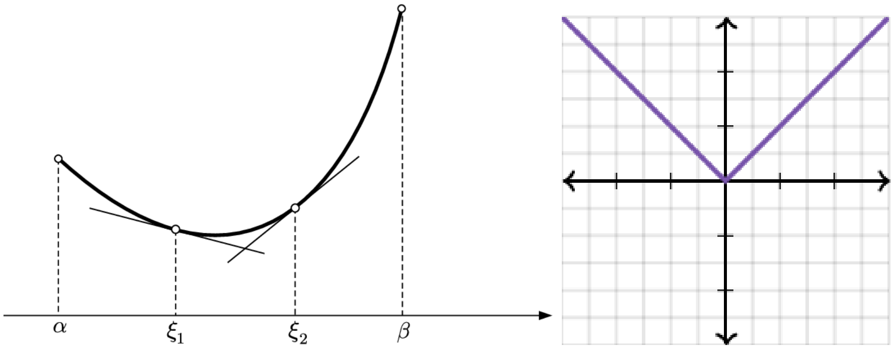
(Y – β'X) is similar as Yi- $\widehat{Yi}$ and we call them **residuals** also.

We can also write this total error as E(Y – β'X)$^2$

Our goal is to **minimize this error, i.e. min E(Y – β'X)$^2$**

**Now the question is why do we take the square of the errors and why not absolute errors?**

This is because, according to calculus, if we want to minimize something, we can set its derivative to zero and we cannot differentiate absolute error, as it has a sharp peak.



*The graph on the left is a convex function and it is differentiable at the lowest/minimum point,  while on the right, not differentiable as it is a peak on the minimum point,*

Now, let’s get back to the discussion on minimizing the error.

To minimize this we just have to take the first derivative of the error and set it to zero.

*First derivative* of min E(Y – β’X)$^2$ is **E(Y-β’X)*X**

So we set **E(Y-β’X)*X = 0**

**The solution for the above equation E(Y-β’X)*X = 0 will give the β1, β2… and minimum error.**

We can also define error **(Y-β’X)** as ϵ.

Replacing **(Y-β’X),** in the equation **E(Y-β’X)*X = 0**, we can write the *first derivative* as:

**E(ϵ)*X = 0**

**Where epsilon is also called the residuals.**

Here we are minimizing the error. To minimize this we have to take the first derivative of the error and set it to zero.

**The first derivative of $E(Y - \beta' X)$ is $E(Y-\beta'X)*X$.**

**As we are finding the minimum , we set $E(Y-\beta'X)*X = 0$ using this equation we find the coefficients($\beta$) that minimizes the error. We can also define error $(Y-\beta'X)$ as $\varepsilon$.**

**Replacing $(Y-\beta'X)$, in the equation gives $E(Y-\beta'X)*X = 0$, we can write the first derivative as,**

**$E(\varepsilon)*X = 0$**

It is very intuitive, we have actual values Y and we estimate some expected value EY as $\beta'X$. Therefore the residual will be $(Y-\beta'X) = \epsilon$**.**

Solving the first derivative will immediately give us the equation:

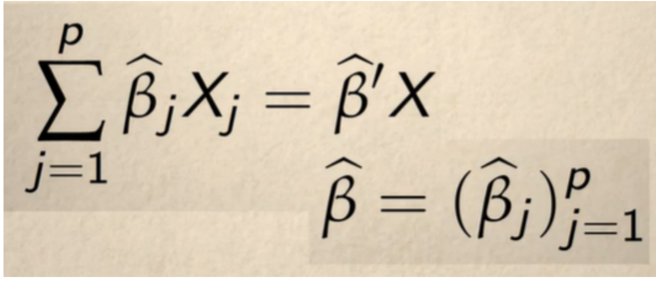
$Y = \beta'X$ + epsilon, epsilon is not correlated with Xs, as it is residual and is not explained by our regression model.

Thus, $\beta'$**X 'is the explained part,** while **epsilon($\epsilon$) residuals are also called the unexplained part** by our regression equation.

So, this is how we get the regression equation for population $Y=\beta'X$, But in actual, we don't have access to population data and we generally work on the sample of data. Samples are drawn randomly from the population, and are denoted as ( (X1,Y1), (X2,Y2), ..... ,(Xn,Yn )), where n is the number of observations.

Formally this means that the observations are obtained as realizations of independently and identically distributed copies of the random vector (Y, X).

We then construct the best linear prediction rule in-sample(train dataset) for Y using X; namely, given X our predicted value of Y will be:



$$\sum_{j=1}^{p} \widehat{\beta}_j X_j = \widehat{\beta}' X$$

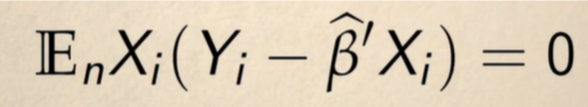
$$\widehat{\beta} = (\widehat{\beta}_j)_{j=1}^{p}$$

This is the same equation as we have for population data, just the representation of changes. For beta is replaced by hat beta, which denotes the sample linear regression parameter/coefficients.

Same as population regression, here we minimize E(Y- $\widehat{\beta}$*X)

Again, we can compute an optimal $\widehat{\beta}$ by solving the Sample Normal Equations:

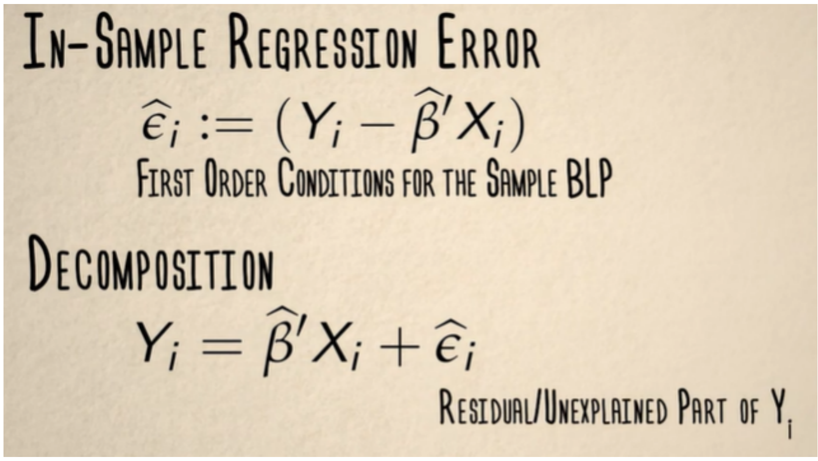
Where we set the empirical expectation as 0.



$$\mathbb{E}_n X_i (Y_i - \widehat{\beta}' X_i) = 0$$

This is again similar to what we did in Population regression.

Defining the in-sample regression error, we obtain the decomposition:



IN-SAMPLE REGRESSION ERROR

$$\widehat{\epsilon}_i := (Y_i - \widehat{\beta}' X_i)$$

FIRST ORDER CONDITIONS FOR THE SAMPLE BLP

DECOMPOSITION

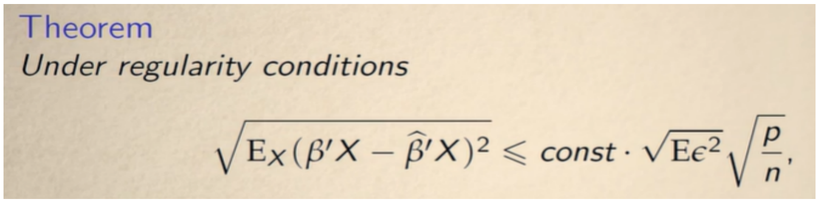
$$Y_i = \widehat{\beta}' X_i + \widehat{\epsilon}_i$$

RESIDUAL/UNEXPLAINED PART OF $Y_i$

Next, we examine the quality of prediction that the sample linear regression provides.

We know that the best linear predictor out-of-sample(test data) is $\beta^*X$. So the question really is: Does the sample best linear predictor $\hat{\beta}X$ adequately approximate the best linear predictor $\beta^*X$? Let's think about it.  Sample linear regression estimates p parameters from $\beta1$ to $\beta p$, without imposing any restrictions on these parameters. So, intuitively, to estimate each of these parameters well, we need many observations per each such parameter. **This means that the ratio of n over p must be large, or, equivalently p divided by n must be small,** where n is the number of observations and p is the number of regressors(X1,X2,X3....Xp).

This intuition is indeed supported by the following theoretical result. which reads: Under regularity conditions, the root of the expected square difference between the best linear predictor and the sample best linear predictor is bounded above by a constant time the level of noise times square root of the dimension p divided by n. Here we are averaging over values of X and the bound holds with probability close to 1 for large enough sample sizes. The bound mainly reflects the estimation error in $\hat{\beta}$, since we are averaging over values of X. In other words, if n is large and p is much smaller than n, for nearly all realizations of the sample, the sample linear regression gets really close to the population linear regression.



Theorem

*Under regularity conditions*

$$\sqrt{\mathrm{E}_X(\beta' X - \widehat{\beta}' X)^2} \leqslant const \cdot \sqrt{\mathrm{E}\epsilon^2}\sqrt{\frac{p}{n}},$$