Constructed Regressor

In the previous modules, we have built a regression model,

$$Y = \beta' X + \epsilon = \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

where X are independent variables or regressors and β's are coefficients associated with them. So we have found $\beta' X$ is the best linear predictor of Y using X.

We have p regressors over here generally we call it p-dimensional. Mathematically we can write as,

$$X = (X)_{J=1}^{P}$$

Now we have p which is much larger than n .i.e number of regressor are larger than the number of observations.

Reasons for high dimensional regressors:

1. The increasing availability of modern rich data or "big" data.
2. Constructed regressors.

Many modern datasets are rich in that they have many recorded features or regressors: for example,

- In house pricing and appraisal analysis, we can make use of numerous housing characteristics such as the building class, Type of road access, General shape of the property, Proximity to the main road or railroad, and many more we can think of.
- In demand analysis, we can rely on price and product characteristics
- In the analysis of court decisions, we can employ many of the judges' characteristics.

Another reason due to which regressors are increasing is because of Constructed regressors.

Constructed Regressors :

Constructed regressors are the regressor that can be constructed from raw regressors using any transformation such as squaring, cubing, etc.

Formally, if we have Z raw regressors or features then the constructed regressors denoted by X are given by the set of transformations $P(z)$ whose components are $P_j(z)$. We sometimes call this set of transformations a dictionary.

$$X = P(Z) = (P_1(Z), ..., P_p(Z))'$$

For instance, in the Wage Example, we used quadratic and cubic transformations of experience, as well as interactions (or products) of these regressors with education and geographic indicators.

Another example is if we have three features x_1, x_2, x_3 then the constructed regressors are,

$$dictionary = \left\{x_1x_2, x_2x_3, x_3x_1, x_1^2, x_2^2, x_3^2, x_1^3, x_2^3, x_3^3\right\}$$

Here we are considering only square, cubic, and interactions between regressors. The set of all these transformations is the dictionary. Earlier there were only three features now after transformation we got 9 new regressors or features. So now it makes a total of 12 features or regressors.

Motivation:

The use of constructed regressors allows us to build more flexible and potentially better prediction rules than just linear rules that employ raw regressors. This is because we are using prediction rules $\beta'X = \beta'P(z)$ that are allowed to be either linear or nonlinear in the raw regressors Z. So it can capture both linear and non-linear patterns. Although this model allows for a nonlinear relationship between *Y* and *X*, we still call the prediction rule $\beta'X$ linear, because it is linear with respect to the parameters β and with respect to the constructed regressors X = P(Z).

What is the best prediction rule for Y using Z?

It turns out that the best prediction rule is the conditional expectation of Y given Z.

$$g(Z) = \mathrm{E}(Y \mid Z)$$

The conditional expectation is nothing but if we are doing Ordinary least squares for linear regression then we are finding the best linear prediction which is $\beta' X$ then we will write as $E(Y|X) = \beta' X = \beta_1 X_1 + \beta_2 X_2 \ldots$. it says we need to find the expected value of Y given X.

We denote this prediction rule by g(Z), and we call it the regression function of Y on Z. This is the best prediction rule among all rules, and it is generally better than the best linear prediction rule.

Indeed, it can be shown that g(Z) solves the best prediction problem, where we minimize the mean squared prediction error (or MSE) among all prediction rules m(Z) (linear or nonlinear in Z).

$$\min_{m(Z)} \mathrm{E}(Y - m(Z))^2$$

That is the mean squared prediction error is equal to the mean squared approximation error plus a constant that does not depend on b so that minimizing the former is the same as minimizing the latter. We now conclude that the BLP $\beta' P(z)$ is the Best Linear Approximation (or BLA) to the regression function g(Z).

$$\mathrm{E}(Y - b'P(Z))^2 = \mathrm{E}(g(Z) - b'P(Z))^2 + \text{const.}$$

By using a richer and richer dictionary of transformations, P(Z), we can make the best linear predictor approximate better. Let's understand with an example,

Suppose Z is uniformly distributed on the unit interval (0,1) and the true regression function is,

g(Z)=exp(4.Z)

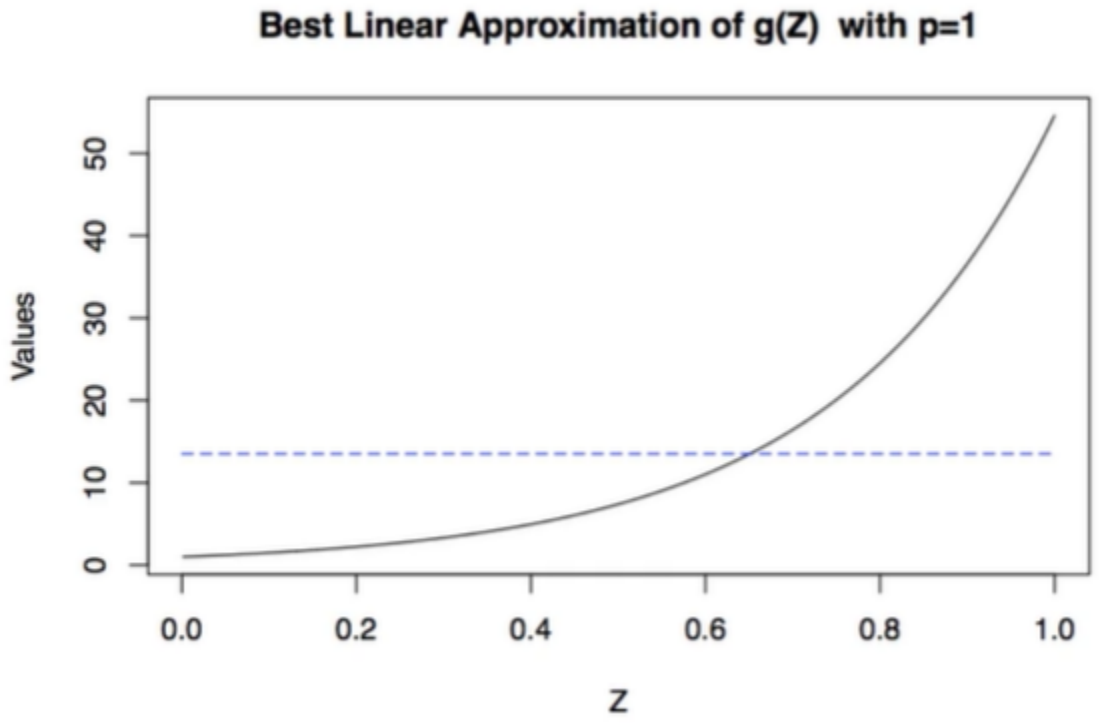
Suppose we don't know this and we use the linear forms $\beta' P(z)$ to provide approximations to g(Z).

Suppose we use P(Z) that consists of polynomial transformations of Z, consisting of the first p terms of the polynomial series:

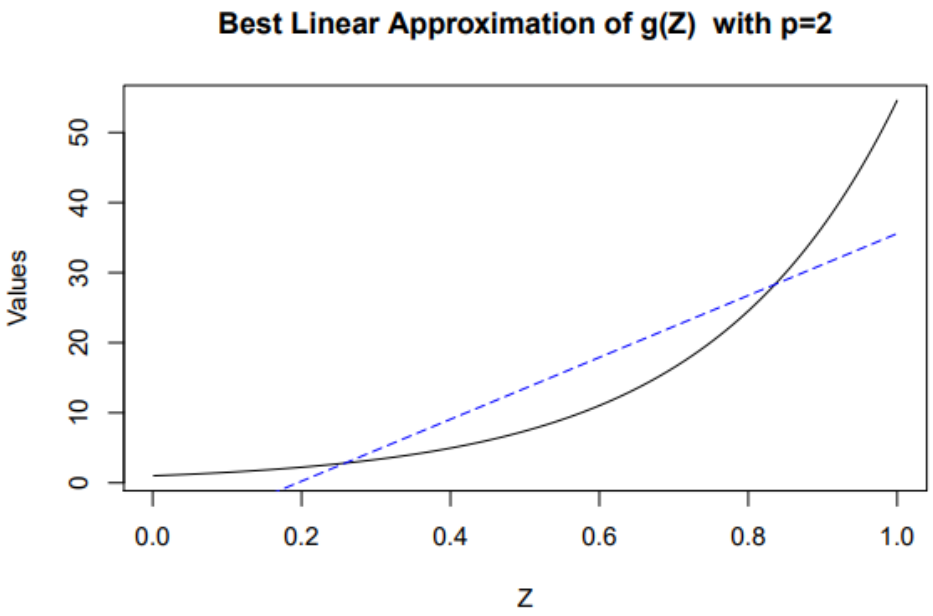
$$P(Z) = \underbrace{(1, Z, Z^2, \ldots, Z^{p-1})'}_{p \text{ terms}}$$

So we use this dictionary to build the best linear approximation.

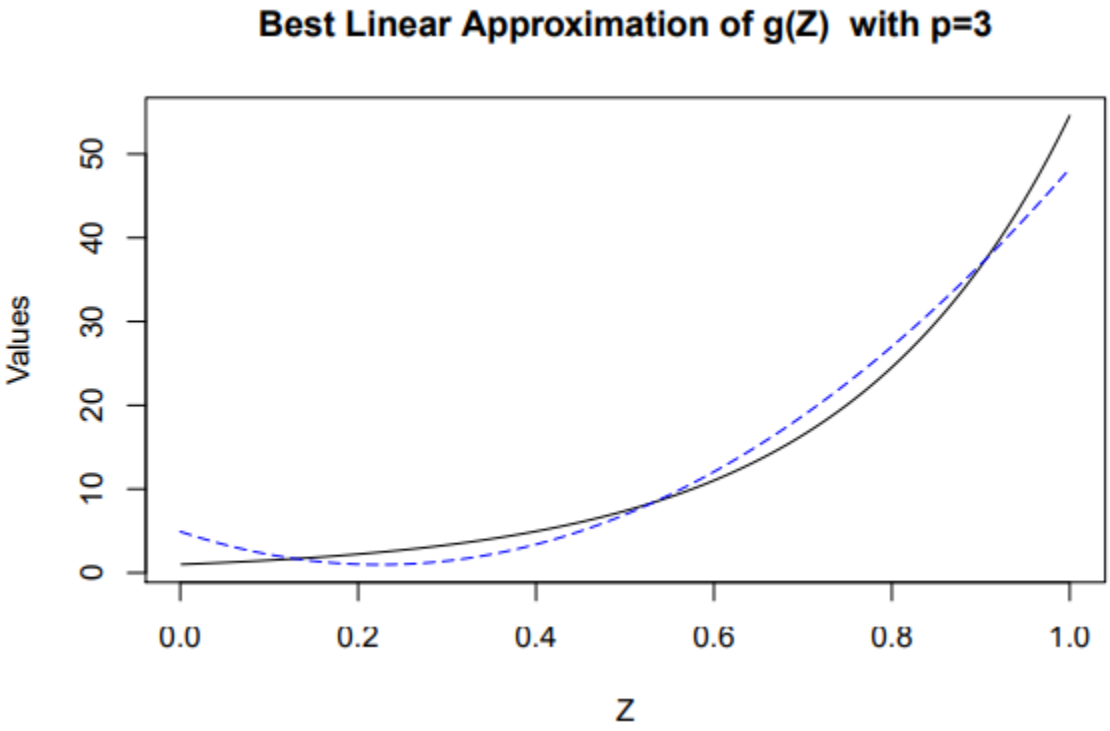
We now build a sequence of figures that illustrate the approximation properties of BLA or BLP corresponding to p equals 1, 2, 3 and 4.



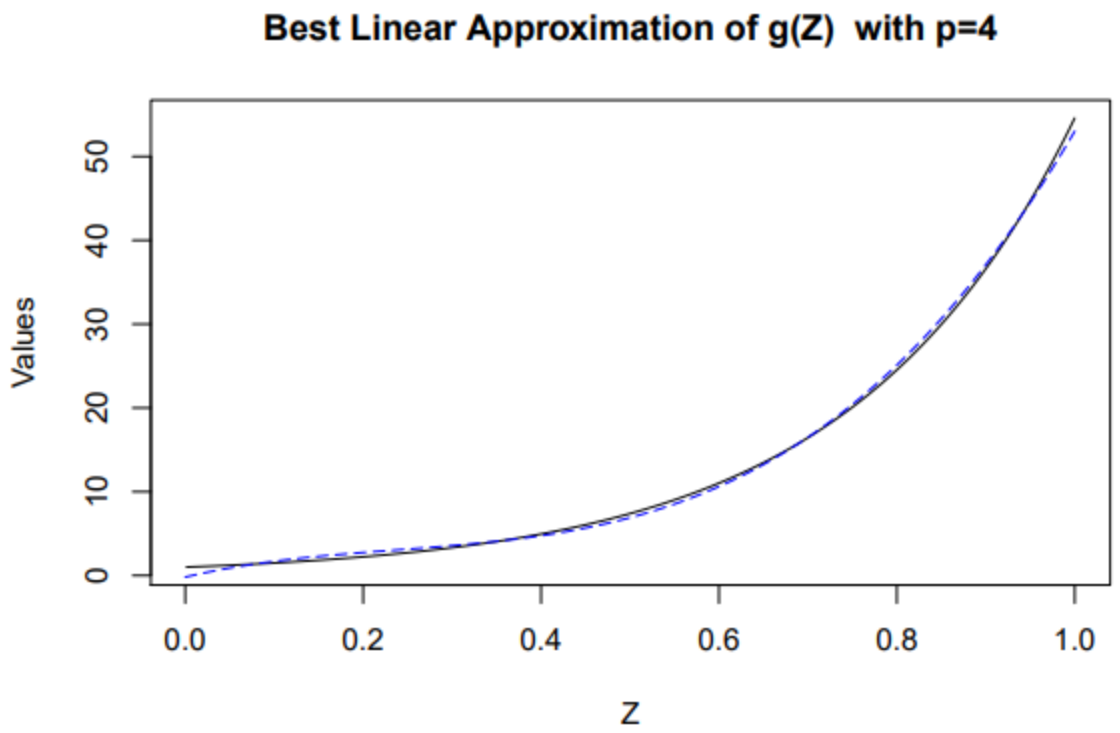
In the above image, the black line represents the equation, **g(Z)=exp(4.Z).** We need to find the best linear predictor for this curve. If we had used p=1 it's nothing but the horizontal line which is a very poor approximation.



Here we have increased the value of p to 2 which is the original regressors. We can still see the quality of this approximation is still very poor.



Here we have increased the value of p to 3 which is the square of all the regressors. Now, the quality is quite good.



Here we have increased the value of p to 4 which is a cubic-in-Z approximation to g(Z), and the quality of the approximation becomes simply excellent. This further stresses the motivation for using nonlinear transformation of raw regressors in linear regression analysis. This explains why using nonlinear transformation of raw regressors is a good idea.

To summarize,

1. The first motivation is that modern data sets have high-dimensional features that can be used as regressors.
2. The second motivation is that we can use non-linear transformations of features or raw regressors and their interactions to form constructed regressors. This allows us to better approximate the ultimate and best prediction rule.