Linear Regression

In this project we draw a synthetic dataset with linear relation out of a population defined by a generating linear relation and a superposed gaussian noise (mean of zero and standard deviation of 5). The equation to construct our dataset is:

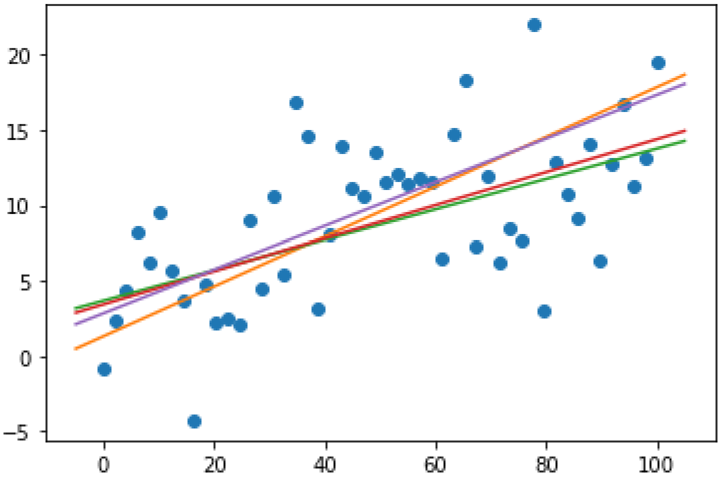
$y = 0.1x + 4 + \xi(\text{noise})$

$\beta_0=4$ and $\beta_1= 0.1$

We want to see if our regression model is able to reconstruct the generating equation or not.

With one time running regression model we get one set of regression coefficients (one result is not reliable) . If we repeat several times the process, each time we get a different set. This is because every time a new training validation set is drawn.

We want to know how confident we are about these results. We want to see the variability of the regression coefficients. For example by running four times the regression model we get four different regression lines.



We repeat the process (1000 times), every time the database divided in a training and test set, selected randomly, with a proportion of 80% training and 20% for the test set. We tack the mean and standard deviation of our process to compare with the generating equation.

For each process we calculate the MSE (mean squared error). This error represent the noise in our dataset.

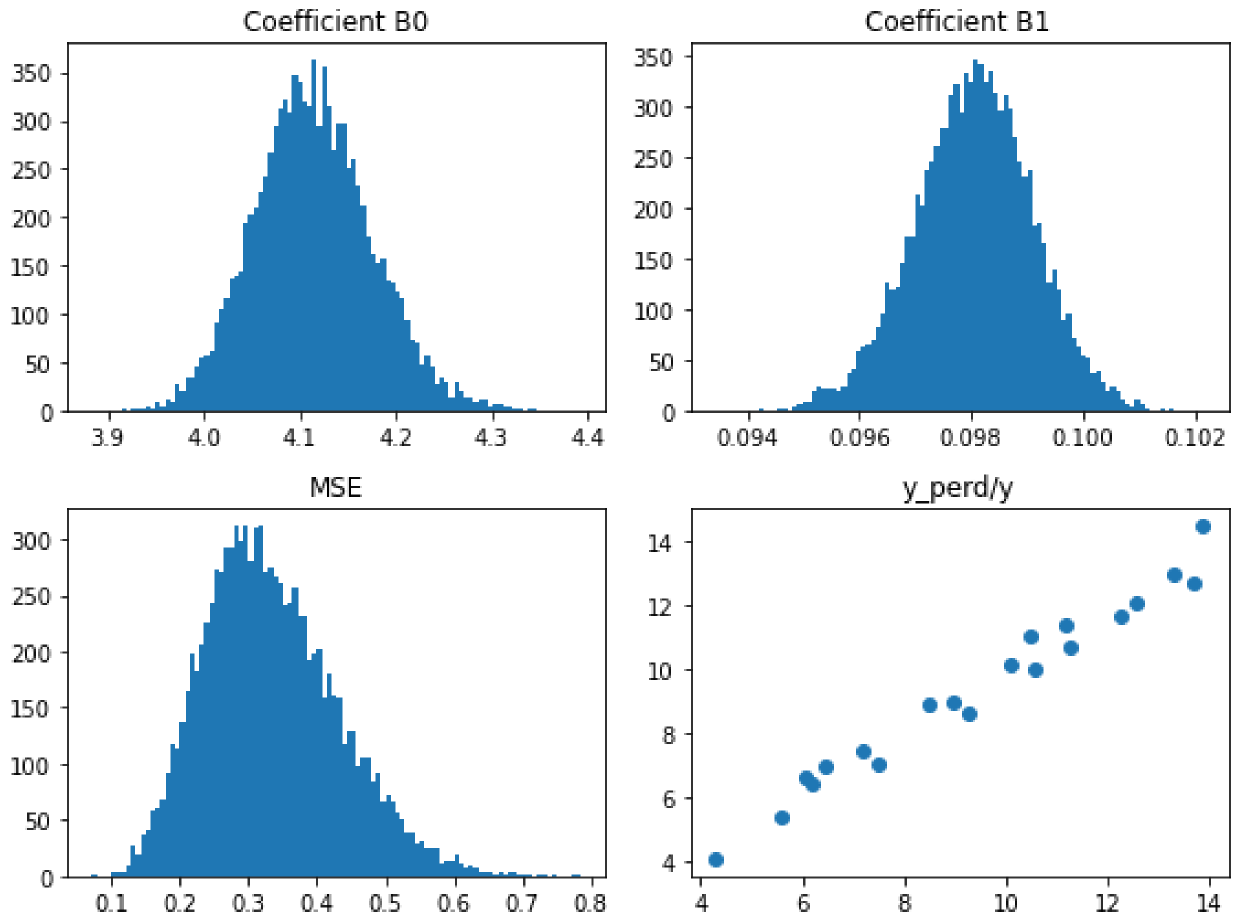
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

N = number of data points

Y_i= observed values

$\hat{Y}_i$ = predicted values

As we see in the output results (graphs), the estimation of each coefficient of regression model follows the gaussian distribution.



The mean of graph is the best estimator of these coefficients. For example a confident interval to accept β0 is:

β0+- one Standard deviation

in our result:

Mean β0 = 4.0088411047971695

Std dev on β0 = 0.04967255263703209

Therefor a confident interval to accept β0 is:
3.96 ≈ β0 ≈ 4.05

We see that our estimated β0 is almost equal to β0 of our generator equation.