Exercise 5 (python code + text):

Consider a regression problem where both the independent and dependent quantities are scalars and are related via the following linear model

$$y = \theta_o \cdot x + \eta$$

where η follows the zero mean normal distribution with variance σ^2 and $\theta_o = 2$ (thus, the actual model is $y = 2 \cdot x + \eta$).

(a) Generate $d = 50$ data set as follows:

- Generate a set D_1 of $N = 30$ data pairs $(y_i{}', x_i)$, where $y' = 2 \cdot x$.
- Add zero mean and $\sigma^2 = 64$ variance Gaussian noise to the $y_i{}'$'s, resulting to y_i's.

- The **observed** data pairs are (y_i, x_i), $i = 1, \ldots, 30$, which constitute the data set D_1.

Repeat the above procedure d=50 times in order to generate 50 different data sets.

(b) Compute the LS linear **estimates** of θ_o based on $D_1, D_2, \ldots, D_d$ (thus, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ numbers/estimates will result).

(c) Consider now the random variable $\hat{\theta}$ that models $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ (that is, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ can be viewed as instances of the random variable $\hat{\theta}$)[1] and

(c1) compute the $MSE = E\left[\left(\hat{\theta} - \theta_o\right)^2\right]$ and

(c2) depict graphically the values $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d$ and comment on how they are spread around θ_o.

<u>*Hint:*</u> For (c) approximate MSE as $MSE = \frac{1}{d}\sum_{i=1}^{d}\left(\hat{\theta}_i - \theta_o\right)^2$.