Validation

Anahita Zarei

Overview

- Case for the Validation set
- Utilization of validation set
- Hold-out validation
- K-fold cross-validation
- Reading: 4.3 from “Learning from Data” by Abu-Mostafa

Validation Set

- Minimize E_{out} rather than just E_{in}.
- Of course E_{out} isn't available to us, so we need an estimate based on information available to us in sample.
- We've seen the idea of a test set before, where a subset of D that is not involved in the learning process is used to evaluate the final hypothesis.
- The idea of a validation set is *almost* identical to that of a test set. The difference becomes clear shortly.

The Validation Set

- Partition the data set D into a validation set (K points) and a training set (N-K points) at random.
- Validation Set: $(X_1,y_1),\ldots,(X_K, y_K)$
- The held out set is effectively out-of-sample, because it hasn't been used during learning.
- The error for a single point in the validation set is e(h(X),y).
  - squared error = $(h(X)-y)^2$,
  - classification error = $\begin{cases} 1 & h(X) \neq y \\ 0 & h(X) = y \end{cases}$
- E_{val} then is

$$E_{val}(h) = \frac{1}{K}\sum_{k=1}^{K} e(h(X_k), y_k)$$

How reliable is E_{val} in estimating E_{out}?

- The validation error is an unbiased estimate of E_{out} , because the final hypothesis was created independently of the data points in the validation set.
- Mathematically, you can show that expected value of E_{val} equals E_{out} .

$$\mathbb{E}\left[E_{\text{val}}(h)\right] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\mathrm{e}(h(\mathbf{x}_k), y_k)\right] = E_{\text{out}}(h)$$

- But unbiased doesn't imply high reliability. How about variance?

$$\text{var}\left[E_{\text{val}}(h)\right] = \frac{1}{K^2}\sum_{k=1}^{K}\text{var}\left[\mathrm{e}(h(\mathbf{x}_k), y_k)\right] = \frac{\sigma^2}{K}$$

- Conclusion: Increasing the size of the validation set results in a better estimate of E_{out} .

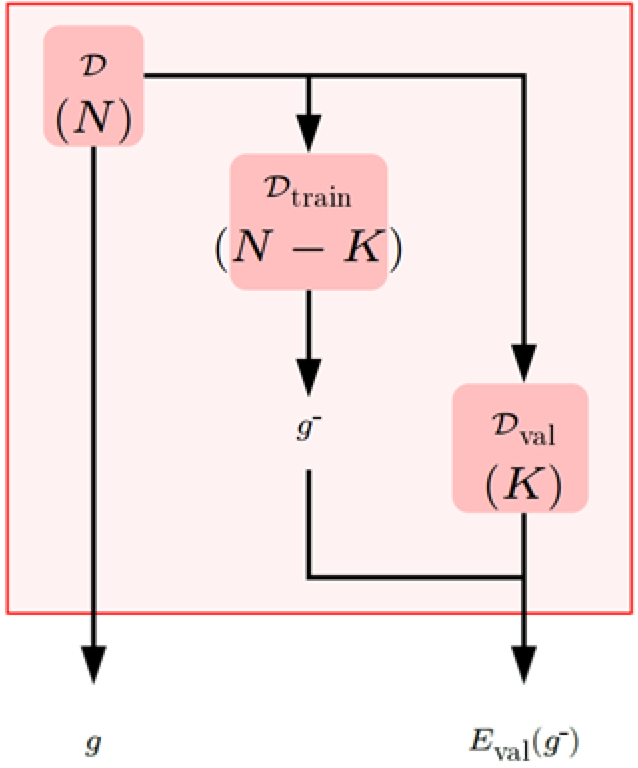
$$E_{val}(h) = E_{out}(h) \pm O\left(\frac{1}{\sqrt{K}}\right)$$

How big should the validation set be?

- The derivation on the last slide tells us that a **LARGE K** will result in a **BETTER** estimate of E_{out}.
- Is there a downside in choosing a large K?
- Yes! There is a price to be paid for increasing K: when we set aside more data for validation, there are fewer training data points.
    - K validation points => N-K training points
- Overfitting is a function of number of points in the training set.
- If number of training data points becomes critically **SMALL**, we're **HURTING** the model performance.

How big should the validation set be?

- We established two conflicting demands on K.
  1. It has to be big enough for E_{val} to be reliable.
  2. It has to be small enough so that the training set is big enough to get a decent hypothesis.
- Though we said that taking out K points for validation and using only N-K for training will cost us in terms of getting a better hypothesis, we do not have to pay that price.
- We first train with N-K data points, validate with the remaining K data points and then retrain using ALL the data points to get a better hypothesis.
- Therefore, you report the E_{val} on a reduced hypothesis, not on the final hypothesis. If K isn't too large, the E_{val} of the reduced and full model are close.
- Rule of thumb: **Use about 20% of the data for validation.**

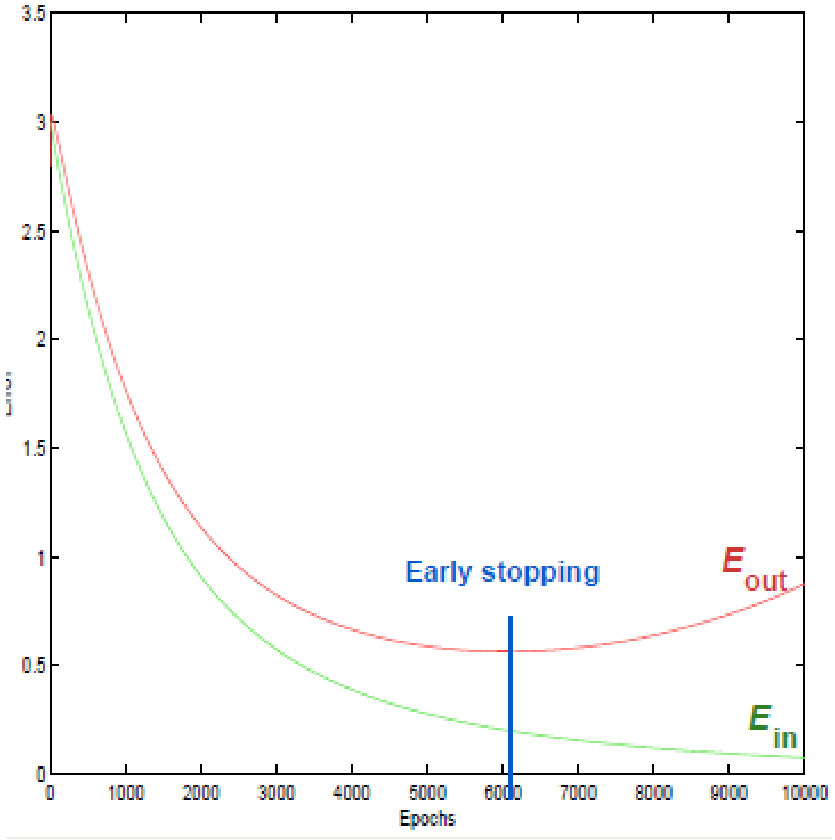


Model Selection

- One of the most important use of validation is for model selection.
- Example: Choice between a linear model vs. a nonlinear model, the choice of the order of polynomial in a model, between two different neural networks with different topologies, etc.
- M models: $H_1,\ldots, H_M$
  - Use D_{train} to learn a hypothesis for each model.
  - Evaluate hypothesis using D_{val}: $E_m=E_{val}(h_m)$ m=1,…M
  - Pick model with smallest E_m.

Difference Between the Test Set and Validation Set

- If we treat the validation set as a way to estimate E_{out}, **without involving** it any decisions that affect the learning process, then there is no difference between the test set and validation set.
- As soon as you start using the set to make decisions about the learning process (e.g. early stopping) or model selection, then it's no longer a test set.



Cross Validation

- Earlier, we expressed the dilemma for selecting K: We want the K to be small so the hypothesis on full data is close to the one on reduced data. We'd also like K to be large so Eval provides a good estimate of E_{out} .

$$\underbrace{E_{\text{out}}(g) \approx E_{\text{out}}(g^-)}_{\text{(small } K\text{)}} \approx \underbrace{E_{\text{val}}(g^-)}_{\text{(large } K\text{)}}$$

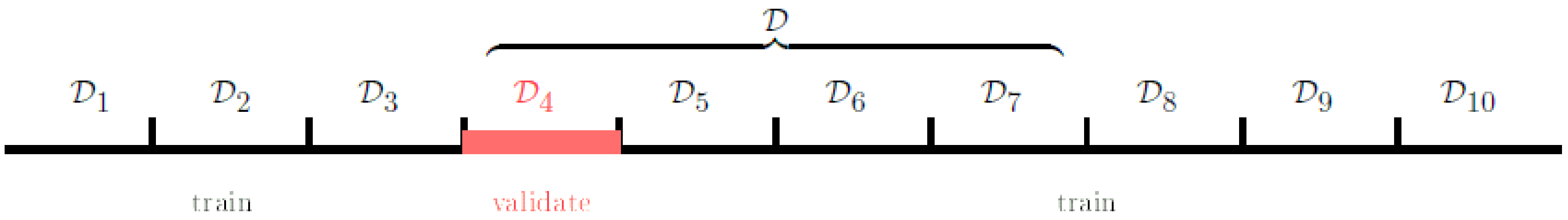
- One solution to this dilemma is cross validation.
- The simplest form of cross validation is leave-one-out.

Leave One Out

- Set N-1 points for training and 1 point for validation. This means that your estimates for E_{out} (g) and E_{out} (g-) are close. (But Eval isn't a reliable estimate for E_{out}.)
- D_n: $(X_1,y_1),\ldots(X_{n-1},y_{n-1})$, $\cancel{(X_n,y_n),}$ $(X_{n+1},y_{n+1}),\ldots(X_N,y_N)$
- Final hypothesis from D_n is h_n.
- $e_n=E_{val}(h_n)$
- Repeat this for all points. Every estimate is out of sample with respect to hypothesis that's used to evaluate.
- Now define Cross Validation Error as $\boxed{E_{CV} = \frac{1}{N}\sum_{n=1}^{N} e_n}$ which is a descent estimate for E_{out}.

K-fold Cross Validation

- Leave-One-Out results in N training session.
- Instead break the data to a number of K folds.
- K training session on the remaining points each time.
- Rule of Thumb:10-fold cross validation => 10 training sessions



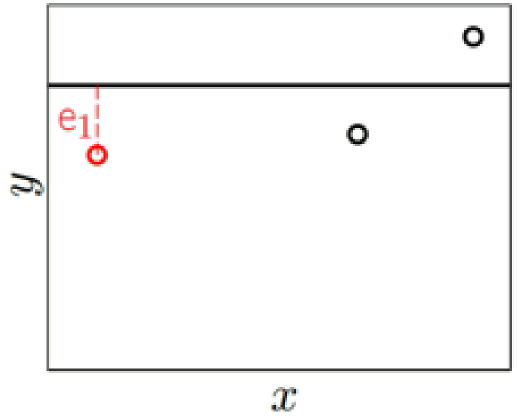
Toy Example

Use leave-one out cross- validation to determine between a constant model or a linear model (y=mx+b) for the following data points.

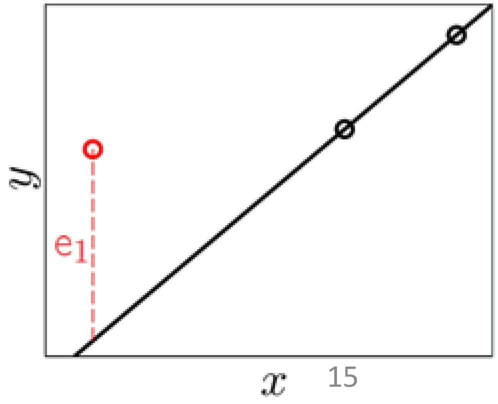
$$(x_1 = 1, y_1 = 3)$$
$$(x_2 = 4, y_2 = 3.5)$$
$$(x_3 = 6, y_3 = 5)$$

Solution

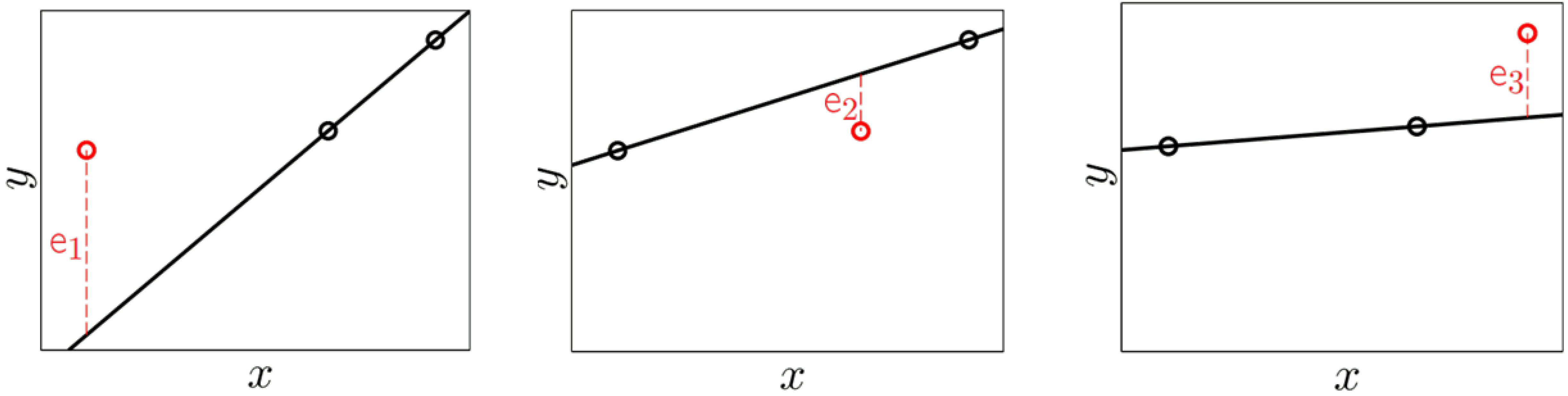
- Leave point 1 out:
- $(x_1 = 1, y_1 = 3)$
- $(x_2 = 4, y_2 = 3.5)$
- $(x_3 = 6, y_3 = 5)$
- $y = \frac{5+3.5}{2} = 4.25$
- $e_1 = (3 - 4.25)^2 = (-1.25)^2$



- $(x_1 = 1, y_1 = 3)$
- $(x_2 = 4, y_2 = 3.5)$
- $(x_3 = 6, y_3 = 5)$
- $m = \frac{5-3.5}{6-4} = \frac{3}{4}$
- $y - 5 = \frac{3}{4}(x - 6) \Rightarrow y = \frac{3}{4}x - \frac{1}{2}$
- $e_1 = \left(3 - \frac{3}{4} \times 1 + \frac{1}{2}\right)^2 = (2.75)^2$

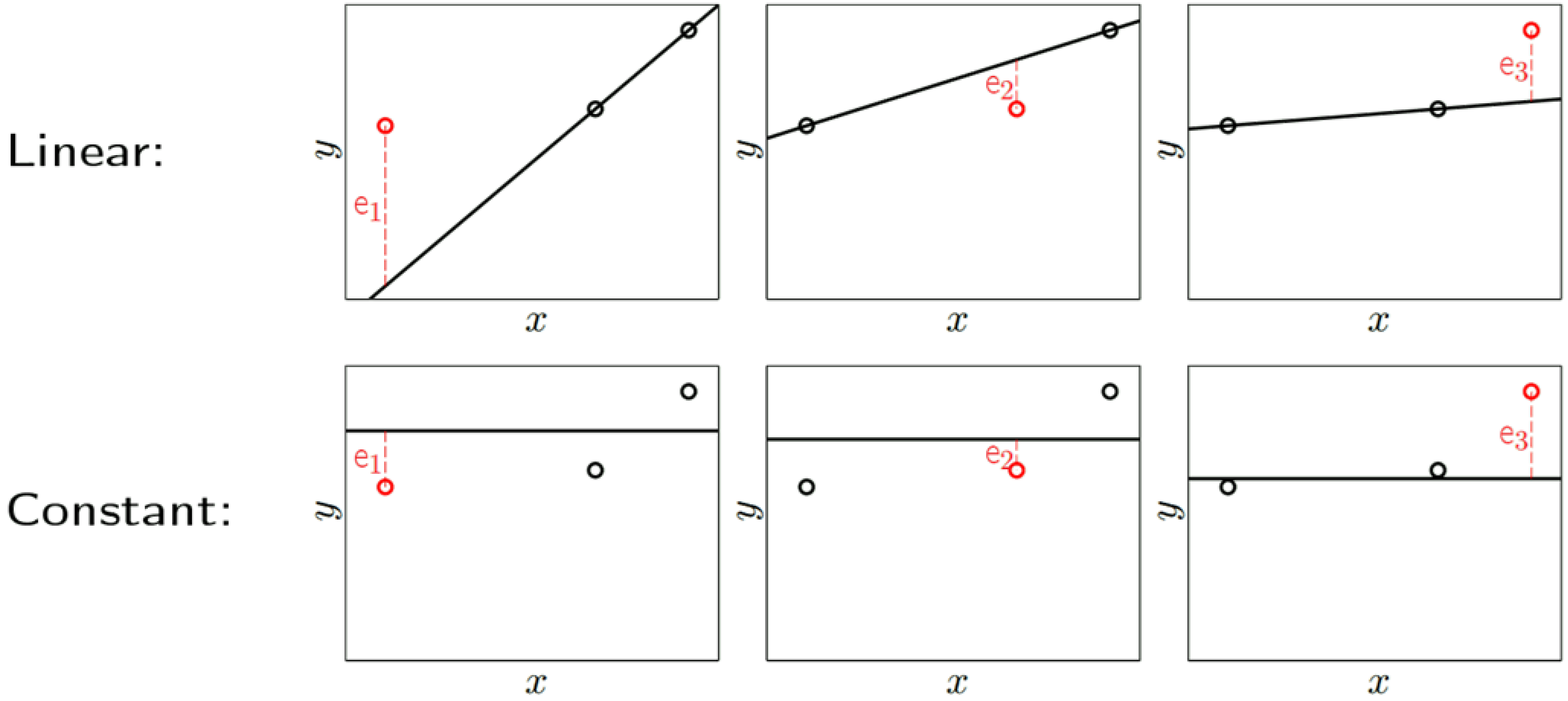


Example of Leave-one Out Cross Validation



$$E_{\text{cv}} = \frac{1}{3}\left(e_1 + e_2 + e_3 \right)$$

Model Selection using Leave-one Out Cross Validation



$x_1 = 1,\ y_1 = 3$
$x_2 = 4,\ y_2 = 3.5$
$x_3 = 6,\ y_3 = 5$

Linear

$y = mx + b$

Constant

① $m = \frac{5 - 3.5}{6 - 4} = \frac{1.5}{2} = 0.75$

② $m = \frac{5 - 3}{4 - 1} = \frac{2}{3}$

$m = \frac{3.5 - 3}{}$

$Y_1 = \frac{3.5 + 5}{2} = 4.25$

$E_1 = 4.25 - 3$

$= 1.25^2 = 1.56$

$Y_2 = 4$

$E_2 = 4 - 3.5 = 0.5^2 = 0.25$

$Y_3 = 3.25$

$E_3 = 3.06$

$E = 1.62$

$y = 0.75x + b$

$5 = 0.75\ 6 + b$

$y = 0.75x + 0.5$

$y = 0.75\ 1 + 0.5$

$y = 1.25$

$E = 3 - 1.25$

$= 1.75$

$Y = \frac{2}{3}x + b$

$3 = \frac{2}{3}\ 1 + b$

$b = 4.5$

$Y = \frac{2}{3}(6) + 4.5$

$y = 8.5$

$E = 8.5 - 5$

$E = 3.5$

MSE = 1.62

MSE = 1.638