1. Use the provided formula (exercise).

2. Use the R function lm( ).

3. Compute the L^2-error for the least square solution.

4. Compute the root mean square error (RMSE) $\frac{\|r\|_2}{\sqrt{n}}$, where r is the residual and n is the number of observations.

Solution.

$$
\begin{aligned}
\text{mean}(x) = 0.5, \qquad \text{mean}(y) &= -\frac{3}{16} = -0.1875 \\
(x_i - \text{mean}(x)) &= (-3-0.5, -2-0.5, -1-0.5, 0-0.5, 1-0.5, 2-0.5, 3-0.5, 4-0.5) \\
&= (-3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5) \\
(y_i - \text{mean}(y)) &= \Big(-2-(-\frac{3}{16}), -1-(-\frac{3}{16}), \dots, -1.5-(-\frac{3}{16}), 1-(-\frac{3}{16})\Big) \\
&= (-1.8125, -0.8125, -1.8125, 0.1875, 1.1875, 3.1875, -1.3125, 1.1875) \\
\sum (x_i - \text{mean}(x)) \cdot (y_i - \text{mean}(y)) & \\
&= (-3.5)(-1.8125) + (-2.5)(-0.8125) + \cdots + (3.5)(1.1875) = 17.25
\end{aligned}
$$

□

Project 6 - part 1 (2 points, due April 29)

Let $\{(x_i, y_i)\} = \{(1,1), (2,2), (0,0), (3,2)\}$ be the data points.

1. Find the best line of linear regression using formulas (1) and (2).

2. Find the best line of linear regression using the R function lm( ).

3. Compute the L^2-error for the least square solution.

4. The root mean square error (RMSE) is $\frac{\|r\|_2}{\sqrt{n}}$, where r is the residual and n is the number of observations. Compute the RMSE for your model.

1.2 Higher dimensional case

A variable Y is said to be linearly dependent on variables $X = [X_1, X_2, \dots, X_m]$ if there is a column constant vector $\beta = [\beta_1, \beta_2, \dots, \beta_m]^T$, so that

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m.$$