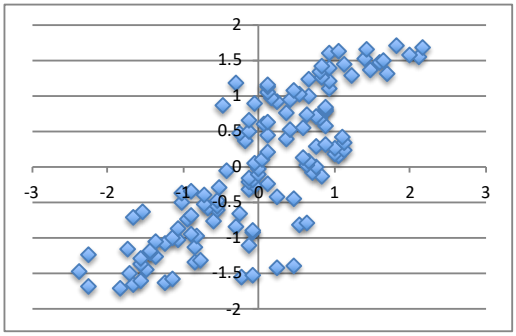
A warmup exercise

The line of best fit for modeling the relationship between 2 variables in a dataset passes through the origin (see example in scatter plot). Find the equation of that line by minimizing the sum of the square of the errors. Assume you are given all the (x, y) values.



Here are the steps:

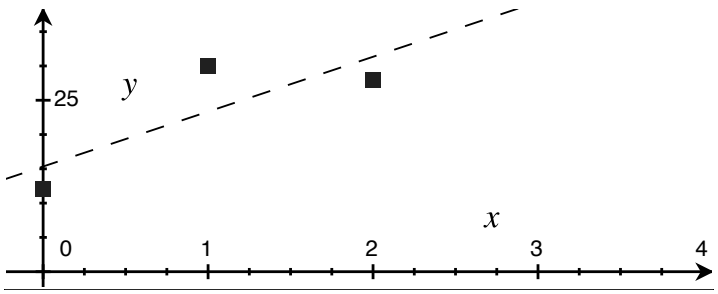
1. Let the equation of the line be: $\hat{y} = mx$.
2. Denote the given set of data points as (x_i, y_i) for $i = 1, 2, 3, \ldots, n$.
3. For each i, find the error: $e_i = y_i - \hat{y}_i$
   Each e_i will be a function of m only.
4. Compute the function: $f = \sum(e_i)^2$
   This function is the sum of the square of the errors.
5. Minimize f using calculus, and find the corresponding value of m.
6. Plug it into $\hat{y} = mx$.

Keep this result in mind! We will come back to it after the next warmup.

Another warmup exercise

In a previous class you found the line of best fit for the following data set

x_i	y_i
0	12
1	30
2	28



Let's explore the effect of standardizing the data (unless you already know and remember its effect from your previous stats class!). Here are the steps:

1. Convert all the x-values to z-scores.
2. Do the same with the y-values.
3. Compute the mean and SD of each variable after standardizing.
4. **What is the moral of the story?!**

Put together with previous warmup!

Key observation:

Line of best fit in z_x-z_y coordinates is _____

Summary of key ideas

- In the 1st warmup we looked at the special case of scatter plots centered around the origin. For this case the best-fit line has the form: $\hat{y} = mx$.
- If the (x, y) data are known, m can be found by minimizing the sum of the square of the errors, which is given by the function: $f = \sum(y_i - mx_i)^2$
- To get the critical point: $f' = 0 \;\Rightarrow\; \sum -2x_i(y_i - mx_i) = 0$.
  This gives: $\boxed{m = \dfrac{\sum x_i y_i}{\sum x_i^2}}$
- In the 2nd warmup we learned that
  - "standardizing" a data set means converting all values into z-scores.
  - any standardized variable has mean=0 and standard deviation=1.
  - a scatter plot between any two standardized variables will be centered around the origin.
- A good way to think about linear regression is to combine these insights
  - Suppose we want to find the line of best fit for a given set of data points (x_i, y_i), for $i = 1, 2, 3, \ldots, n$.
  - Consider the scatter plot of the standardized form (z_{x_i}, z_{y_i}).
  - Then the line of best fit is: $z_y = mz_x$.
  - Notice that $m = \dfrac{\sum z_{x_i} z_{y_i}}{\sum z_{x_i}^2} = \dfrac{\sum z_{x_i} z_{y_i}}{n-1} = r =$ correlation coefficient!
    ($\sum z_{x_i}^2 = n - 1$ because the mean=0 and SD=1)
  - Thus, the line of best fit in standardized coordinates is: $\boxed{z_y = rz_x}$
- Finally, to transform the regression line back to the original coordinates, replace z_x, z_y by their known formulas: $z_x = \frac{x-\bar{x}}{s_x}, z_y = \frac{y-\bar{y}}{s_y}$.
- This gives: $z_y = rz_x \;\Rightarrow\; \frac{y-\bar{y}}{s_y} = r\frac{x-\bar{x}}{s_x}$
  Upon simplification, we get the familiar result: $y = \left(r\frac{s_y}{s_x}\right)x + \left[\bar{y} - r\frac{s_y}{s_x}\bar{x}\right]$