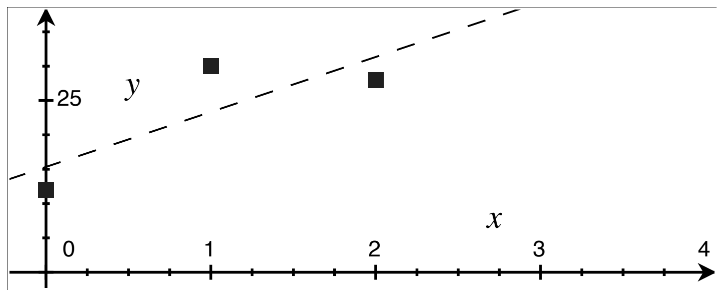
CW Exercise

Recall the following exercise from a recent class, where we found the line of best fit for a dataset containing only 3 points?

x_i	y_i
0	12
1	30
2	28



A very small dataset consisting of only 3 observations is shown in the table, together with a scatter plot of y vs x. We want to fit a straight line approximation to the plot, as shown by the dotted line.

Consider the straight line: $\hat{y} = mx + b$.
Our goal is to find numerical values of m and b that give the "best" straight line approximation. Here are the steps:

1. For each i, find the error: $e_i = y_i - \hat{y}_i$
   Each e_i will be a function of m and b (only!).
2. Compute the function: $f = \sum(e_i)^2$
   This function is the sum of the square of the errors.
3. Next, we want to minimize f – remember calculus?!
   Find the derivative of f with respect to b.
   Then, find the derivative of f with respect to m.
4. Set $\frac{df}{db} = 0$ and $\frac{df}{dm} = 0$, and solve simultaneously for m and b.
5. Well, then that is your best straight line!

CW Exercise (continued)

We want to extend that same strategy to 3 dimensions: Our dataset now contains 2 predictor variables (x, y) and 1 response variable (z). The goal is to find the plane of best fit using the least squares method – i.e., minimize the sum of the square of the errors. Here are the steps:

x_i	y_i	z_i
0	0	12
1	0	20
0	1	42
3	3	30

Assume the plane has the equation: $\hat{z} = b + mx + ny$

1. For each i, find the error: $e_i = z_i - \hat{z}_i$
   Each e_i will be a function of b, m, n.

2. Compute the function: $f = \sum(e_i)^2$
   This function is the sum of the square of the errors.

3. Next, we want to minimize f using calculus.
   Find the derivative of f with respect to b.
   Then, find the derivative of f with respect to m.
   Then, find the derivative of f with respect to n.

4. Set $\frac{df}{db} = 0, \frac{df}{dm} = 0, \frac{df}{dn} = 0$, and solve simultaneously for m, n, b.

5. Well, then that is your best-fit plane!

MLR: Key ideas

- Suppose we have a dataset containing k predictor variables $(x_1, x_2, \ldots, x_k)$ and 1 response variable (y).
- We want to model the relationship between y and the predictor variables using a "hyper-plane" of best fit

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

  where the $b_0, b_1, b_2, \ldots$, etc, are constants that we must find to best fit the data.
- How? In principle, it is straightforward to extend the exercise we just did in 3 dimensions: We formulate and minimize the sum of the square of the errors, which is a function of the form: $f = \sum (e_i)^2$
- Many ideas from simple linear regression extend readily to multiple linear regression:
  1. The assumptions & conditions
  2. Interpretation of slope(s)
  3. Residuals
  4. R-squared
  5. Inference strategies

MLR coverage: 4 key aspects

In this course we will cover the following aspects of multiple linear regression:

1. How to setup models with multiple predictors; how to interpret results.

2. How to do inference for MLR.

3. Model selection/refinement.

4. Model diagnostics.