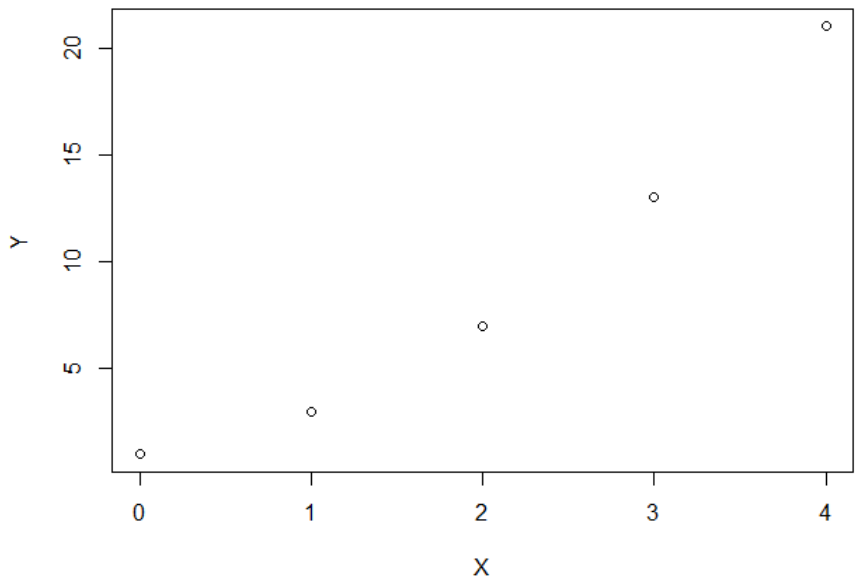
Computing regression parameters (closed form example)

The data

Consider the following 5 point synthetic data set:

```
X   Y
0   1
1   3
2   7
3   13
4   21
```

Which is plotted below:



What we need

We want the line that "best fits" this data set as measured by residual sum of squares -- the simple linear regression cost. We have a closed form solution that involves the following terms:

- The number of data points (N)
- The sum (or mean) of the Ys
- The sum (or mean) of the Xs
- The sum (or mean) of the product of the Xs and the Ys
- The sum (or mean) of the Xs squared

Then once we have calculated all of these terms, we can use the formulas to compute the slope and intercept. Recall that we first solve for the slope and then we use the value of the slope to solve for the intercept. The formula for the slope is a fraction with:

```
numerator = (sum of X*Y) - (1/N)*((sum of X) * (sum of Y))
denominator = (sum of X^2) - (1/N)*((sum of X) * (sum of X))
```

Note that you can divide both the numerator and denominator by N (which doesn't change the answer!) to get:

```
numerator = (mean of X * Y) - (mean of X)*(mean of Y)
denominator = (mean of X^2) - (mean of X)*(mean of X)
```

Hence, we can use either the sum or the means.

The formula in action

Method 1: (using sums)

- N = 5
- The sum of the Ys = 45
- The sum of the Xs = 10
- The sum of the product of the Xs and the Ys = 140
- The sum of the Xs squared = 30

So that:

```
numerator = [(140) - (1/5) * (45*10)] = 50
denominator = [(30) - (1/5) * (10*10)] = 10

```

hence:

```
slope = 50/10 = 5
```

Method 2: (using means)

- The mean of the Ys = 9
- The mean of the Xs = 2
- The mean of the product of the Xs and the Ys = 28
- The mean of the Xs squared = 6

So that

```
numerator = 28 - 9*2 = 10
denominator = 6 - 2*2 = 2

```

hence:

```
slope = 10 / 2 = 5
```

Then, we can use this computed slope to compute the intercept:

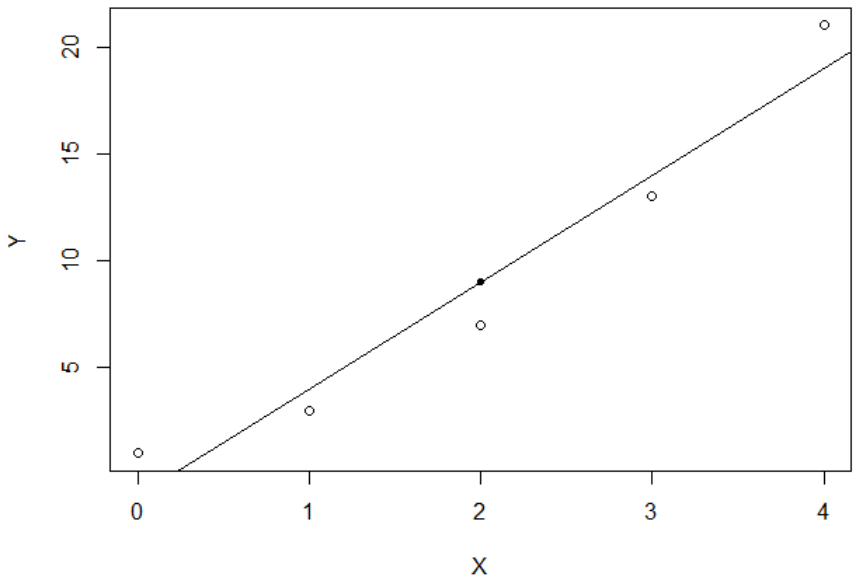
```
intercept = (mean of Y) - slope * (mean of X)
intercept = 9 - 5 * 2 = -1
```

(Food for thought: what if Y and X both have mean 0?)

In summary, we have:

slope = 5, intercept = -1

Finally we can add the line to our plot from above:



The solid black point included in this plot is the point (mean of X, mean of Y). You'll notice that this point falls exactly on the regression line!

(Food for thought: is this always true? Hint: try plugging in (mean of X) as input into prediction = intercept + slope * input) where you use the formula for intercept).

✓ Complete