Algorithm for Finding the Best-Fit Line

Suppose we have a set of points (more than one) in the X-Y plane. Usually there will not be a line that actually does go through all of them. We want to find the line that comes closest to passing through all the points.

Each point has coordinates (X, Y). We read through all the points in the set and calculate the following:

```
Count = the number of points

SumX = sum of all the X values

SumY = sum of all the Y values

SumX2 = sum of the squares of the X values

SumXY = sum of the products X*Y for all the points
```

Now we can find the slope M and Y-intercept YInt of the line we want:

```
XMean = SumX / Count

YMean = SumY / Count

Slope = (SumXY - SumX * YMean) / (SumX2 - SumX * XMean)

YInt = YMean - Slope * XMean
```

The equation for the line is:

```
Y = Slope * X + YInt
```

The algorithm will fail if the data points are all on one vertical line, so they all have the same X-coordinate. (When we try to find Slope, we try to divide by 0.)

If the data points are all on the same horizontal line, so they all have the same Y-coordinate, there is no dificulty, and we will have Slope = 0.

Example

Suppose we have the points (1, 1), (2, 3) and (3, 4). Then:

```
Count = 3

SumX = 6

SumY = 8

SumX2 = 14

SumXY = 19

XMean = SumX / Count = 2

YMean = SumY / Count = 2.666

Slope = (SumXY - SumX * YMean) / (SumX2 - SumX * XMean) = 1.5

YInt = YMean - Slope * XMean = -0.333
```

Thus for this set of points, the best-fit line is:

```
Y = 1.5 * X - 0.333
```

Notice that the best-fit line does not (in this case) actually pass through any of our original points.

Why is this the best line?

The theory of this is that we want to minimize the sum of the squares of the vertical distances from the points (X, Y) to the points on the line (X, Slope * X + YInt). That is, we have two unknown quantities Slope and YInt, and we have a function:

```
        Sum(Slope, YInt) = the sum of (Y - (Slope * X + YInt))**2 for all data points (X, Y)
```

We want to find Slope and YInt to minimize the function Sum.

This is a problem in the calculus of two variables and is beyond the level of the prerequisites for this course. The algorithm given above does solve the problem.

Why do we use the squares of the vertical distances instead of the vertical distances themselves? The answer is that some vertical distances will be positive and others negative, and we don't want them canceling each other out.

This is also sometimes called the "least-squares" technique.