Business Trend Analysis [Draft * 2008-09-23]

In business, as in life, it is usually helpful to be able to predict/create the future. The purpose of this tutorial is a little more modest though, since my intent here is just to show you how to find a 'best' straight line through a set of 2-dimensional data points that you can *then* use to predict or create the future! That best line, called the *trend* line, is used to establish or detect a limited, although important *trend* or direction of some process of interest. The idea of using a simple straight line is a *robust* way to represent the underlying complexity of many processes whose complete explanation needs to wait for additional information (or expertise!). Further, initial interest in some process might simply consist of whether it will increase or decrease some important metric, and so merit further investigation. A trend line might be able to supply that -to simply establish a *direction* rather than a magnitude.

> The trend line is of great value in any context where you want to use (a simplified version of) history to predict what happens next.

I am aiming at beginning researchers here and others who would like to see some simple explanations of the "least squares" approach to finding a *trend* line (or even just need a refresher on the characteristics of straight lines). I will apply the perspective of Exploratory Data Analysis (EDA) initially, to set the stage for the later calculations, then I will go through several alternative ways to find the equation of the trend line, that is, to get its defining slope and intercept. I have categorized those ways as:

- Plot the data set as a "scatter plot", eyeball the resultant graph and draw a freehand straight line through the points. That freehand line is your initial best estimate of the trend line and actually may be good enough! I would recommend you do this anyway, just to provide a check on later calculations. An example of a scatter plot is shown in Figure 1.
- Plot a 3-D picture showing least squares values as a consequence of choosing pairs of slope and intercept. (This is just for intuition, since this approach needs a math package to draw the picture, but from this picture you can almost pick out the particular slope and intercept that result in the smallest least-squares value and so the trend lines' slope and intercept.)
- Use cookbook algebra to calculate the parameters of the trend line, that is, its slope and intercept. If you are in a hurry, this is the way to go.
- Use calculus to derive the same results as the cookbook formulas, but with resultant psychological satisfaction.
- A follow-on illustration of using the medians of box & whisker plots as trend line data points.This is an EDA tool that will give you a rough way to not only show a trend, but the variation in a data set at each time value along the way.

Note that an optional section of this tutorial, "Correlation and Coefficient of Determination" on page 9, shows several ways to calculate how well your trend line fits the data, that is, the various correlation coefficients.

As a side note, you will see the phrase Exploratory Data Analysis (EDA) frequently in this document. The perspective of EDA is to use a mix of geometry, visualization, and algebra to gain special insight into questions. I have prepared several tutorials on the fundamentals of this perspective and references can be found at the end of this document. See John Tukey's work (Exploratory Data

Analysis, 1977) for a deeper look. O.k., enough preliminaries, lets cut to the chase!

A Business Research Question

Suppose you are a business/systems analyst for a consulting company and, at the request of your CEO, are asked for an estimate of what revenues she can expect from her training department over the next two months. You have available revenue data from the department over the last three months that looks like this (next month will be month 4):

month 1 @ $4000

month 2 @ $3000

month 3 @ $6000

After talking with the training department personnel, you think that calculating the *best* straight line through this given data, and then extending that line for two more months into the future, will be the most robust answer to the CEO's question. In short, the training department personnel are confident that these three months are indicative of what the near future will be like.Taking them at their word, you decide to calculate a "trend line" and extrapolate it for two more months.

Finding the *best* line through this data set means finding that unique line that is the closest possible to the three revenue values that are given. Note: when you find that best line, that *is* the trend line. As you know, two points determine a line so, unless that third point 'lines' up, the line can't go through all three, but we *can* find a line whose distance from the three points is minimized. That is, we can calculate the squared distance from each given revenue point to a *test* line and then find the particular *test* line that makes this sum the smallest. This is called a *Least Squares Line and in our context, this is the trend line.* Doing some calculations, as shown below, will determine the two parameters that will pin down the *trend* line, namely its slope and its intercept.(I will explain further as the example develops).

It is always a good idea to plot out whatever you can (this is the message of Exploratory Data Analysis), and, in this case I have created the graph "Training Department 3-Month Revenue Stream, 'A Scatter Plot of the Revenue Points'" on page 3. The horizontal axis is the time axis, in months, while the vertical axis is the $revenue amount (in thousands). For this case, at month 1 we see a height of "4", meaning $4000 revenue for month 1, $3000 for month 2, and $6000 for month 3. The graph is also a general example of a 'Scatter Plot', which is very helpful for getting an initial "feel" for the data's distribution.

Going 'Graphical' as a First Cut at Finding the Trend Line

Of course, for these three points you might just want to draw a *test* line by eye for starters. In fact, doing this is good idea in general, since this gives you a 'sanity' check on more sophisticated procedures you might want to try later. Notice that we can go a long way in determining the desired "trend line" just by using some simple semi-graphical techniques that are part of the Exploratory Data Analysis toolkit. You will see how drawing in a test line, setting up some simple sums of squares, and plotting those, will give us a graphical picture of the desired solution.

> In general, a useful initial approach to understanding opportunities/problems is to use visual tools and simple algebra to get an idea of what to do, and perhaps guides for how to do it, and then, use other math tools like calculus, linear algebra, and statistical procedures, to do the nitty gritty production calculations.

So let's start with some graphical insight to help construct this trend line.

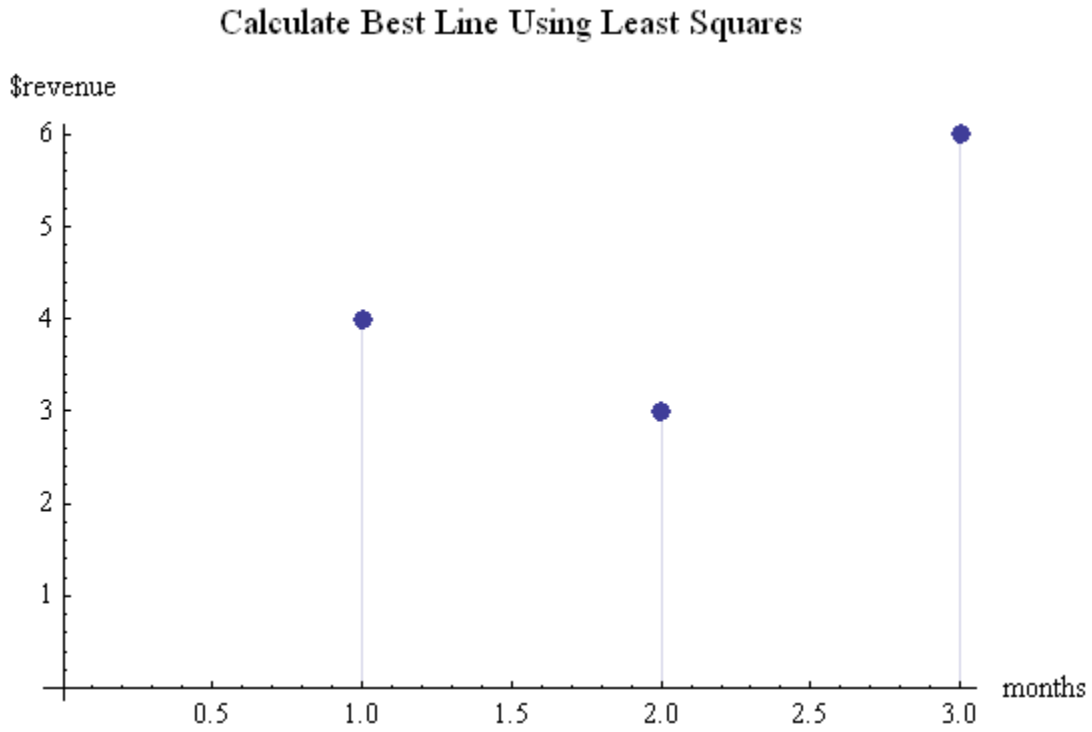


FIGURE 1. Training Department 3-Month Revenue Stream, 'A Scatter Plot of the Revenue Points'

Sketching a Test Trend Line

Starting with the three actual revenue points, I want to draw the closest line I can to these points. To help me visualize what I need to do, look at the next diagram, "Training Department Revenue Data with a Test Trend Line" on page 4, and see that I have drawn a *test* line that I can use to guide my calculations. In this diagram I have indicated the distance (point 'revenue' vertical coordinate minus line 'revenue' vertical coordinate) between the test line and the known revenue points by the symbol "d". Since some distances will be plus and some minus, it is convenient mathematically to work with the square of these distances, since this will make them all zero or better. The idea then is to find a line that minimizes the sum of these squared distances from that *test* line to the given points. (This is called "fitting a least squares line").

The General Equations of the Test Trend Line (An Exploratory Perspective)

The general equation for a line requires two parameter: its "slope" (rise over run, that is, for an increase on the horizontal axis (run), what is the change along the vertical axis (rise)) and where it intercepts the vertical axis is called its "intercept".[This is the most basic starting point for everything that follows, so be sure you follow this part closely]. For our example, the general *test* line can be written as: [where I have indicated the units in square brackets].

revenue[$] = slope[$/month] * month + intercept [$]

To give a little more detail to this equation, notice that the units on the 'slope' parameter would be $dollars per month and the 'intercept' parameter would have the units of $dollars. This general equation works for any values of month that you plug in, known or unknown, whole numbers or not. Of course, the dilemma is that we don't know either the actual slope or the intercept. Those two parameters are what we still need to find.

Setting Up to Find the Slope and Intercept

From the figure below, I have indicated the distances from a revenue point to the *test* line. What I really want is to make these squared distances as small as I can by picking the right slope and intercept for the *test* line. But again, at this stage of analysis we don't know the slope and intercept. In symbols, I can write out these distances as:

d1 = 4 - (slope * 1 + intercept)

d2 = 3 - (slope * 2 + intercept)

d3 = 6 - (slope * 3 + intercept)

You can see that we have three equations and two unknowns, the 'slope' and the 'intercept'.

If I square each of these distances and add them up I get:

Sum of Squares= $d1^2 + d2^2 + d3^2$

Notice that the distances d1, d2, and d3 will be both positive and negative, but squaring them makes them all positive, which is more convenient for the analysis.

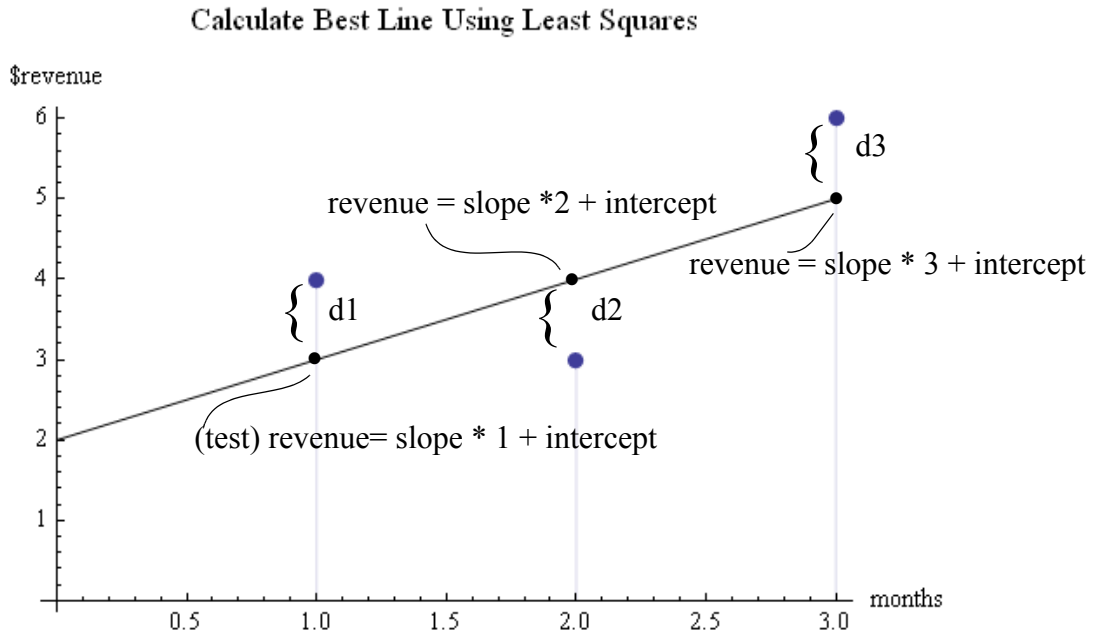


FIGURE 2. Training Department Revenue Data with a Test Trend Line

Going 3-D Graphical to Find the Slope and Intercept of the Trend Line

Lets start with a 3-D geometric approach to finding the slope and intercept of the best *test* line. *Remember, the best test line is the trend line.* In the diagram below, I have plotted the sum of squares value against various combinations of slope and intercept. The vertical coordinate is labeled "sum of squares" and represents the sum, $d1^2 + d2^2 + d3^2$ as noted in the previous section.

Looking closely, you can almost pick out the combination of particular slope and intercept that give the lowest point of this 'bowl'. That combination is what we are looking for. Can you almost see that a slope of "1" and an intercept of 7/3 = 2.33, gives the lowest point of the bowl? No, well don't be discouraged, this diagram was intended as just another intuitive perspective on the task!

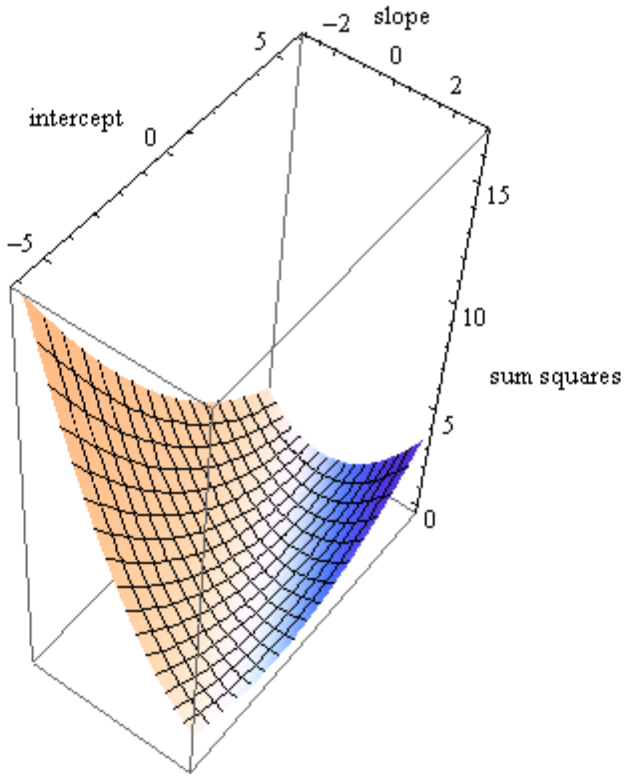


FIGURE 3. Sum of Squares Plotted against combinations of Slope and Intercept

The sum of squares depends on two parameters/variables, slope and intercept. If you could visualize ($d1^2 + d2^2 + d3^2$), "Sum of Squares" you would see a shallow (Japanese modern?) bowl shape and we would be looking for that lowest point of the bowl. That lowest point is where a particular slope and a particular intercept combine to make the sum of squares a minimum. In fact, this shape looks exactly like the picture above! As an aside, I plotted values of *slope* from -3 to +3 and *intercepts* from -6 to 6 since I had an idea of the possible ranges from the graphic "Training Department 3-Month Revenue Stream, 'A Scatter Plot of the Revenue Points'" on page 3. [You might want to think about how you personally alternate between visualization and calculation as you do problems].

What you are seeing here is a parametric 3-dimensional plot with the height at each (slope, intercept) pair given by Sum of Squares/50. (I scaled the vertical axis by 50. just to keep it on the page). I used the *Mathematica* 6.0 math package to make this plot.

Using the Cookbook Formulas to Calculate the Intercept and Slope

Most of the time you won't need the calculus since the algebraic formulas presented next reflect the results of the calculus. I will be a little bit general here so you can use these formulas for any simple regression problem (I show our specific numbers as well so you can follow along):

let the observed values be the vector y = {y[1], y[2],.. . ,y[n]}

let the corresponding independent/predictor variable be the vector x ={x[1], x[2], . . . , x[n]}

b is the slope (to be calculated)

a is the y-intercept (to be calculated)

xbar is the mean/average of the x values

ybar is the mean/average of the y values

The general formulas below work for any number of points:

General equations to calculate slope and intercept

$$\text{eq3:} \quad b = \frac{\sum_{i=1}^{n} (-\text{xbar} + x[i])\,(-\text{ybar} + y[i])}{\sum_{i=1}^{n} (-\text{xbar} + x[i])^2}$$

$$\text{eq4:} \quad a = \text{ybar} - b * \text{xbar}$$

For the tutorial example, specific calculations give:

for x={1,2,3}, mean/average =xbar = 6/3, where x[1]=1, x[2]=2, and x[3] = 3

for y={4,3,6} mean/average = ybar = 13/3 where y[1]=4, y[2]=3,and y[3]=6

Note: Its not necessary to subtract off the ybar in the equation below, it cancels out (of course you can put it in if you like)

$$b = \frac{(-\text{xbar} + x[1])\,y[1] + (-\text{xbar} + x[2])\,y[2] + (-\text{xbar} + x[3])\,y[3]}{(-\text{xbar} + x[1])^2 + (-\text{xbar} + x[2])^2 + (-\text{xbar} + x[3])^2}$$

$$b = \frac{(1 - 2)*4 + (2 - 2)*3 + (3 - 2)*6}{(-1)^2 + 0^2 + 1^2} = 1$$

$$a = \text{ybar} - b\ \text{xbar} = 13/3 - 1 * 2 = 7/3$$

FIGURE 4. Algebraic calculations for slope and intercept

Using the Calculus to Find the Slope and Intercept of the Trend Line

While you could get the optimal slope and intercept from a careful inspection of the earlier 3-D picture, or by applying the algebaic formulas noted above, a most valuable message is the need to learn general tools that will allow you to find maximums or minimums, that is, *optimums*, for various functions representing measures of interest. In this trend line finding case, the optimum consists of finding the *minimum* of a least squares function- at the bottom of a bowl like structure. Now I would like to do an algebraic/calculus approach to finding these minimizing parameters.

The *calculus* is designed to *efficiently* find that particular slope and associated intercept that combine to yield the lowest value of this sum of squares. It turns out that my 'eyeball' test line is pretty close to the 'best' line. After some calculations, we will see that the (scaled) optimum slope is "$1 per month" and the scaled) intercept is "$7/3". So, the best trend line turns out to be: (inserting the dimensions of each term and the scale factor of $1000)

*Trend Line: revenue[$dollars] = 1000($dollars per month) * month + 7000/3 ($dollars)*

This tells me, for any month, how much revenue is to be expected. Notice that this line can't exactly go thru the three given points of months 1,2, and 3, but it's as close as you can get. Be careful in interpreting this line though, since it will match the past three months as accurately as possible, but the future is another matter!

Finally, the CEO Gets Her Answer!

To answer the CEO's question:

- (anticipated) Revenue for month 4 = 1* 4 + 7/3 = 19/3 or $6,333.33
- (anticipated) Revenue for month 5 = 1 * 5 + 7/3 = 22/3 or $7,333.33

I continue this analysis in a later section and calculate the correlation coefficient to see how good this linear fit is to the data, seeSee "Back to the CEO's Trend Analysis Review" on page 12.

The Details of the Calculus

For those who would like to know how I got the slope and intercept values using the calculus, read on!

Keep in mind that to find minimums or maximums of a function of a variable, you find where the function has a horizontal tangent line. This occurs when the derivative (rate of change) of that function is zero. So, setting the derivative to zero and solving for that value of the variable that makes it zero, locates a max or a min (there are always exceptions but don't worry about those right now!). In our case the function to be considered is a sum of squares depending on *two* variables, *slope* and *intercept*. So, we take the derivative of the Sum of Squares function with respect to each variable in turn, set that derivative to zero and write down the resulting equations. The result is a *set* of equations called the "Normal Equations" that, when solved, will give us the minimizing slope and intercept.

You will see all this below:

Here is the sum of squares I want to minimize (that is, find the least value of this sum)
To do that I could try different values for both slope and intercept until I get the smallest value
There is an easier way though, using the calculus, and that's the route I take below.

```
ss = d1^2 + d2^2 + d3^2
(6 - intercept - 3 slope)^2 +
 (3 - intercept - 2 slope)^2 + (4 - intercept - slope)^2
```

Now I take the derivative of 'ss', with respect to the intercept variable (keeping the slope var constant.)

```
derivativeIntercept = D[ss, intercept]
-2 (6 - intercept - 3 slope) -
 2 (3 - intercept - 2 slope) - 2 (4 - intercept - slope)
```

I set this derivative equal to zero since I know the tangent line at the bottom of the bowl is horizontal, that is, zero.Doing that gives me the next equation

```
eq1: 3 intercept + 6 slope == 13
```

Now take the derivative of 'ss' with respect to the slope, holding the intercept variable constant. Setting that expression to zero yields the second equation that we need.

```
derivativeSlope = D[ss, slope]
-6 (6 - intercept - 3 slope) -
 4 (3 - intercept - 2 slope) - 2 (4 - intercept - slope)
```

```
eq2: 3 intercept + 7 slope == 14
```

**Subtract eq1 from eq2 and you will see that the slope is "1".
**Plug that value back into eq1 and find that the intercept is "7/3".

For you calculus buffs, I used the chain rule as well as the derivative of a power to calculate. these derivatives.

(To check my work I used a math package called *Mathematica* that does literally anything I can specify!)

FIGURE 5. Calculating the Slope and Intercept values Using Least Squares

Trend Lines With Multiple Values at Each Time Point (Time Series)

[This next discussion is just a preliminary comment on analyzing batches of data, and is intended to be expanded by investigating more deeply some of the EDA and calculus tools at a later time. I have written several EDA tutorials along these lines that might be worth taking a look at, see references].

In the business example that started off this tutorial, I worked with three data points representing the revenue outcomes for a single department. Suppose though, that there were eight departments that reported net revenue each month. Now I have eight data points for each month. How can I handle that? A good way is to take each data set, calculate the median, and use that as the point for that

month. Given this single point for each month, that analysis matches what we have done above. Of course, you could plot each department's revenues on the same graph and do comparisons that way. (Showing your clients important features of the data, that they might have otherwise missed, is the signature of a top rate data analyst).

Suppose the three data sets for each month are as follows.

month1 = {1,2,3,3,5,6,6,9}

month2 = {2,2,3,3,3,6,6,8}

month3 = {3,3,4,4,8,8,9,10}

So, just by taking the median of each data set, we can reduce this problem to the previous approach involving single data points per month. You will notice that I made the median values match the simple example we started with, that is, {4, 3, 6}. So for these data sets, the analysis is identical to the earlier example. To get more out of an example like this though, you might want to learn more about the details of Box and Whisker Plots in particular, and time series in general.

Box & Whisker Plots of Three Months of 8 Departments

The box and whisker plots below show the median revenue as the cross bar within each box and matches the values we have been using all along. The extent of the box shows the interquartile range while the extended lines go out to the max and min values in each data set. More detail can be found in other tutorials by myself or from the extensive literature on EDA.

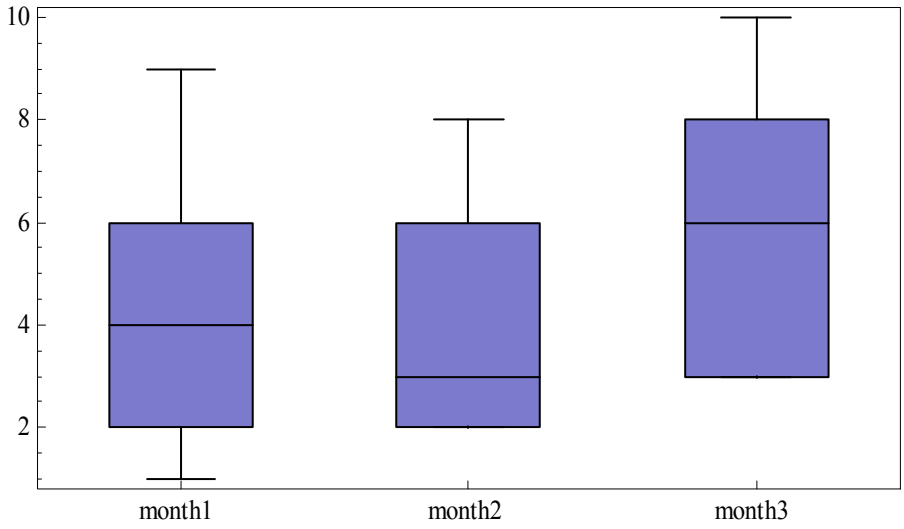


FIGURE 6. Box & Whisker Plot of Three Months Revenue for Eight Departments

Easing Into the Geometry of Regression Analysis

Thinking of the simplest way to *geometrically* develop and present the regression equations, I have made up the exercise below using just three points for the independent x-values and associated dependent y-values. The reason I want to do this is that I want to use x-values to predict y-values. That is, if I pick an x-value that's not in the given set, what would be a reasonable y-value? For example, for the data given below, what would be a good guess for the 'y-value' if I considered an 'x-value' of 4? Since that x-value is not in the given data set, I will have to estimate its effect on y,

that's where the straight line comes in. That is, if I have the line equation, I can simply plug in 'x=4' and calculate 'y'. Check out the diagram below for a visual illustration of the following text description.

From the earlier discussions in this tutorial you know that a line needs a constant and a slope for its specification. So, I need to estimate what that *constant* (also called the y-intercept) and *slope* should be to get as close as possible to the three given y-value points.

Some Super Simple Data for Illustration

Ok, here are the made up sets. The 'raw' designation means the original data, as given.

xraw = {1, 2, 3}, yraw = {4, 8, 9}

xbar= mean[x] = 2, ybar = mean[y] = 7

Centering the Data

Now, it will make the analysis much simpler if I break up the raw data into its constant part (its mean) and its deviation part. Taking each xraw value and subtracting off its mean gives me the deviations which I call the centered variable: That is, *xc* is the centered x variable while *yc* is the centered y variable. Notice that these variable are actually 3-component vectors and can be plotted in 3-dimensional space as in the diagram below.

Ok, here is the x -deviation vector that I will call xc.

xc = {-1, 0, 1}, yc = {-3, 1, 2}

I get these centered variables by performing the following operations, just subtract the means.

{1,2,3} - {2,2,2} = {-1, 0, 1} and {4,8,9} - {7,7,7} = {-3, 1, 2}

So using this idea of breaking up the raw variables into their constant part and deviation part gives lets me write

for the xraw values:

1 = 2+ -1

2 = 2 + 0

3 = 2 + 1

or, in general, xraw = xbar + xc

Similarly I can write the yraw values as:

yraw = ybar + yc

4 = 7 - 3

8 = 7 + 1

9 = 7 + 2

Getting as close as possible to yraw by using a + b xraw*

eq1: yraw ≅ a + b * xraw

this equation #1 says that yraw is going to be approximated by the expression on the right, a straight line.

To carry out this approximation I am going to re-write eq1 using the idea that yraw = ybar + yc and that xraw = xbar + xc. Notice that these are equalities while eq1 expresses only an approximation.

ybar + yc ≅ a + b* (xbar + xc)

Ok this eq1 with both sides re-expressed. It is still not an equality, but instead states that I want to

get as close as possible to the left hand side of the ≅ symbol by calculating the right hand side: a + (xbar + xc).

Notice that *ybar* on the left and *a + b * xbar* on the right are *constants*. Those constant vectors, with all components equal, must lie in the *constant* space spanned by the vector u1 = {1,1,1}. The *deviation* parts are *yc* on the left and *b* xc* on the right. These vector lie in the complement of the constant space, that is, they lie in a perpendicular space.

Can you see from the diagram below that the deviation vectors xc and yc are perpendicular to the constant space containing xbar and ybar. Further, can you see that if you add xbar + xc you will get the xraw vector! Similarly, ybar + yc adds up to the yraw vector. What is of most importance is that the xbar and ybar vectors are in a space that is perpendicular to the space in which the xc and yc reside. This means we can calculate what best fits the constants *separately* from what best fits the deviations. That's a really big deal, trust me.

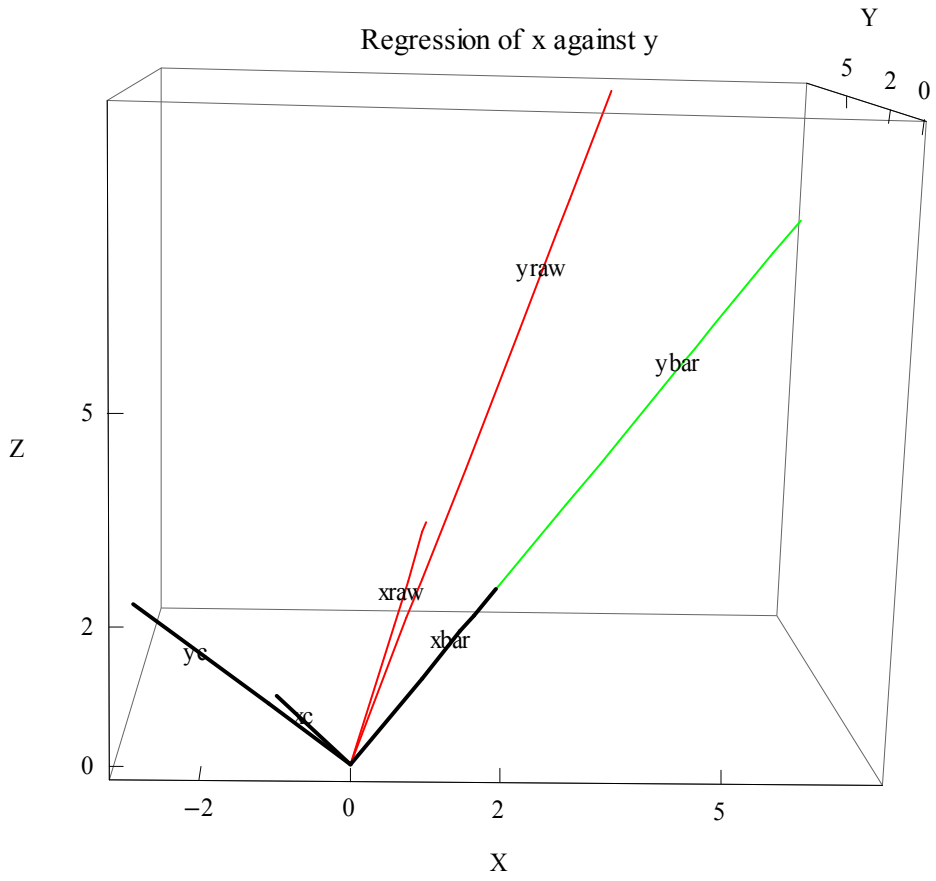


FIGURE 7. Constant Vector Space and Its Complement, the Deviations Space

The Constant Space is Perpendicular to the Deviation Space

It turns out that the space in which xbar and ybar lie is perpendicular to the space in which xc and yc live as you can see from the diagram. These spaces are independent of each other which indicates that our task breaks down into two parts:

Part I; *Get as close as possible* to constant y- values, represented by ybar, by adjusting a + b * xbar

Part II: *Get as close as possible* to the deviation y-values represented by yc, by adjusting b* xc.

Constant Part must match as close as possible (adjust 'a' and 'b')

ybar ≅ a + b * xbar

Deviation Part must match up as close as possible (adjust 'b')

yc ≅ b * xc

So, I'll work on this part first since I can find 'b' easily:

Note I will use the symbol "•" to stand for the Dot Product. (See the tutorial *Vector Spaces and Vector Operations* on this site for derivation and illustrations of the Dot Product).

To find 'b

Geometrically, what you want to do is to extend xc by some multiple until it is exactly underneath the tip of yc. (Mentally do that using the diagram). That multiple is the 'b' we are looking for. The condition that guarantees that b * xc is exactly underneath yc is that the vector from the tip of b * xc up to yc is perpendicular to xc. That vector is commonly called the error vector and is equal to yc - b* xc.

This perpendicular geometric condition means that the dot product of b* xc and the error vector is zero, since they are at 90 degrees from each other.

so: xc • (yc - b * xc) == 0 (this condition results in the so-called normal equations, where 'normal' means perpendicular)

Solving for 'b' yields (you have seen this before in eq3: of Figure 4 on page 6)

b = (xc • yc)/ (xc • xc) = {-1,0,1} •{-3,1,2}/({-1,0,1} •{-1,0,1}) = 5 / 2

now solve for 'a' to get

ybar = a + b * xbar

a = ybar - b xbar = 7 - 5/2 * 2 = 2

Finally:

yraw = 2 + 5/2 xraw

done!

Back to the CEO's Trend Analysis Review

(I have included a simpler discussion of this geometric approach in a subsequent section, (See "Easing Into the Geometry of Regression Analysis" on page 9. below since it was developed later as an in-class exercise).

Now we move a little closer to saying how good the fitted trend line actually is. Take a look at "The Geometry of Regression and the Correlation Coefficient" on page 14, This is a visualization of the vector space in which the observations lie. If the revenue data points closely hug the line we calculated, then the error vector will be small and so the angle denoted as theta, will be small. The

cosine of that angle is the *correlation coefficient*. From your knowledge of trig, the cosine of an angle goes from -1 to +1. Values close to -1 or +1 are 'good', suggesting that knowing the month helps to know the revenue. Values close to zero suggest that knowing the month is not helpful in knowing what the revenue will be. Values in between require additional data and informed judgment!

The square of that cosine value is called the *coefficient of determination*. This last parameter tells us the proportion of the total variation accounted for by the fitted line.

> *Cut to the Chase:* The correlation coefficient, 'r', is the Cosine of the angle between the vector (revenue-revenueBar) and the vector (month-monthBar). The Coefficient of Determination (R^2) is the square of the correlation coefficient.

How Close Do the Observed Points Hug the Regression Line? Correlation Knows!

[Preliminary knowledge required: To find this correlation coefficient, by the geometric method presented below, you will need to know what a *dot product* is, but that's it.

> *Micro-Refresher:* Let V = {v1,v2,v3} and W = {w1,w2,w3}, |V| = length of V and |W| length of W, then the dot product is defined as: V•W = v1 * w1 + v2 * w2 + v3* w3 = |V| * |W| Cos[theta], where theta is the angle between the vectors.[You could also check out the tutorial *VectorSpaces* for dot products and *TrigNotes* for the details of trig functions if you are a little rusty.]

Once we have the line fitted to the data points (admittedly a tiny sample), we could go on and ask: How good is this fit? How much does knowing the month value explain the revenue value. Does knowing the month help at all? You can see that if all the revenue values were say $6000, (a horizontal line would fit these data points), then knowing the month wouldn't tell you anything new and the *correlation* between month and revenue would be zero. In our case though, it looks like knowing the month is of some help in estimating revenues. The CEO sure hopes so!

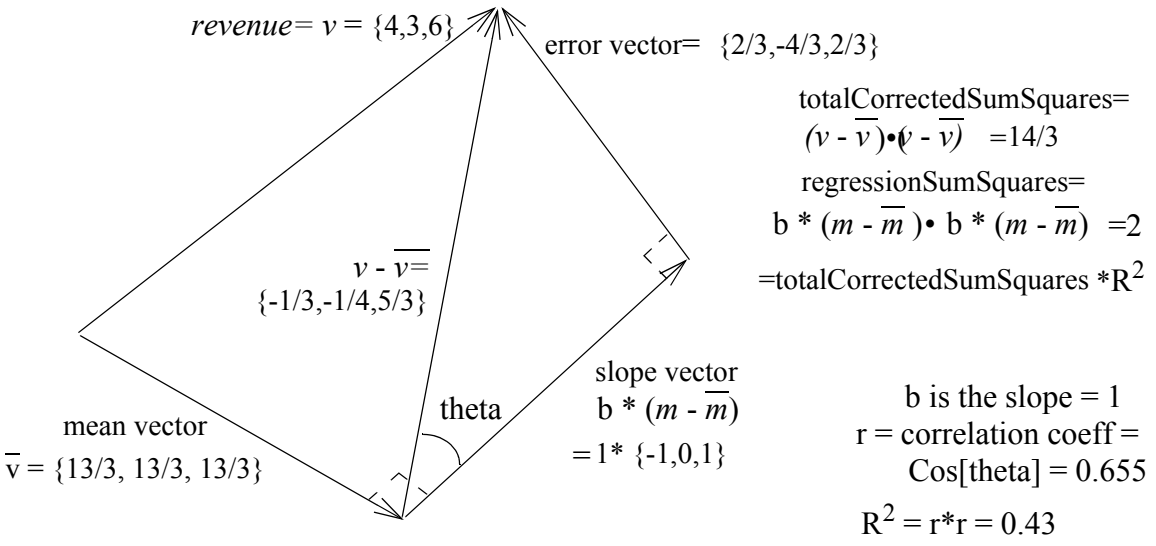
First, let me find some averages and lengths that we will need and some abbreviations to keep the drawings uncluttered. It is helpful in analyses and calculations to subtract off the mean of a data set which is called centering the data.

m = {1, 2, 3} months

$\overline{m}$ = (1+2+3)/6 = 2, average of the month values,

v= {4, 3, 6} revenues per month

$\overline{v}$= (4+3+6)/3 = 13/3 average of the revenues.



The orthogonal form of the regression line is: revenue = $\bar{v}$ + b* ($m - \bar{m}$) = 13/3 + 1* (month -2)
The non-orthogonal form is: revenue = 7/3 + 1 * month (this is the same as Trend Line of Least Squares)

FIGURE 8. The Geometry of Regression and the Correlation Coefficient

The picture above shows the geometry of the regression analysis. Note the right angles between the mean vector and the slope vector and the error vector. These right angles allows me to say, for example that v = mean vector + slope vector + error vector, and most importantly, allows me to use Pythagoras' theorem to equate sums of squares.

Correlation Coefficient and Coefficient of Determination

The Correlation Coefficient (denoted as 'r') is *defined* as the:

Cosine of the angle between $v-\bar{v}$ and $m - \bar{m}$.

That is, the definition of 'r' is the dot product of the above two vectors.

r = $v-\bar{v}$ • $m - \bar{m}$ = *Cos[theta]*

where theta is the angle between the two vectors.

In our case the dot product gave me 0.655, so:

r = 0.655 and, knowing the Cosine of the angle enabled me to calculate theta, which worked out to be 49 degrees. This means the regression line doesn't hug the data very closely, since a zero angle would be much better!

Finally, the Coefficient of Determination

To go a little further, r^2 is called the *coefficient of determination* and is usually written as R^2

As you will learn (or already know), the total corrected sum of squares is given by the dot product of $v-\bar{v}$ with itself to yield= 14/3, while the regression sum of squares comes from the dot product of b * ($m - \bar{m}$) with itself to yield = 2. The usefulness of the R^2 is that it represents the proportion of total variation accounted for by the regression sum of squares. That is, length of $v-\bar{v}$ squared times R^2 yields the regression sum of squares. For out case, R^2 is $(0.655)^2 = 0.43$. This means that 43% of the total variability is explained by the regression line I calculated. Not so hot!

Summary

This tutorial used the ideas of EDA (Exploratory Data Analysis), Calculus, and a bit of Linear Algebra to answer a common business question - "what can we say about the near (linear) future"? In our example question we analyzed a graph of three revenue points, found the best fit line to these points, and extrapolated that line to answer the question: "what will the next two month's revenues be"? The interpretation here is that we constructed a "trend line" that we then used to estimate two months in advance. As a follow-on, I showed the geometry of the trend line in its natural setting. From the vectors representing *revenue* and *months*, I calculated the angle between them. The cosine of that angle is the correlation coefficient and measures how close the vectors are to each other.

Among topics touched on by this tutorial were: EDA, time series, box plots, scatter plots, trend lines (linear regression), maximum and minimum determinations using the calculus and the geometry of the correlation coefficient.

So, try out these ideas in your everyday work and surprise everyone with your grasp of future trends!

References

Rucker, Rob. (2007) *Exploratory Data Analysis Introduction,* available from the author.

Ibid. (2007) *Five Number Summaries & Box Plots,* available from the author.

Ibid. (2008) *How Good is Your Trend Line?*, (in preparation) available from the author

Tukey, John (1977) *Exploratory Data Analysis*, Addison Wesley