

Solving Linear Regressions $\hat{Y} = a + b \cdot X$, using geometry to get a & b

**draft 2009-08-20, rob r. **



This note is to get you going with the simplest possible *linear regression* example so you can *understand* the vocabulary and the algebra you see in textbooks. I will use simple geometry to show how to calculate the regression coefficients (b =slope and a =intercept) that fit a straight line approximation to your data set of Y 's, predicted by your data set of X 's. Also, as a bonus, the *correlation coefficient* will emerge naturally as a feature of the geometry as we go along.

The *prediction* line will be denoted as the *vector* equation: $\hat{Y} = a + b \cdot X$. Notice that the equation is *not* $Y = a + b \cdot X$, since no straight line will exactly pass through all your Y data points, so, we will have to approximate this set of Y 's with the values of the closest straight line possible, that's our $\hat{Y} = a + b \cdot X$. The X 's will be considered the *independent* set of values while the Y 's will be considered as the *dependent* set.

■ Variable Space Representations (Scatter plot viewpoint)

You will see from the two-panel diagram below all of the elements of a regression problem, and its geometric solution. I show two complementary approaches to understanding and solving regression questions. The left hand panel shows the *scatter-plot* approach using what is called *variable space* representation. It is called that since the axes are labeled with the *variables* under consideration, X and Y , and the observations are placed on this variable grid using their X - Y coordinates. Each $\{X, Y\}$ point in space now represents an *observation*. This scatter plot perspective is good for seeing what patterns the *observations* form, although it does obscure what the pattern of the *variables* are, which brings me to:

■ Subject Space Representations (Vector space viewpoint)

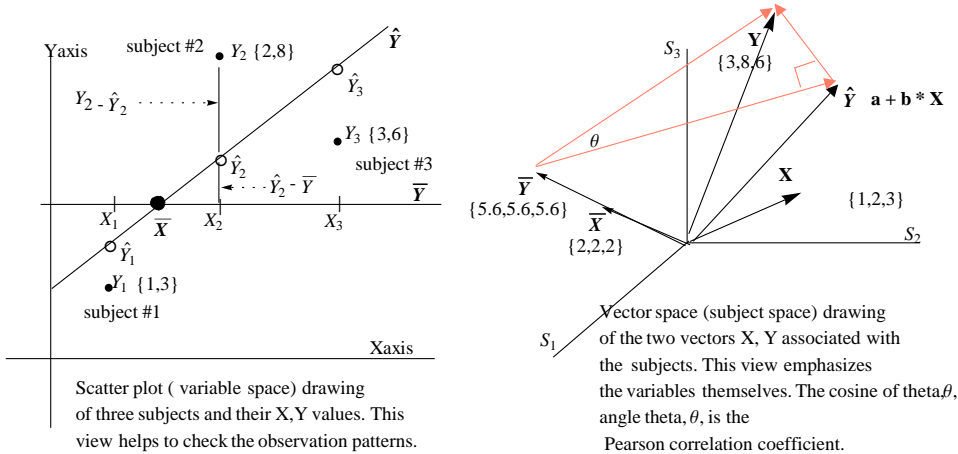
Complementing the scatter plot view, the right hand panel shows the second approach using what is called *subject space* representation. Here, each subject (e.g. each survey respondent or some measures taken on an entity) scores a certain value on X and Y . What is plotted in this representation is the X scores, taken as a vector, together with the Y scores taken as a vector. Now, the axes for this graph use the *subjects* rather than the variables. This view allows you to concentrate on the *variable* patterns, rather than the observation patterns. Both variable space and subject space diagrams are essential in order to get a comprehensive view of how variables and subjects interact and, it is very helpful to be able to track back and forth between them

Cut to the Chase

For those with very little time, consider the table below showing a few made up values for subjects S_1 , S_2 , and S_3 . (For those with more time, consider reading the sections following Cut to the Chase). Let X and Y be the values associated with each subject as shown in the table below. X is assumed to be the independent variable and Y is dependent, that is, knowing the X values should allow some prediction (with more or less accuracy) of the Y values. We would say that X is a predictor of Y , or, in the best possible world, X causes Y . So this table is read as: $X_1 = 1$, $X_2 = 2$, and $X_3 = 3$. $Y_1 = 3$, $Y_2 = 8$, $Y_3 = 6$. These are approximately keyed to the diagram below. Notice that reading the table 'across' maps into the variable space view (left panel) while reading the table 'down' maps into the subjects space view (right panel).

	X	Y
S1	1	3
S2	2	8
S3	3	6

■ Comparison between Variable Space (left panel) and Subject Space (right panel)



From the left panel *scatter plot* you can see a pattern of the *observations*, roughly moving up and to the right which means that X is roughly *positively correlated* with Y. (Of course this is a toy data set so not too much should be read into this, but this is what you would look for). Also, with more data, you might see that a more complex curve would fit the observations better than a straight line. This means that you should look for a different model or do some transformation that would bring the variables closer to a linear relationship. It is also the case that the regression line always goes through the point $\{\bar{X}, \bar{Y}\}$, as indicated by the large black dot. (I will derive this later in the tutorial, but notice for now that if \bar{X}, \bar{Y} are both zero, then the regression line goes through the origin, $\{0,0\}$, which means that the intercept "a" would have value zero).

From the right hand panel, you can see the pattern displayed by the variables themselves. This will allow you to see directly how close the X vector is to the Y vector. Closer is better since I can get a better prediction of Y using X. From the 'red' triangle, which turns out wonderfully to be a *right triangle*, you can get an idea of how helpful knowing X is in order to estimate Y. A measure of this closeness is to get the cosine of the angle between $\mathbf{Y} - \bar{Y}$ and $\hat{\mathbf{Y}} - \bar{Y}$. That cosine is the ratio of the lengths of the adjacent leg of the triangle over the hypotenuse length and is called the *correlation coefficient*, r_{yx} . This is also known as the *Pearson product moment correlation coefficient* and its square is the *coefficient of determination*. Note: Digging into the interpretation of this 'red' right triangle will allow you to extend your understanding beyond these two vectors, X, Y to multiple vectors and multivariate analysis with no change in perspective. This triangle also illustrates ANOVA in its simplest format. This 'triangle' knowledge will also be required when you venture into areas such as path analysis and structural equation modeling (SEM), but that is for another tutorial!

■ How to get from the left panel of individual components, to the right hand panel of vectors

Look at the diagram and trace $Y_2 - \bar{Y}$
 $Y_2 - \bar{Y} = (Y_2 - \hat{Y}_2) + (\hat{Y}_2 - \bar{Y})$. You can do the same for each of the $\{X, Y\}$ points ..

$$Y_1 - \bar{Y} = (Y_1 - \hat{Y}_1) + (\hat{Y}_1 - \bar{Y})$$

$$Y_2 - \bar{Y} = (Y_2 - \hat{Y}_2) + (\hat{Y}_2 - \bar{Y})$$

$$Y_3 - \bar{Y} = (Y_3 - \hat{Y}_3) + (\hat{Y}_3 - \bar{Y})$$

In general, you can see that $Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y})$ for each point,

and these components make up the three vectors: $Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y})$

These three vectors are shown in the right hand panel as the 'red' vectors. It will turn out that these components are mutually perpendicular thus allowing a divide and conquer approach to regression solutions. That is, the regression line \hat{Y} breaks up the total deviation, $Y - \bar{Y}$, into two perpendicular components.

■ Standard textbook notation between left and right panels

As you read the linear regression section of your textbooks, you will see algebraic formulas for the various components. Below I show three of these equivalences. More connections are shown later.

$$\text{Sum Squares errors } SS_{\text{error}} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}}) \cdot (\mathbf{Y} - \hat{\mathbf{Y}})$$

$$\text{Sum Squares regression } SS_{\text{reg}} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) \cdot (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})$$

$$\text{Sum Squares total } SS_{\text{total}} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = (\mathbf{Y} - \bar{\mathbf{Y}}) \cdot (\mathbf{Y} - \bar{\mathbf{Y}})$$

■ A Stripped Down Look at the Regression Triangle

Here is another look at the regression triangle. The upper-case letters denote the raw variables in vector format while the lower case letters represent the vector modeled, centered variables, that is, the raw variables with their means subtracted. Dealing with centered variables means that the y-intercept value 'a', of the regression line is zero and so we can just concentrate on a line through the origin, making it easy to find 'b'.

Note: $x = X - \bar{X}$ and $y = Y - \bar{Y}$ where x and y are the 'centered vector variables'.

Looking at the 'centered' triangle below, the regression task is to find the scalar 'b' so that $b * x$ is as close as possible to the tip of y . That means that $b * x$ must be perpendicular to the 'red' error vector e since that is as close as we can get to y , given that we can only scale x . That condition allows me to calculate b from the geometry. If $b * x$ is perpendicular to $e = y - b * x$ then the dot product is zero. That condition determines b .

$$\mathbf{x} \cdot \mathbf{e} = 0 \text{ or writing it out, } \mathbf{x} \cdot (\mathbf{y} - b * \mathbf{x}) = 0, \text{ using the distributive property of the dot product, I get: } \mathbf{x} \cdot \mathbf{y} - b * \mathbf{x} \cdot \mathbf{x} = 0$$

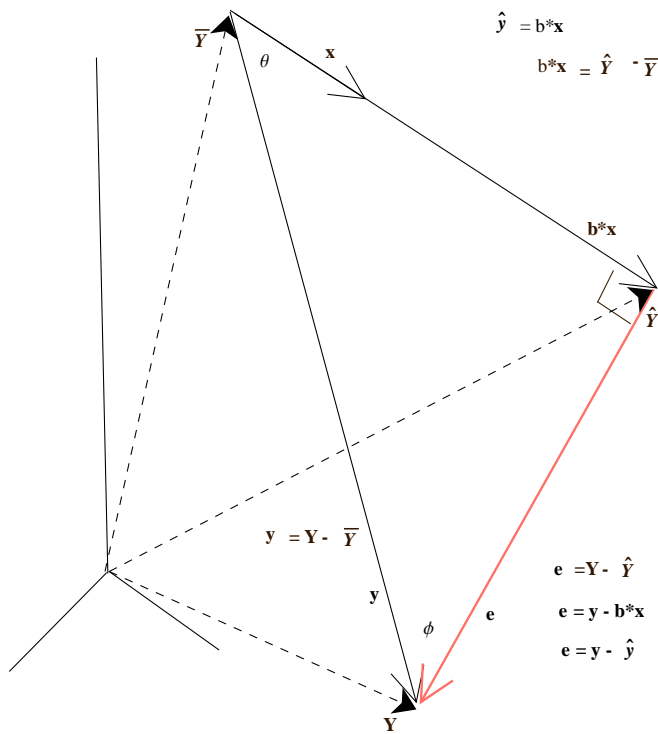
$$b = (\mathbf{x} \cdot \mathbf{y}) / \mathbf{x} \cdot \mathbf{x}$$

To solve for the original regression equation involving the raw data, we have:

$$a = \bar{Y} - b * \bar{X}$$

So the original equation, using the raw values is

$$\mathbf{Y} = \bar{Y} - b * \bar{X} + b * \mathbf{X} \text{ where } b \text{ is the same value as found from the centered variables.}$$



■ **Interesting Observation**

The reader will find out that the triangle shown above is simply the centered triangle with each corner displaced by \bar{Y} . That is, the centered right triangle consists of sides: y , $b \cdot x$, and e while the triangle shown above is the centered triangle + \bar{Y} .

Note that I can write this as:

$y = Y - \bar{Y}$,

$b \cdot x = \hat{Y} - \bar{Y}$, and

error vector $e = y - b \cdot x = Y - \hat{Y}$.

So, the triangle shown in the diagram above is simply the centered triangle displaced by \bar{Y} .

(Whoa, what about e , that isn't displaced by \bar{Y} !) Well, if you write out $Y - \hat{Y}$ as $Y - (a + b \cdot X)$ then you will see, after replacing "a" by $\bar{Y} - b \cdot \bar{X}$. Check it out!

■ **Trig notions and correlation connections**

If you look closely at the triangle you will see two angles noted, theta (θ) and phi (ϕ). The cosine of θ is the correlation coefficient of X and Y, usually designated as r_{yx} while cosine ϕ (which is the sine of θ) is the correlation coefficient of Y and Error, r_{ye} .

Since $\text{Cos}[\theta]^2 + \text{Sin}[\theta]^2 = 1$, we have $(b \cdot x) \cdot (b \cdot x)/(y \cdot y) + e \cdot e/(y \cdot y) = 1$ which translates to $\text{SS}_{\text{regression}}/\text{SS}_{\text{total}} + \text{SS}_{\text{error}}/\text{SS}_{\text{total}} = 1$

You do the math!!

**** * End Cut to the Chase Section**

Detailed Discussion

The next sections are a bit more technical since I would like to show you where the above *Cut to the Chase* conclusions come from.

The scatter plot is a valuable way to look at the data if you want to focus on the just the pattern of the observed Ys. The approach I have taken in the discussion below though, focuses on the variables, X and Y, represented as vectors. Both approaches are valuable, but the vector approach allows a bit more insight into the inter - relationships of the *variables* as opposed to the interrelationships among the *observations*. So, that's the way I'm going work out the regression equations below.

■ Background Notes

The plan for a regression analysis is to predict the next value (s) of a previously observed variable based on one or more influencing variables. The connection between the influencing variables and the observed variable is in the form of some kind of relation, maybe interpreted as causal, but perhaps only correlational. When you use the influencing variables (called the predictors) to construct a line that approximates the observed values, you are doing linear regression. When you only have one predictor and one observed variable, it is called *Bivariate Regression*, that's what this note is about. From another perspective, it is common to use this kind of analysis to try and determine if there is some kind of causal connection between the predictors and the observed variable where prediction by itself is of little interest. (That kind of investigation is carried out in a subsequent tutorial on this web site called *Path Analysis* which is currently under construction).

All my focus in this note is geometrical, since that is the way statistics is most understandable and actually the way multivariate statistics was originally presented by Sir Ronald Fisher in the 1920' s and 30' s. Over time though, the need for quick, efficient formulas and computer oriented calculations, pushed the underlying geometry aside. Algebra and computer packages took the place of geometric insight. Today, we all use computer packages (think Excel, SAS, or SPSS) that produce reams of numbers with a mouse click. That's fine and I myself use computer packages routinely, but, knowing the underlying geometric structure gives me insight into the algebraic equations and the computer outputs.

The heart of the geometric approach is to look at data sets as vectors and to embed those vectors inside their natural habitats, called vector spaces. This perspective gives you a rich tool set to understand and develop all statistical results. The vector approach is the one I will use here but if you don' t know about vectors, you can start now, and, as I mentioned above, check out my tutorials or look on the Internet.

(See my web site milagrosoft.com and the Vector Operations Quick Look tutorial for other helpful background information about vectors and vector spaces).

■ Predicting Y (observed variable) from X (an influencing , predictor variable)

Let me set up a fanciful context here, just so the numbers have some meaning. Suppose you are a *Lean 6-Sigma for Services* Green belt looking to make a name for yourself and move up to Black! How about if you could predict what next month's company revenue, or the month after, will look like? You'd be a hero!

If you have some reason to think that a straight line might be an adequate representation of the underlying process that's driving revenue, and you have a little historical data, then *linear regression* is just the technique for you. Revenue is what you want to predict, but what are you going to use as a *predictor*? Read on!

Two - Variable example (Bivariate Regression)

If my observed data were a revenue stream consisting of 5000, 3000, and 7000 USD over months 1, 2, and 3, I could write that as a vector (scaled by 1000) :

$$\mathbf{Y} = \{ 5, 3, 7 \};$$

Y is the *dependent* variable and I want to see if I can predict what month 4's revenue (and beyond) will look like. Y is also called the *outcome or dependent variable*, and X , the monthly variable, will be called the *independent* variable, the *predictor*. This is an extremely common occurrence - I am using time (month number) to predict revenue. Plotting the revenue on a graph with months as the horizontal axis and revenue on the vertical axis gives me a *time series* that I can approximate by a straight line.

Written as a vector, the monthly predictor variable looks like: (where the units are months).

$$\mathbf{x} = \{ 1, 2, 3 \};$$

My assumption here (subject to additional confirmation) is that *knowing X will help me to predict Y*. So I write a straight line *regression equation* ($\hat{Y} = \mathbf{a} + b \mathbf{x}$) and find the parameters a and b that make it the best match to the \mathbf{Y} data. Notice that I have set up a new variable \hat{Y} since I can't actually match \mathbf{Y} , I can only approximate it with the variable I'll call \hat{Y} . (often

referred to as "Yhat")

The straight line regression equation is then :

$$\hat{Y} = \mathbf{a} + b \mathbf{x};$$

a and b are parameters that will be calculated from the data so that the values predicted by \hat{Y} are as close as possible to the values of \mathbf{Y} . As I have set this up, "a" is the Y-intercept while "b" is the slope. Note: the bold letter a, represents the scalar 'a' multiplied by the special equi-angular vector $\mathbf{1} = \{ 1, 1, 1 \}$.

That's the trick here, use the historical given data to find a and b so that the line $a + b * \mathbf{X}$ is as close as possible to the \mathbf{Y} values. Since we only have historical data, that's what we will use to calculate the a and b parameters.

■ A simpler representation leads to simpler calculations

So far, I have given you raw data, as gathered from historical records or perhaps from surveys. A simplifying procedure is to subtract the *mean/average* from each variable vector yielding what is called a *centered variable/vector*. The resulting representative vector is called a *centered vector*. I have denoted the raw variables with capital letters and will use lower case letters to denote centered variables. Notice that the mean/average of \mathbf{Y} is 5 and for \mathbf{X} it is 2. Centering is carried out below (also see the accompanying diagram below):

Centering subtracts off what is common to a variable/vector, leaving the variation to be considered separately.

$$\mathbf{Y} = \{ 5, 3, 7 \}; \mathbf{y} = \{ 5 - 5, 3 - 5, 7 - 5 \}; \mathbf{y} = \{ 0, -2, 2 \};$$

$$\mathbf{x} = \{ 1, 2, 3 \}; \mathbf{x} = \{ -1, 0, 1 \};$$

The reason for doing this is to take a raw vector like \mathbf{X} or \mathbf{Y} and break it up into perpendicular pieces that can be worked on separately. For example, the raw vector $\mathbf{Y} = \{5, 3, 7\}$ can be written as the combination of a constant vector $\bar{\mathbf{Y}} = \{5, 5, 5\}$ plus a vector showing the deviations around that mean value, $\mathbf{y} = \{0, -2, 2\}$.

(Check out the diagram below and see these vector relationships).

So, $\mathbf{Y} = \bar{\mathbf{Y}} + \mathbf{y}$, Similarly, I have: $\mathbf{X} = \bar{\mathbf{X}} + \mathbf{x}$,

Of supreme importance is that the Mean vector $\bar{\mathbf{Y}}$ is perpendicular to \mathbf{y} and similarly,

the mean vector $\bar{\mathbf{X}}$ is orthogonal (perpendicular)

to the centered vector \mathbf{x} . Another way to express this important relationship is to say that

the *dot* product of $\bar{\mathbf{Y}}$ and \mathbf{y} is zero and that the *dot* product of $\bar{\mathbf{X}}$ and \mathbf{x} is zero as well. This implies that the angle between them is 90 degrees. The usefulness of this is that we can now write \mathbf{Y} and \mathbf{X} each in terms of two vectors that are perpendicular to each other. This allows us to analyze \mathbf{X} along two independent directions $\bar{\mathbf{X}}$ and \mathbf{x} as well as analyze \mathbf{Y} along its two independent directions, $\bar{\mathbf{Y}}$ and \mathbf{y} .

■ Technical digression (you can omit this right now, look back later)

Technically, the constant vectors, $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ lie in the same subspace spanned by the equiangular vector

$\mathbf{1} = \{1, 1, 1\}$. Let me call this vector subspace V_{constant} . The centered variables, \mathbf{x} and \mathbf{y} ,

lie in the complementary subspace which I will call $V_{\perp\text{constant}}$. Note the symbol \perp to indicate perpendicularity. Complementary

subspaces mean that all the vectors in one space are perpendicular to all the vectors in the other space. Similarly, $\hat{\mathbf{Y}}$ can be

broken up into a constant vector $(a + b * \bar{\mathbf{X}})$ that resides in the V_{constant} space, plus a vector

$b * \mathbf{x}$ that resides in $V_{\perp\text{constant}}$ space. The overall objective, looking at things this way,

is to have $(a + b * \bar{\mathbf{X}})$ match as close as possible to $\bar{\mathbf{Y}}$ in the constant subspace, V_{constant} ,

while having $b * \mathbf{x}$ match as close as possible to \mathbf{y} in the complementary subspace $V_{\perp\text{constant}}$. That's what

the regression equation solutions ensure. Because of the orthogonality of the subspaces,

these operations can be carried out independently.

■ How about a diagram?

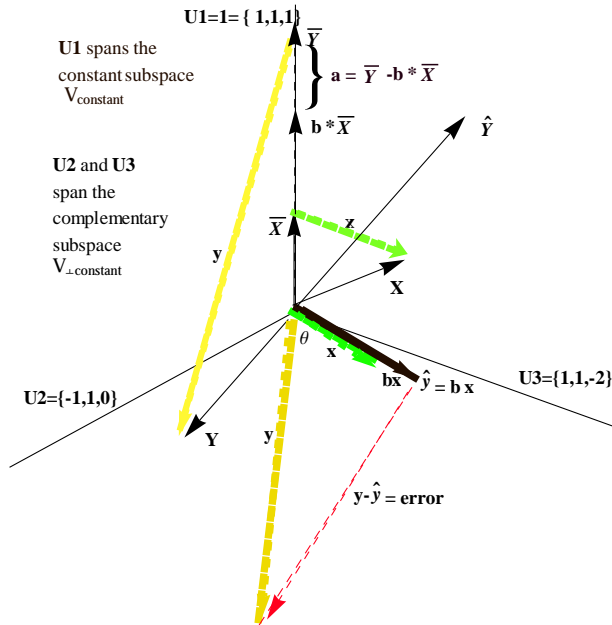
Ok, here is a visual breakdown of what it takes to do a regression of \mathbf{Y} on \mathbf{X} . First off, notice how I have broken up the raw vectors \mathbf{X} and \mathbf{Y} into their orthogonal components. That is, notice that (apart from my drawing skills)

$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{x}$, and $\mathbf{Y} = \bar{\mathbf{Y}} + \mathbf{y}$.

The constant \mathbf{a} looks like a simple scalar number but, as a vector it is considered to be $\{a, a, a\}$

or $a * \{1, 1, 1\} = a * \mathbf{1}$. The

'b' parameter however, is just a scalar although $b * \mathbf{x}$ is a vector.



Because the variables can be decomposed into two parts, we can work on each part separately.

First realize that the observed vector \mathbf{Y} can be broken down into a combination of $\bar{Y} + \mathbf{y}$. Similarly, $\mathbf{X} = \bar{X} + \mathbf{x}$

The plan is to fit the regression equation, $\hat{\mathbf{Y}} = \mathbf{a} + b * \mathbf{X}$ to this breakdown as close as possible. That is, $\hat{\mathbf{Y}}$ can be broken down into two pieces, $(\mathbf{a} + b \bar{X})$ and $b \mathbf{x}$ that will fit as close as possible to the breakdown of the \mathbf{Y} vector into \bar{Y} and \mathbf{y} .

$\hat{\mathbf{Y}} = \mathbf{a} + b \mathbf{X} = \mathbf{a} + b (\bar{X} + \mathbf{x}) = (\mathbf{a} + b \bar{X}) + b \mathbf{x}$ and so breaks into a constant part and a variable part.

Constant part: $(\mathbf{a} + b \bar{X})$

Variable part: $b \mathbf{x}$

Now the task is to fit the constant $\mathbf{a} + b \bar{X}$, part as close as possible to \bar{Y} , and the variable part $b * \mathbf{x}$, as close as possible to \mathbf{y} .

■ How to do the fitting work

Here is what we need to do, and in subsequent sections I will show how to do these tasks.

1.) Task 1 : what value of 'b' do I need to multiply \mathbf{x} by, so as to get as close as possible to \mathbf{y} . That is the problem of finding the regression coefficient "b". This task becomes, find "b" such that the distance from $b * \mathbf{x}$ to \mathbf{y} is minimized. That will become the perpendicularity condition.

2.) Task 2 : Notice that \mathbf{a} , \bar{X} and \bar{Y} are along the same line (1 – dimensional subspace), generated by the vector $\mathbf{1} = 1 = \{1, 1, 1\}$. This is the line that holds all vectors with constant components such as these mean vectors and multiples of them.

■ Task 1. Fit \mathbf{y} as close as possible to $\hat{\mathbf{y}} = b * \mathbf{x}$

From the diagram above, the closest we can get to \mathbf{y} is to get to its tip, at which point $b * \mathbf{x}$ is *perpendicular* to the error vector $(\mathbf{y} - b * \mathbf{x})$.

$$\mathbf{x} = \{-1, 0, 1\}; \mathbf{y} = \{0, -2, -2\};$$

Here is the perpendicularity condition, $\mathbf{x} \perp (\mathbf{y} - \hat{\mathbf{y}})$, $\mathbf{x} \perp (\mathbf{y} - \mathbf{b} * \mathbf{x})$

This condition, that puts $\hat{\mathbf{y}} = \mathbf{b} * \mathbf{x}$ as close as possible to the tip of \mathbf{y} , will allow us to write the equations that determine \mathbf{b} .

That equation is: $\mathbf{x} \cdot (\mathbf{y} - \mathbf{b} * \mathbf{x}) = 0$

$$\{-1, 0, 1\} \cdot (\{0, -2, 2\} - \mathbf{b} \{-1, 0, 1\}) = 0$$

$$2 - 2\mathbf{b} = 0$$

The solution is wonderfully simple, $\mathbf{b}=1$, that's the slope of the regression line.

$$\mathbf{b} = (\mathbf{x} \cdot \mathbf{y}) / (\mathbf{x} \cdot \mathbf{x})$$

■ **Task 2. Fit $\bar{\mathbf{Y}}$ as close as possible to $\mathbf{a} + \mathbf{b} * \bar{\mathbf{X}}$**

Since I found "b" from the first task, all I need now is the constant "a".

Since $\mathbf{b} = 1$, $\bar{\mathbf{Y}}=5$, and $\bar{\mathbf{X}} = 2$ the equations below lead to a solution of $\mathbf{a} = 3$.

$$\bar{\mathbf{Y}} = \mathbf{a} + \mathbf{b} \bar{\mathbf{X}}$$

$$\mathbf{a} = \bar{\mathbf{Y}} - \mathbf{b} \bar{\mathbf{X}}$$

■ **We're done!**

The final regression equation is :

$$\hat{\mathbf{Y}} = 3 + 1 * \mathbf{X}$$

So, for $\mathbf{X}=1$, $\hat{\mathbf{Y}}=4$, for $\mathbf{X}=4$, $\hat{\mathbf{Y}}=7$, and so on. These are the approximations to the actual historical values of \mathbf{Y} as well as predicted values for values beyond the historical record.

■ **How well did we do? (the correlation coefficient tells us)**

Ok, we now have the best (regression) line possible, given the data. But just how good is this *linear* approximation? Granted this is a toy example, but I can still show you how the interpretation works in general. It turns out that the most useful measure of the goodness of fit is a number called the (Pearson's Product Moment) *correlation coefficient*. Fortunately, that number has a very simple geometric interpretation, it is the *cosine of the angle* between the prediction vector $\mathbf{b} * \mathbf{x}$, and the observation vector \mathbf{y} . If you look back to our diagram, you will see the symbol θ , which is the angle between $\mathbf{b} * \mathbf{x}$ and \mathbf{y} . This is the same angle as between \mathbf{x} and \mathbf{y} . The cosine of that angle is the correlation coefficient. The symbol for that coefficient is r . So, $r = \text{Cos}[\theta]$. That's it!

When r is close to +1 or -1, then the prediction vector, $\mathbf{b} * \mathbf{x}$, is close to the observation vector \mathbf{y} and so provides a good approximation. As r tends toward zero, the prediction based on the values of \mathbf{x} become less helpful. When $r=0$, then \mathbf{X} is useless as a predictor of \mathbf{Y} .

■ **Calculating the correlation coefficient r , (which turns out to be 0.50, not so hot!)**

Since the *dot* product has the Cosine of the angle between two vectors as part of its definition, I will use that to find the Cosine between \mathbf{x} and \mathbf{y} and so determine the correlation coefficient R .

Since we have, by definition, for any two vectors with the same number of components:

$$\mathbf{x} \cdot \mathbf{y} = |\mathbf{x}| * |\mathbf{y}| \cos[\theta], \text{ now solve for } \cos[\theta]$$

$$\cos[\theta] = (\mathbf{x} \cdot \mathbf{y}) / (|\mathbf{x}| * |\mathbf{y}|) = r \text{ (the correlation coefficient)}$$

where $|\mathbf{x}|$ and $|\mathbf{y}|$ are the lengths of the respective vectors

$$\text{note that } |\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}} \text{ and } |\mathbf{y}| = \sqrt{\mathbf{y} \cdot \mathbf{y}}$$

In the example :

$$r = \cos[\theta] = \{-1, 0, 1\} \cdot \{0, -2, 2\} / (\text{Sqrt}[2] * \text{Sqrt}[8]) = 1/2$$

So, the Cosine of the angle is 1/2 which is the correlation coefficient, r.

Also, the angle between the vectors is: $\text{ArcCos}[1/2] = 60^\circ$. Notice that this will let me draw the relative positions of vectors with any number of components. This ability to visualize variable inter-relations in space becomes more and more valuable as the number of variables increase.

■ Going a little further to match textbook notations

Since \mathbf{y} , $b * \mathbf{x}$, and the error vector $(\mathbf{y} - \hat{\mathbf{y}})$ form a right triangle, Pythagoras theorem applies.

So, from the diagram, $\cos[\theta]$ is simply the adjacent side

$$b * |\mathbf{x}|, \text{ over the hypotenuse, } |\mathbf{y}|, \text{ so, we can also write } r = \cos[\theta] = b * |\mathbf{x}| / |\mathbf{y}|$$

Also, from right triangle rules, (see next section for an interpretation of each of these factors)

$$|\mathbf{y}|^2 = (b * |\mathbf{x}|)^2 + (\mathbf{y} - \hat{\mathbf{y}})^2$$

$$\mathbf{y} \cdot \mathbf{y} = b^2 * \mathbf{x} \cdot \mathbf{x} + (\mathbf{y} - \hat{\mathbf{y}}) \cdot (\mathbf{y} - \hat{\mathbf{y}})$$

■ The Text book approach

■ A digression showing variance and standard deviation as simply vector lengths and their squares

Note, the variance of a variable/vector is the square of its length, scaled by the number of its components minus 1. For sample statistics, the scale factor is n-1 where 'n' is the number of components of the vector. So in our notation,

$\text{Variance}[\mathbf{x}] = \mathbf{x} \cdot \mathbf{x} / (n-1)$, and

$$\text{StandardDeviation}[\mathbf{x}] = \sqrt{\text{Variance}[\mathbf{x}]}$$

$$\text{So, StandardDeviation}[\mathbf{x}] = \sqrt{\mathbf{x} \cdot \mathbf{x} / (n-1)} = \sqrt{\mathbf{x} \cdot \mathbf{x}} / \sqrt{n-1} = |\mathbf{x}| / \sqrt{n-1}$$

$$\text{and StandardDeviation}[\mathbf{y}] = \sqrt{\mathbf{y} \cdot \mathbf{y} / \sqrt{n-1}} = |\mathbf{y}| / \sqrt{n-1}$$

$$\text{Finally, StandardDeviation}[\mathbf{x}] / \text{StandardDeviation}[\mathbf{y}] = |\mathbf{x}| / |\mathbf{y}|$$

■ Textbook notation equivalences

- $R = \cos[\theta] = b * |x| / |y| = b * \text{StandardDeviation}[x] / \text{StandardDeviation}[y]$
- $R = \cos[\theta] = \text{correlation between } x \text{ and } y$
- $R^2 = \text{coefficient of determination and is the proportion of total sum of squares explained by sum of squares of regression.}$
- $SS_{\text{total}} = |y|^2 = y \cdot y$ (Total Sum of Squares also known as the total variance)
- $SS_{\text{regression}} = \hat{y}^2 = (b * |x|)^2 = b^2 * |x|^2 = b^2 x \cdot x$ (Regression Sum of Squares)
- $SS_{\text{regression}} = R^2 * |y|^2 = R^2 * SS_{\text{total}}$ (showing R^2 as representing the proportion of total variance of y , explained by using knowledge of 'x')
- $SS_{\text{error}} = SS_{\text{total}} - SS_{\text{regression}} = SS_{\text{total}} (1 - R^2)$

■ Textbook algebra equivalences

This section will show how the textbook algebra relates to the underlying geometry. The text book equations for least squares are as follows:

$$b = \left(\sum_{i=1}^N (x_i - \bar{x}) * (y_i - \bar{y}) \right) / \sum_{i=1}^N (x_i - \bar{x})^2$$

$$a = \bar{y} - b * \bar{x}$$

■ The Algebra Equivalences of the Textbooks to our approach

It is very easy to relate the textbook algebra with our geometric approach. Simply notice that $(x_i - \bar{x})$ is just one component of the centered vector x .

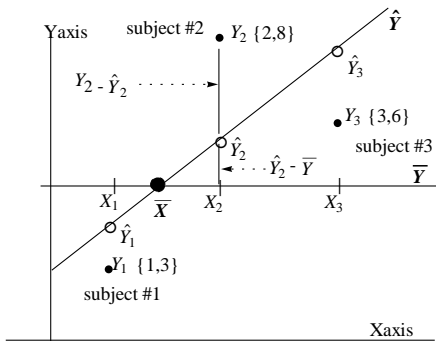
$\sum_{i=1}^N (x_i - \bar{x}) * (y_i - \bar{y})$ is the long way to write the dot product of the centered variables x and y . That is,

the expression above is simply $x \cdot y$

Similarly, $\sum_{i=1}^N (x_i - \bar{x})^2$ is the long way to write the dot product of x with x , $x \cdot x$.

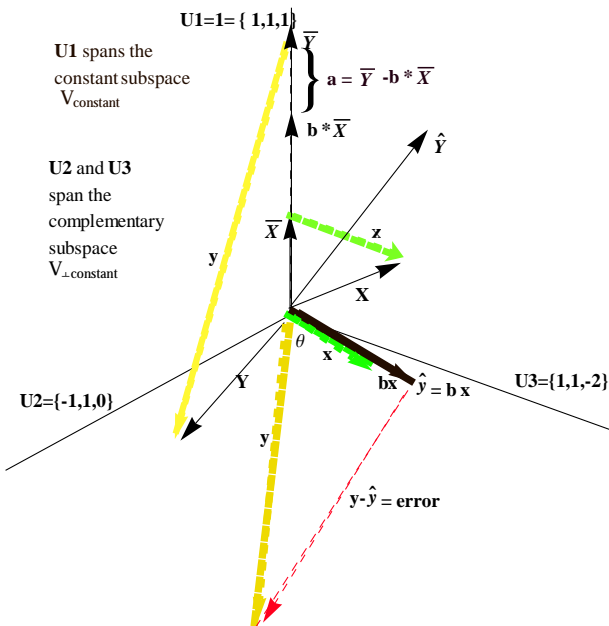
■ **Showing the Scatterplot (variable space)**

The usual way to show a 2 - dimensional regression problem is to show a scatterplot of the Y values along the Y axis and the X values along the X axis. The idea is to draw the best line through the Y values. That line is called the regression line, which I have denoted as $\hat{Y} = a + b * X$ in the above discussions and is labeled below. Notice that the regression line, \hat{Y} , must pass through the point $\{\bar{X}, \bar{Y}\}$. (Remember the derivation above where we came up with a derived, required relationship of $\bar{Y} = a + b * \bar{X}$? Well, that requirement means that the point \bar{X}, \bar{Y} is a point that satisfies the regression equation and so the regression equation must pass through it. To repeat an earlier diagram of the a 3-subject scatter plot:



Scatter plot (variable space) drawing of three subjects and their X,Y values. This view helps to check the observation patterns.

■ **Showing the equivalent vectors from the scatter plot**



Summary

I have tried to show how the geometry of variables, interpreted as vectors, can lead to considerable insight into statistical calculations. These same geometrical ideas work in higher dimensions as well, although it takes a bit of imagination to apply them. Of course, a computer package is essential when more than a few values are to be calculated but is really helps to see what is happening .

■ References

Wickens, Thomas, (2005) *The Geometry of Multivariate Statistics* , Erlbaum.

Saville D. and Wood, (1991) *The Geometric Approach to Statistics* , Prentice Hall.

[BasicStatsBivariateRegression.pdf will be the saved file name, 2009-08=20]

■ ** **N/A -- Extra Stuff & Verbiage (couldn't bear to toss!!) ****

I will start out with the recommended approach of first showing a scatter plot and then use published algebraic equations to solve for the regression coefficients, 'a' and 'b'. All this is standard, and most text books start out like this. I will use an example to show how this scatter plot geometry helps you to avoid the serious error of trying to fit a linear (straight line) equation to data that could very well be curved. Beyond looking at the scatter plot to check for non - linearity, the next step is to plug the X and Y data into equations and solve for a and b. This is o.k. as far as it goes, but I would like to take you a bit farther and show you the underlying geometry that will give you insight into this problem as well as multidimensional one.

■ Extra Equations, Just for grins 2009-08-10!

Was playing around with the stat triangle and was deriving the relating eqns. Wanted to show how the sum of squares worked around the triangle.

$$y.y = bx . bx + (y - bx).(y - bx) = bx .bx + e.e$$

$$y.y = b^2 x.x + y.y + b^2 x.x - 2 b x.y$$

$$y.y = 2 b^2 x.x + 2 b x.y + y.y$$

$$b x.x + x.y ==0$$

$$b = x.y/x.x$$

so initially couldn't equate these but it works when I realize that $b = x.y/x.x$