Analogies: Techniques for Adjusting Them

SCEA National Conference
Manhattan Beach, June 2004

The Society of Cost Estimating and Analysis

Richard L. Coleman, TASC
Jessica R. Summerville, TASC
Shishu S. Gupta, IC CAIG

NORTHROP GRUMMAN
Information Technology
TASC

Table of Contents

Introduction 3
Definitions 3
The Current Method 4
Three schools of thought about the y-intercept 5
School 1: Y-Intercepts must pass through the origin” 5
School 2: “Y-Intercepts must make sense” 7
School 3: “The Y-Intercepts is just what it is” 7
Proposal - Two New Methods 7
The Borrowed Slope Method 8
The Relational Correlation Method 10
Bivariate Normality 10
The Geometry of Rregression 14
The Implications of the Geometry of Regression 17
Implementing Relational Correlation for Analogies (and buildups) 18
Conclusions 19

Analogies: Techniques for Adjusting Them

Introduction

The three basic methods of cost estimation are generally accepted to be Parametric, Analogy and Build-up. Considerable attention is devoted to research in parametric methods, generally in higher-level and oversight organizations, and for early studies. Likewise, great expertise is resident in manufacturing sector in Build-up methods. The technique of analogy receives surprisingly little attention in research or in development of methods. Practitioners are enjoined to "find the best match and adjust it if necessary." The issue of what exactly *constitutes* a best match and *how* to adjust are left to the practitioner. This paper will discuss the usual method, linear adjustment based on a single parameter such as weight or size (often called "J-ing up"), and will recommend two other methods as preferable to straight linear adjustment: one which borrows a slope from a cost estimating relationship, when one exists, and the other that develops a factor from the geometry of correlation, and which arises from a previous paper by the authors. The reasons for these methods being preferable will be presented, and an illustrative case study with actual data will be shown. The problem of finding a best match is not trivial and is not solved in this paper, but it will be discussed briefly for completeness.

Definitions

The below definitions are not put forward to pick a fight with those who define terms more broadly, or differently, but simply to give definitions to ensure clarity of terms. The definitions are not arbitrary, and reflect the usage of the authors, but should the reader use terms differently, the reader is respectfully invited to take the point being made with the definitions in mind, and not let disagreement with the definitions become obstacles.

<u>Cost Estimating Relationship</u>: A relationship between cost and another parameter. CERs involving parameters but not based on statistical analysis are, in this paper, considered either rates or factors.

<u>Parametric Estimates</u>: Estimates made by developing statistical "Cost Estimating Relationships" (CERs) based on one or more parameter and cost. Parametric estimates involving parameters but not based on statistical analysis are, in this paper, called either "adjusted analogies" or "adjusted buildups."

<u>Analogies</u>: Estimation by assuming that the costs of a new system will be equal to (or similar to) the costs of a system that is similar in form. "Adjustments" (defined below) are almost always made to analogies.

<u>Buildups</u>: Physical Bills of Materials (BOMs) and CAD-generated material lists and the like. In this definition, we do not mean "buildups" consisting entirely of Staffing levels multiplied times durations and costed, such estimating techniques are little more than "engineering judgment" in fine detail. Buildups often include "adjustments" to allow for

size differences in their details. Buildups almost always include rates (the product of a cost per unit times a unit) or factors (factors are usually expressed and used as percentatge of per-unit, and are usually cost on cost.)

Composite methods: A method that involves at least two of the three other types is a composite. Most estimates are really composites, although it is the practice to refer to them by the name of their most predominant component, thus estimates with most of the costs derived by parametric will be called parametric.

Adjustments: Scaling of a cost by some physical, performance, or other such attribute is called adjustment in this paper. Scaling parameters are usually countable or measurable and intuitively tied to cost. Scaling is usually (in practice) directly proportional to the attribute, but it need not be.

The Current Method

Adjustments, in the analogy or buildup method, typically rely on an "obvious" characteristic. The characteristic used for adjustment is most often weight; sometimes weight of the new system is not known, and so another characteristic is used (often as a proxy for weight. For example, a characteristic such as bore diameter of a gun might be used. Usually the ratio of the values of the characteristic in the new system to the value in the old system is multiplied by the cost of the old system. This method is sometimes called "j-ing up the estimate." Sometimes the characteristic is transformed in a way that is thought to make it proportional to weight, e.g., the bore diameter of a gun, cubed, might be used to scale up costs.

To see the implications of the current method, we will examine an example adjustment by ratio. Suppose:

> The analogy weighs 300 tons and costs $100M
> The new system weighs 500 tons
> The new system is assumed to cost $166.67M (500/300*$100M = $166.67M.)

The above is a typical and familiar adjustment. What is its implication? Should we be inclined to believe it? Is it in accord with what we believe? Let's look at a graph to see what it implies. There is a surprise there for most of us. Before we look at the implications, force yourself to predict what the line between the analogy and the prediction looks like, and where it crosses the y-axis. To prevent inadvertent glimpsing, the answer is on the following page.

The below graph shows the adjustment discussed previously. The analogy weighs 300 tons and costs $100M. The new system weighs 500 tons and is assumed to cost $166.67M (500/300*$100M = $166.67M.) Note that the line through the 2 points passes through the origin.

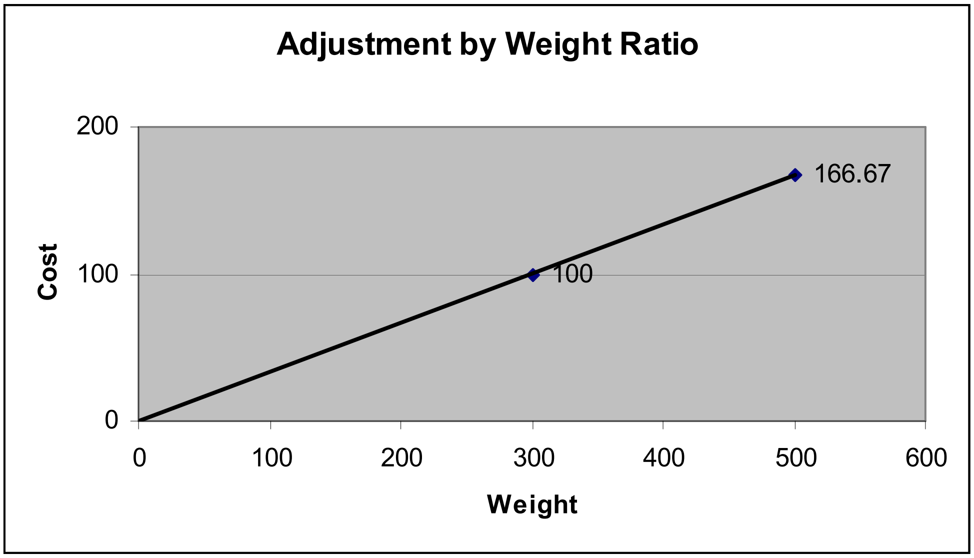

Figure 1: Adjustment by Weight Ratio

Straight adjustments by ratios always pass through the origin! Most observers fail to predict that, even though it is straightforward to show that it must. Is this reasonable?

Three schools of thought about the y-intercept

Before we decide if the fact that adjustment by ratio amounts to a line passing through the origin makes sens, we need to discuss what the three schools of thought are for the y-intercept of a CER. The y-intercept is a litmus test among cost estimators. There are about three schools of thought:

1. CERs "should" pass through the origin
2. CERs that do not pass through the origin must have an explicable y-intercept
3. CERs must be statistically derived, and if done properly, the y-intercept is just "what it is"

We'll discuss each briefly and then assume you are of school 2 or 3. The reader should be warned that almost anyone is from one of these schools of thought at heart. <u>The authors are no exception.</u> Further, the reader should know that the gulf between these schools is wide.

School 1: Y-Intercepts must pass through the origin"

The typical argument posed by estimators form school of thought 1 is:

- "If I spend no money, I get no product." The pros and cons of this school of thought are:

Pro:

- Sounds good

Con:

- The argument doesn't seem to match the data. E. g., the price of Flash Drives, as shown in the below graphs.

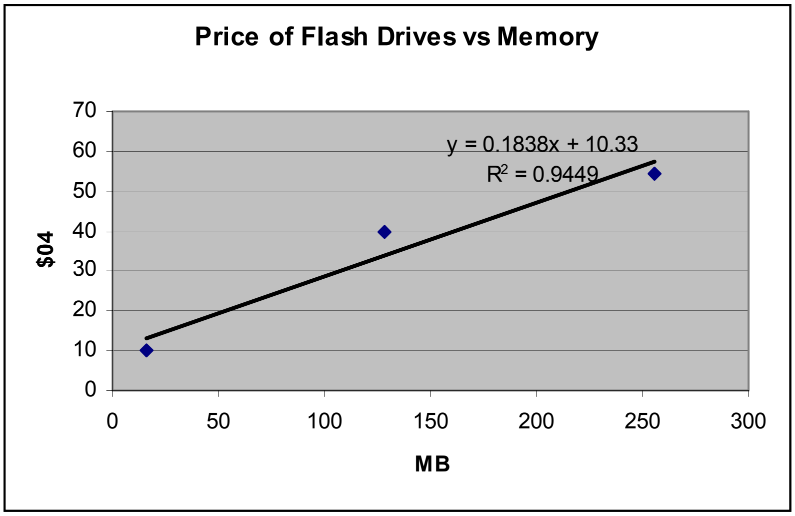

Figure 2: A simple linear fit to 3 actual Flash Drives costs

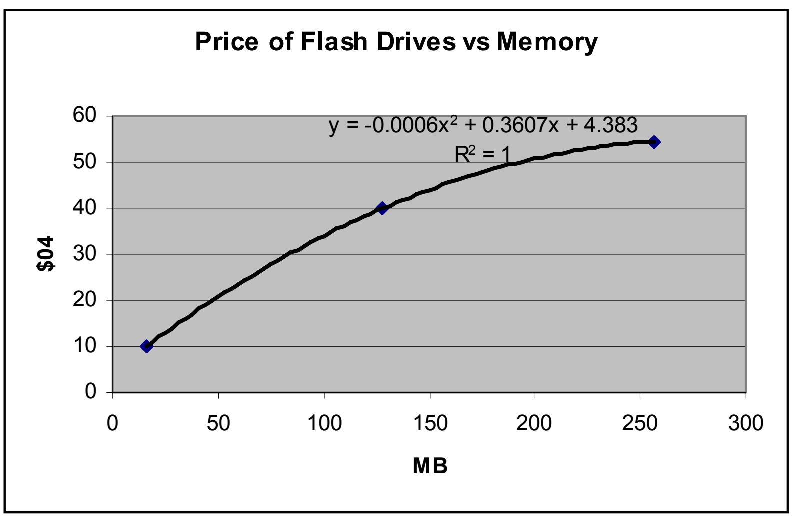

Figure 3: A 2nd-order polynomial fit to 3 actual Flash Drives costs

In the two curve fits, it is noteworthy that although the lowest priced flash drive is quite close to zero capacity, particularly relative to the other two Flash Drives, neither line passes through the origin. It is safe to assume that, in the current technology no matter how small the memory of a Flash Drive, as long as it is non-zero, its price will never be "free" because it will never be free to produce it. CERs are to predict costs of a thing, not the cost of nothing. Before the reader objects to proceeding outside the range of data, the authors freely agree with the desire to keep within the range of data, and the need not to extrapolate needlessly, but this does not prevent the observation that the y-intercept of the best fit line and the best fit polynomial is not zero. Secondarily, before the objection is made that a second degree polynomial is inappropriate to fit to 3 points, the authors are not teaching line fitting, just making a simple point with simple, but real data. Thirdly, if the reader still objects that the y-intercept must be zero, it does not make the authors wrong, but simply places the reader in question into school one, and the authors in another school of thought!

School 2: "Y-Intercepts must make sense"

The typical argument put forward by school 2 is:

- "There must be physics-based arguments for CERs"

Pro:

- It is helpful to think about CERs in this way, within reason.

Cons:

- If practiced to the extreme, good CERs can be rejected just because we do not yet understand them.
- Engineers, who hate cost estimation, can usually talk the analyst to a full stop!

School 3: "The Y-Intercepts is just what it is"

The typical argument put forward by school 3 is:

- We are not trying to predict the y-intercept. We are trying to predict the cost of systems of non-zero size.
    - We should take the best advice the data can give us
    - We should extrapolate as little as we can.
- If the data show that the y-intercept is non-zero, we should not reject a CER just because we do not know why
    - Galileo believed the data he saw, even absent a theory of gravity. It took centuries before Isaac Newton knew why – but Isaac Newton wouldn't even have wondered without Galileo showing that there was an explanation missing.
- This approach is what the practice of statistics currently recommends

Pro:

- Any existing system (i. e., one of the data points underlying the CER) is well predicted.

Con:

- If the analysis is not well done, there may be a better CER.

Proposal - Two New Methods

Two new methods of adjusting a CER are proposed: 1) the borrowed slope method, a variant of the methods for calibrating[1] CERs, which involves adjusting a "trusted

[1] *A Framework for Costing in a C AIV Environment*, R. L. Coleman, TASC; D. Mannarelli, Navy ARO, ASNE 1996, ADoDCAS 1996

analogy" by a "trusted slope," and 2) "relational correlation," taking advantage of the geometry of regression[2], adjusting a "trusted analogy" by a "best guess slope."

The Borrowed Slope Method

The borrowed slope method is based on "calibrating a CER." A CER is adjusted to "more trusted," industry, or company specific data by moving the slope to pass through a point or set of points. This is illustrated in the following figure.

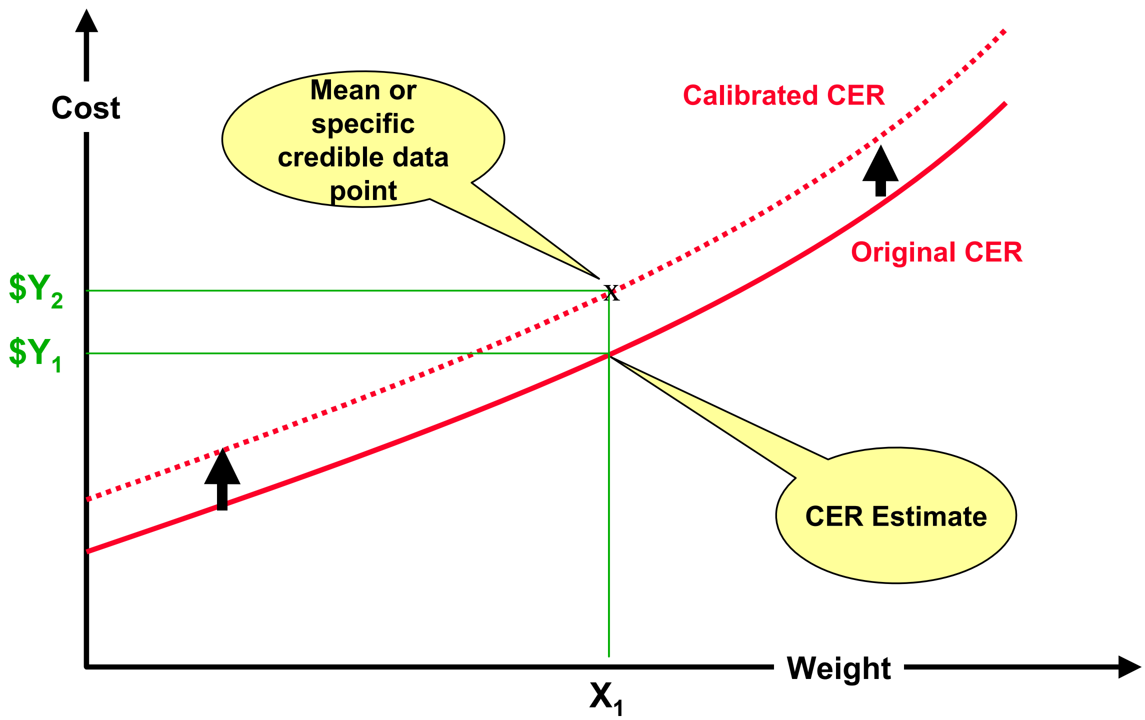

Figure 4: Calibrating a CER

To adjust an analogy, do precisely the same thing, but instead of believing you are adjusting a CER to specific data, think of it as departing from "the most credible point" via "the most credible slope."

With calibration in mind, consider the next figure.

[2] *Relational Correlation, What to do when Functional Correlation is Impossible,* ISPA/SCEA 2001, R.L. Coleman, J.R. Summerville, M.E. Dameron, C.L. Pullen; TASC, Inc., S.S.Gupta, IC CAIG

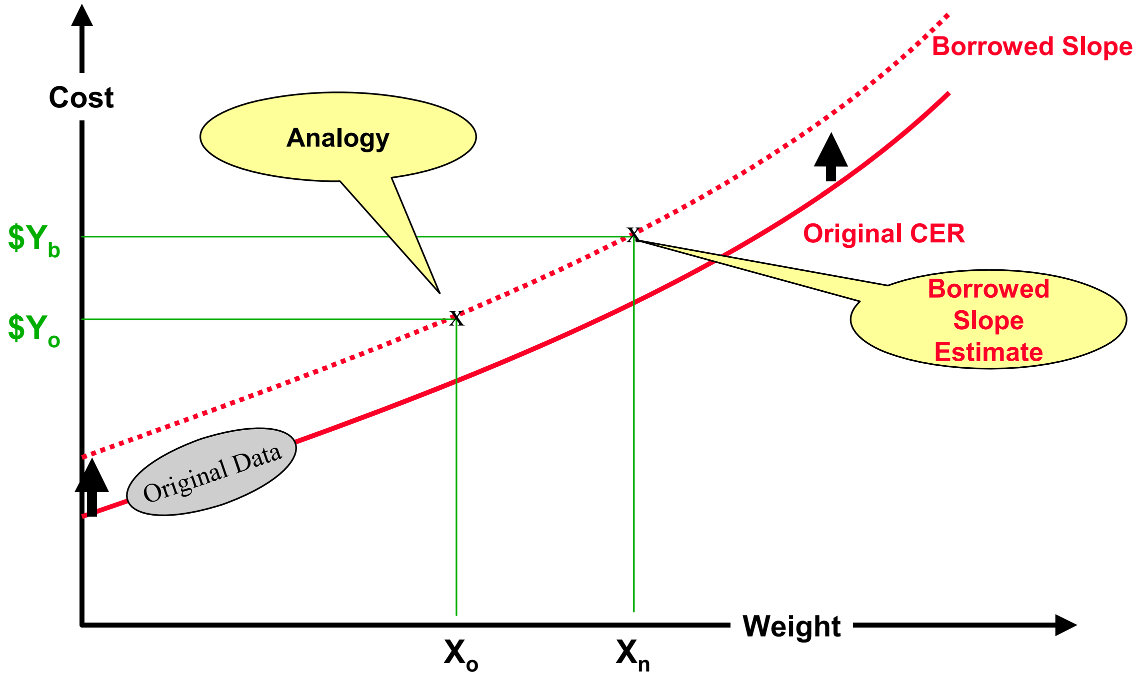

Figure 5: The borrowed slope method

In the above figure, an object with weight of $weight_o$ and cost of $\$Y_o$ is the analogy. This object lies off the best-known CER for reasons that are sensible, in accord with the direction of the offset, and for reasons that are shared by the system being estimated. For example, suppose the CER is based upon industry-wide data, but the analogy system was made by a factory that has known higher costs, and that this factory will make the system being estimated (the reader is requested to accept the example as reasonable, and for purposes of the illustration). Given that the estimator accepts these beliefs, the estimator would revise the CER so as to make it pass through the analogy point, retaining the slope of the CER.

Adjusting by borrowed slope is compared to adjusting by ratio in the following figure.

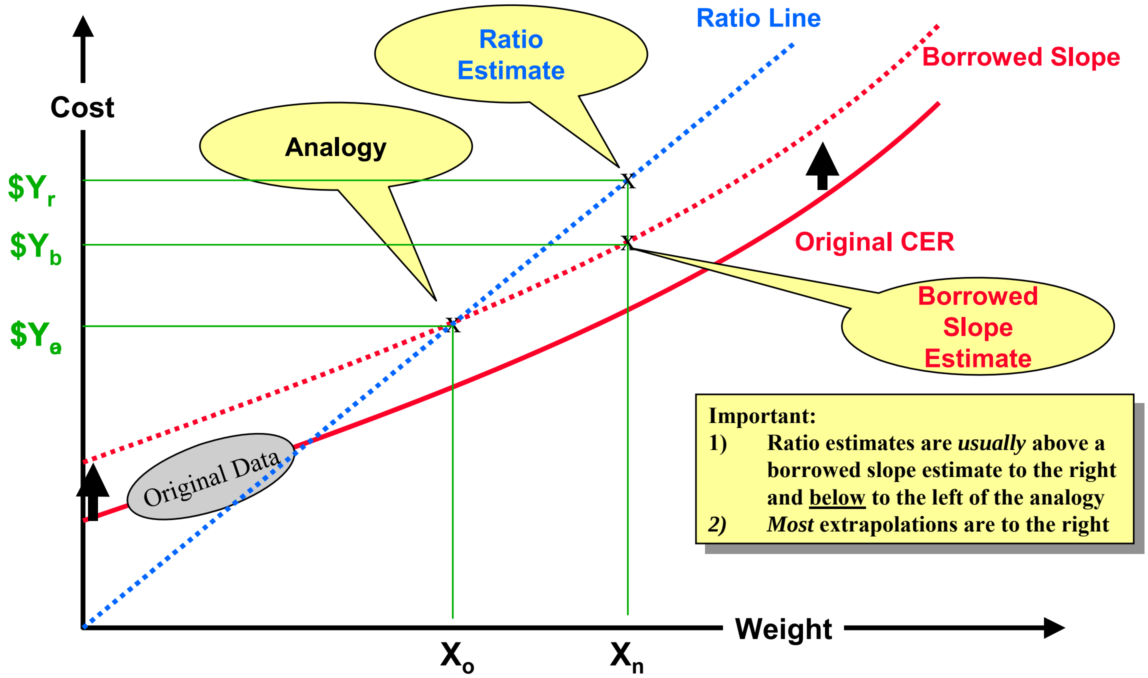

Figure 6: The borrowed slope method compared to the ratio method

As can be seen in the above graphic, there can be considerable difference between a borrowed slope adjustment and a atio adjustment. In general we develop bigger, faster and the like, so we tend to over estimate with the ratio method. We should not, by the way

The Relational Correlation Method

A much more esoteric method is available, which borrows from bivariate normality and the geometry of regression. This method is available when there is no "trusted slope" to borrow.

Bivariate Normality

Let us first consider the case of bivariate normality.

Suppose X and Y are distributed N(μx, σx) and N(μy, σy)

Suppose X and Y are jointly bivariate normal with correlation ρ

Then the graph of X and Y will appear as follows:

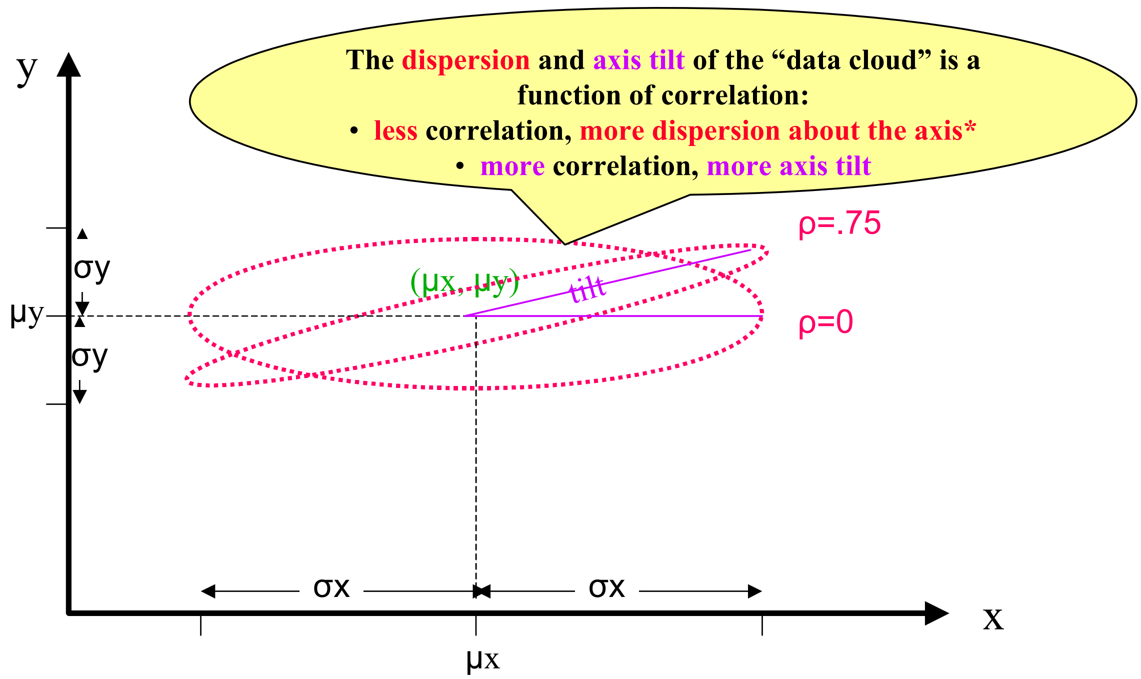

Figure 6: The borrowed slope method compared to the ratio method

We note that the "data cloud" will be shaped something like one of the two red dotted ovals, with 68.3% of the mass of the joint probability distribution inside the ovals which mark the 1-sigma curve, centered at the means of the two variables. The degree of correlation will affect the tilt of the oval. As noted on the illustration, the "fatness" of the data cloud is also connected to correlation.

For further background, we will now consider "the geometry of regression." The below facts are known to mathematicians, but obscure, and not remembered in cost analysis:

For any two jointly distributed variables, there is a regression line

The slope is:

$$m = \rho*(\sigma y/\sigma x)$$

The y intercept is:

$$b = \mu y - \rho(\sigma y / \sigma x) * \mu x$$

If the variables are joint bivariate normal, then ρ is the correlation coefficient. This is best seen by a series of graphics, which follow:

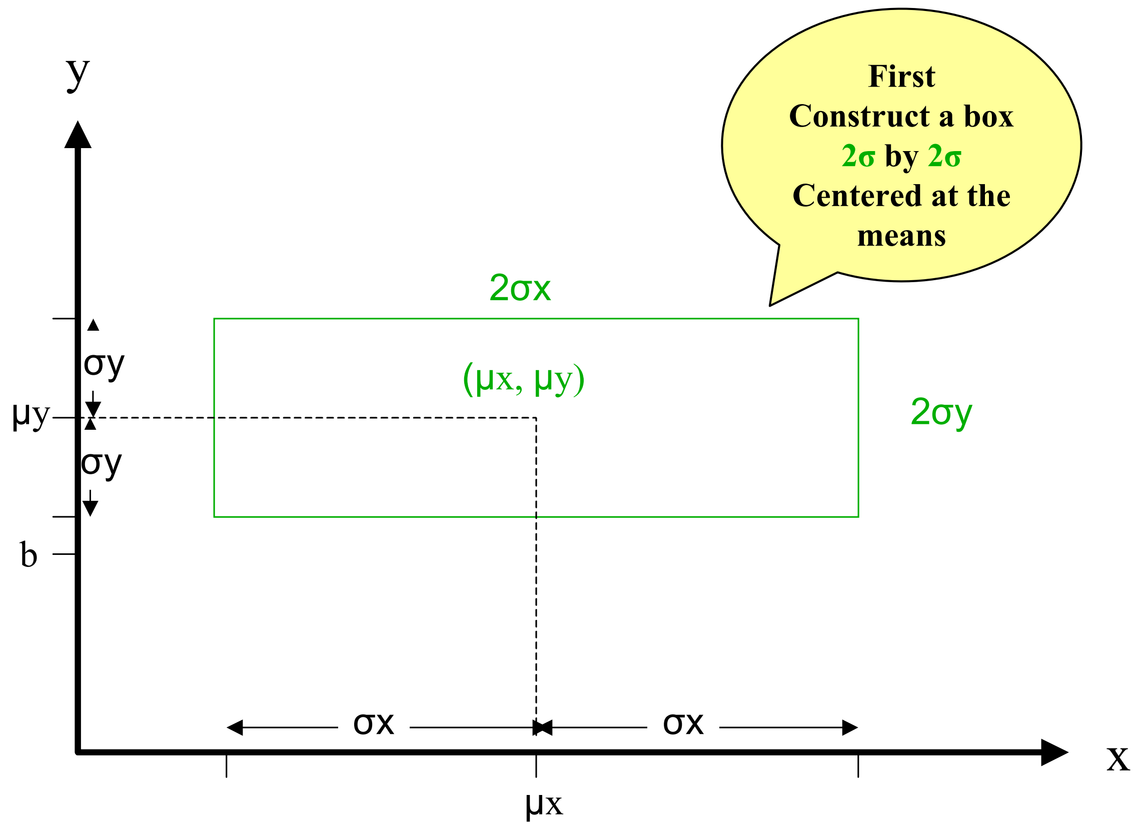

Figure 7: Bivariate normality – constructing the box

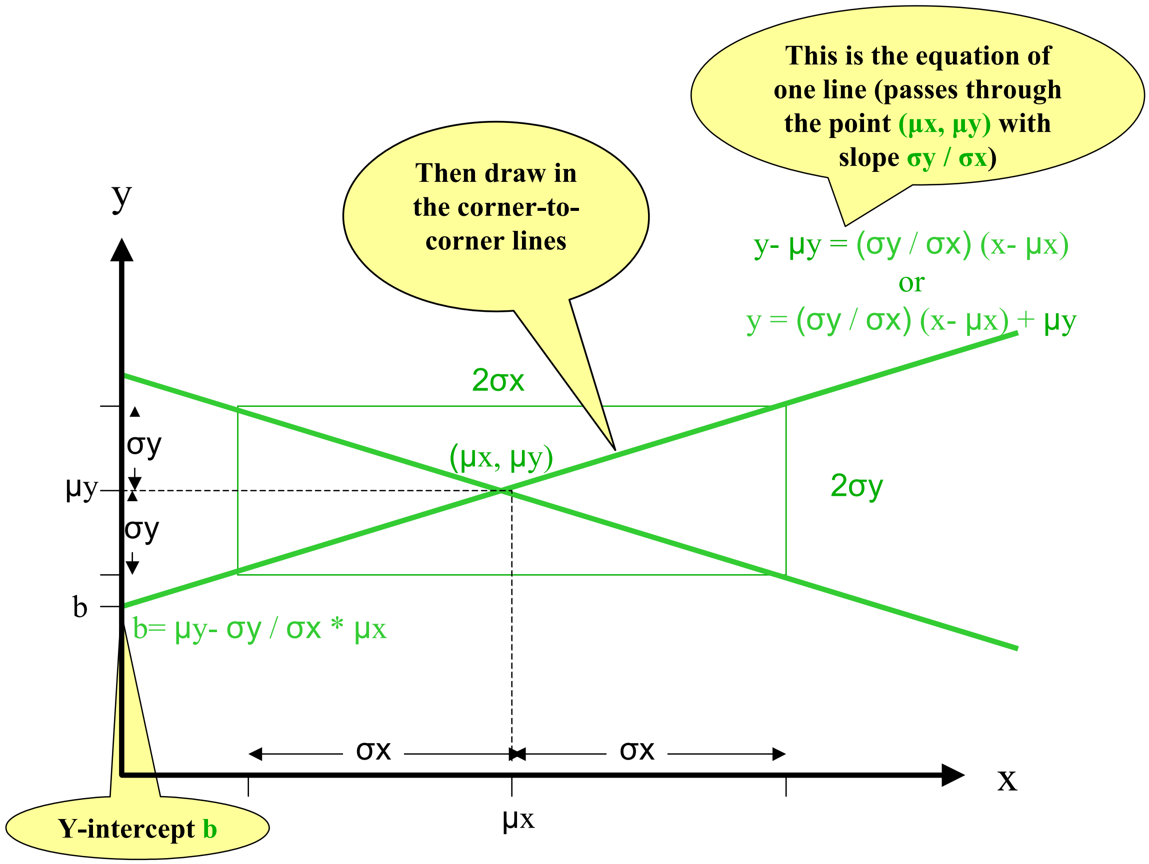

Figure 8: Bivariate normality – inserting the diagonals

Now, populate the box with data, shown here as the already illustrated 'data clouds."

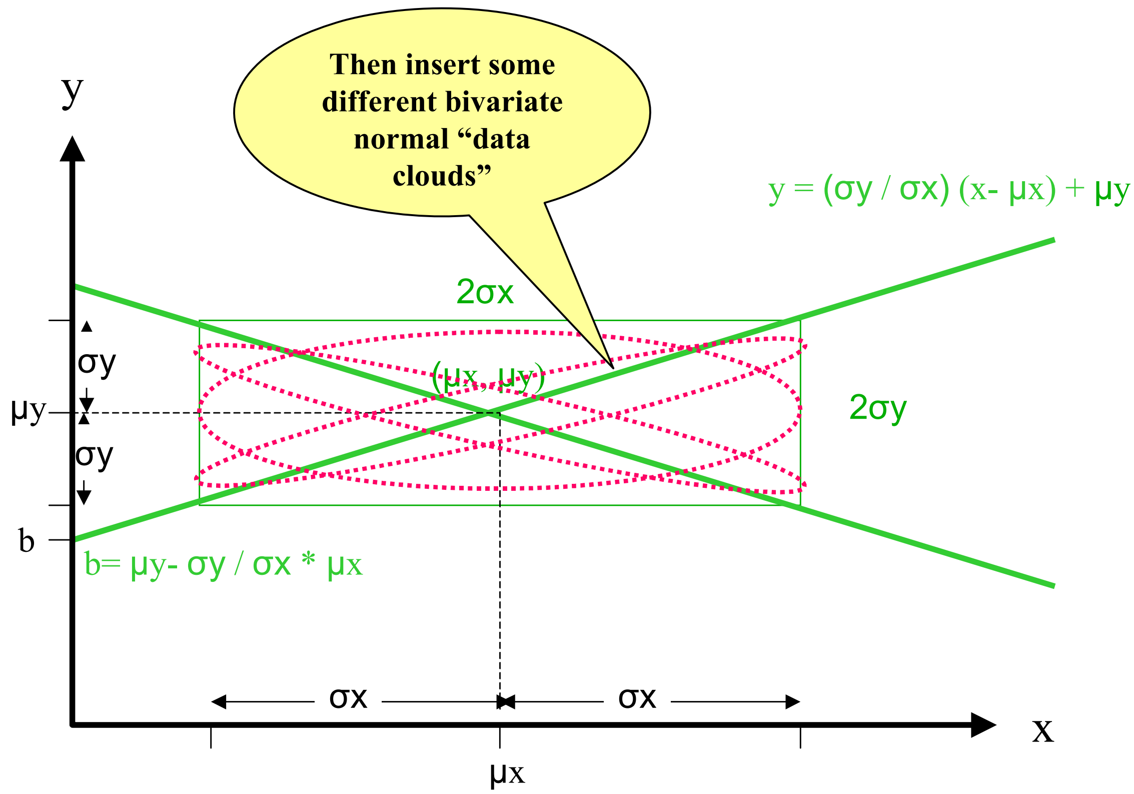

Figure 9: Bivariate normality – inserting the data

Now, let us look at the meaning of what we have constructed. And consider the geometry of regression.

The Geometry of Rregression

We will look at the picture we have constructed and see what the geometry of regression tells us. We should note that two variables need not be jointly bivariate for the regression line to exist – the only addition to our 'picture' is that the slope of the regression line is affected by a parameter called ρ, and if the variables are jointly-distributed bivariate normals, this parameter ρ is their correlation.

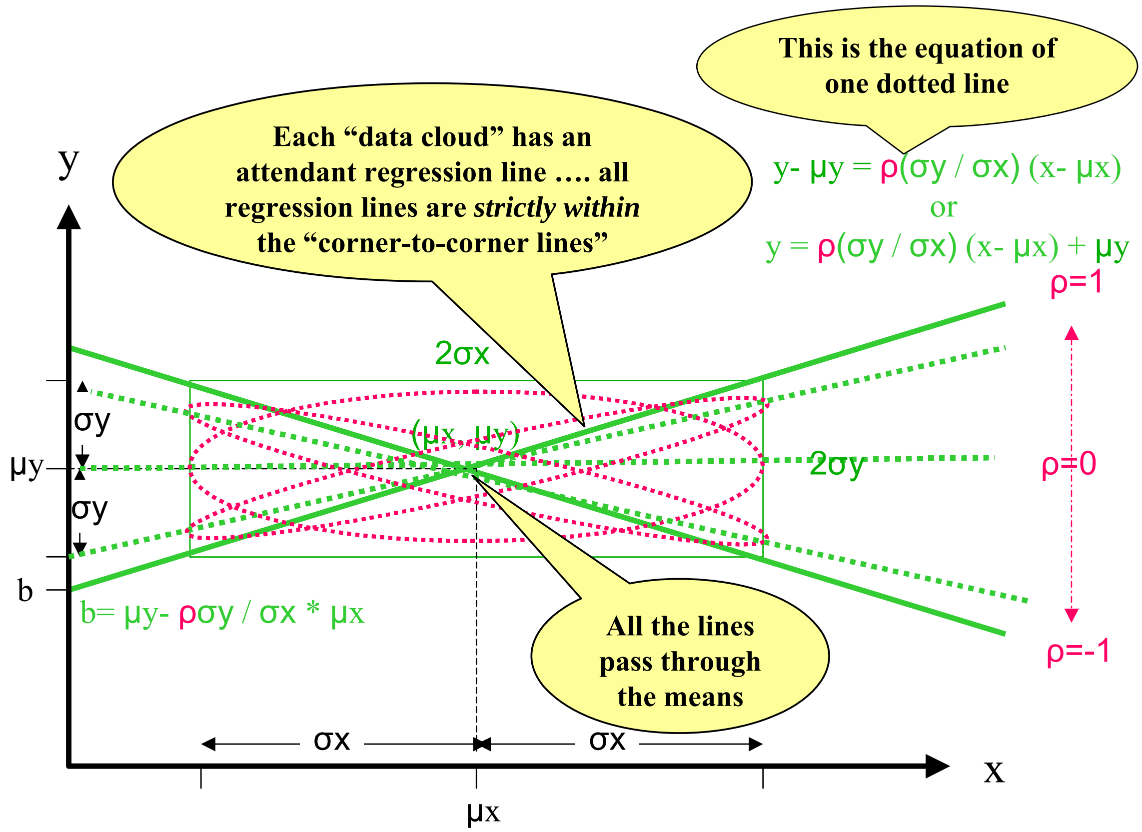

Figure 10: The geometry of regression - the implications

Now let us look at how the parameters ρ, σy and σx affect the regression line.

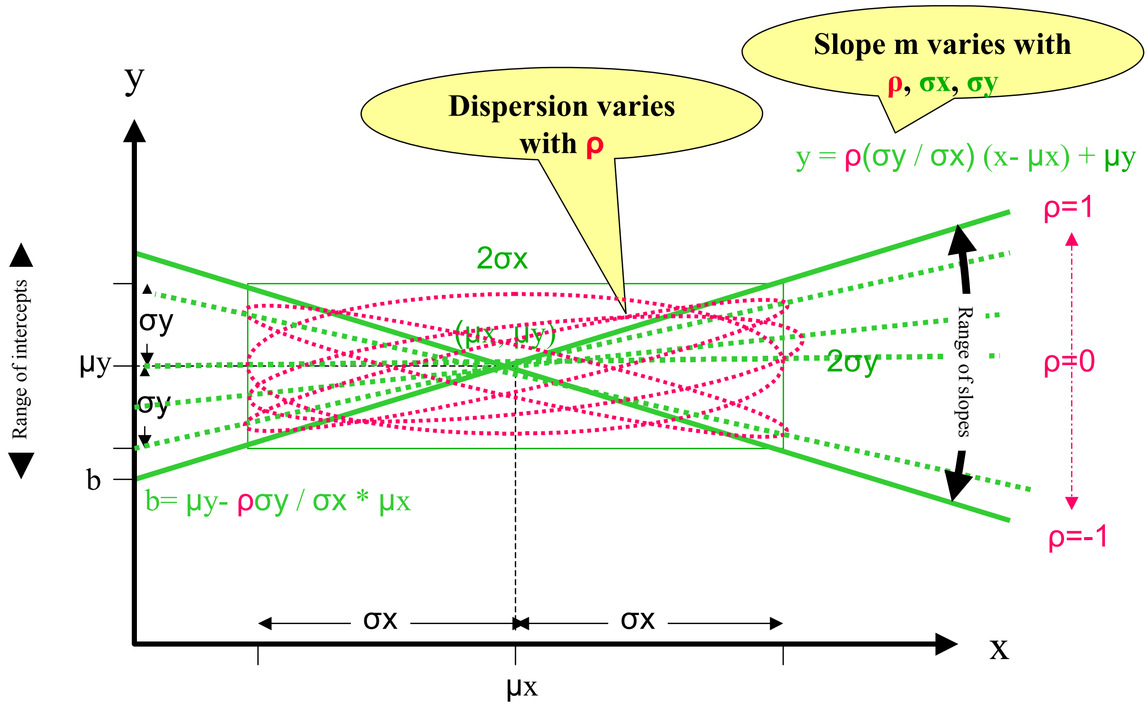

Figure 11: The geometry of regression - the effect of the parameters ρ, σy and σx

Now we will depart briefly from the case we are building and just look at the meaning of r^2. We do this simply because we are already well-versed in the meaning of the geometry of regression, and we can see this important parameter with little additional work. We do not need it for our development, but it's a good thing to know and we may as well know it!

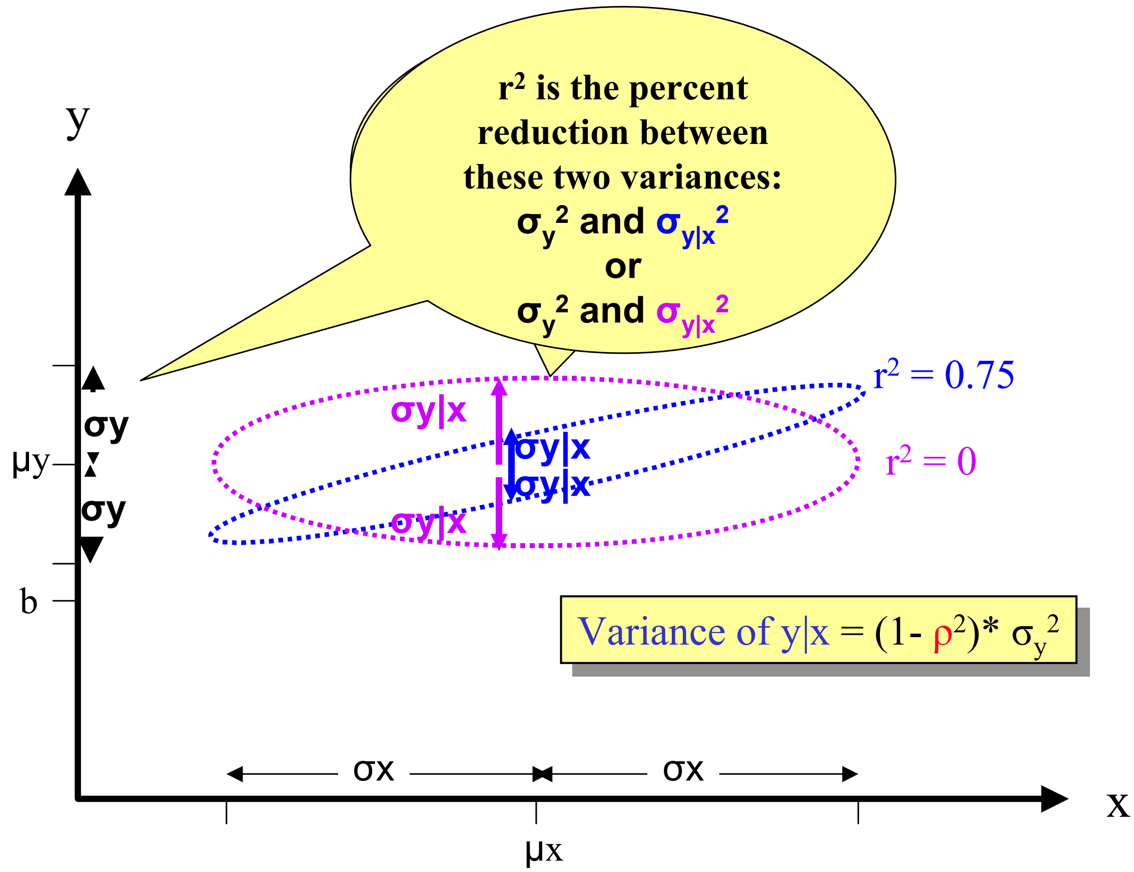

Figure 12: The meaning of r^2

The Implications of the Geometry of Regression

For every regression with apparent slope m, there is an unseen equation with steeper slope m/ρ which is the unseen slope of the two variables, and with an unseen accompanying y intercept. Once we decide upon the means and the variances of x and y, the unseen line is fixed. Once we pick ρ, the regression line is fixed.

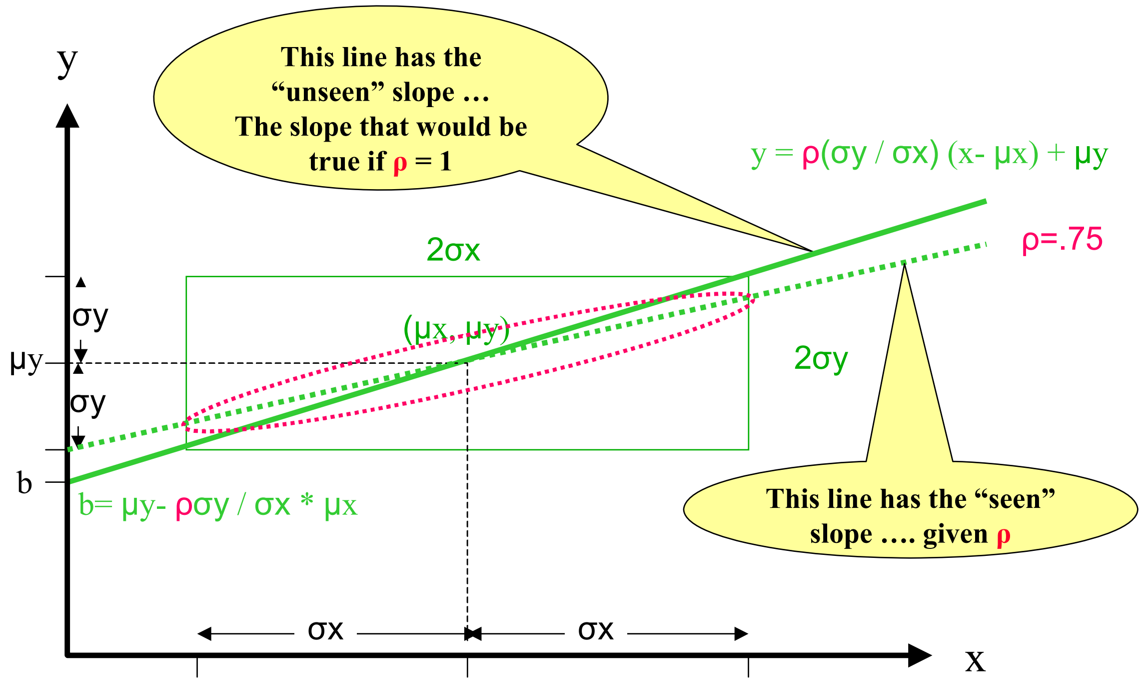

Figure 13: The implications of the geometry of regression

Implementing Relational Correlation for Analogies (and buildups)

For Single Point Analogies
1) Determine a reasonable (preferably historically-based) standard deviation for the x and y variable, e.g., to estimate ship repair parts as a function of tonnage you'll need:
a. The standard deviation for the analogy ship class repair parts cost
b. The standard deviation for the tonnage within the ship class
c. The standard deviation of repair parts for a single ship of the class
2) The ratio of 1 and 2 gives you the unseen slope
3) The relationship of 3 and 1 will yield r2 (Variance of y|x = (1- ρ2) * σy2)

For buildups
For buildups, do as above, but use an analogy for the values of the standard deviations, and apply it to your buildup using percents

We will now look at the next figure to see what this looks like.

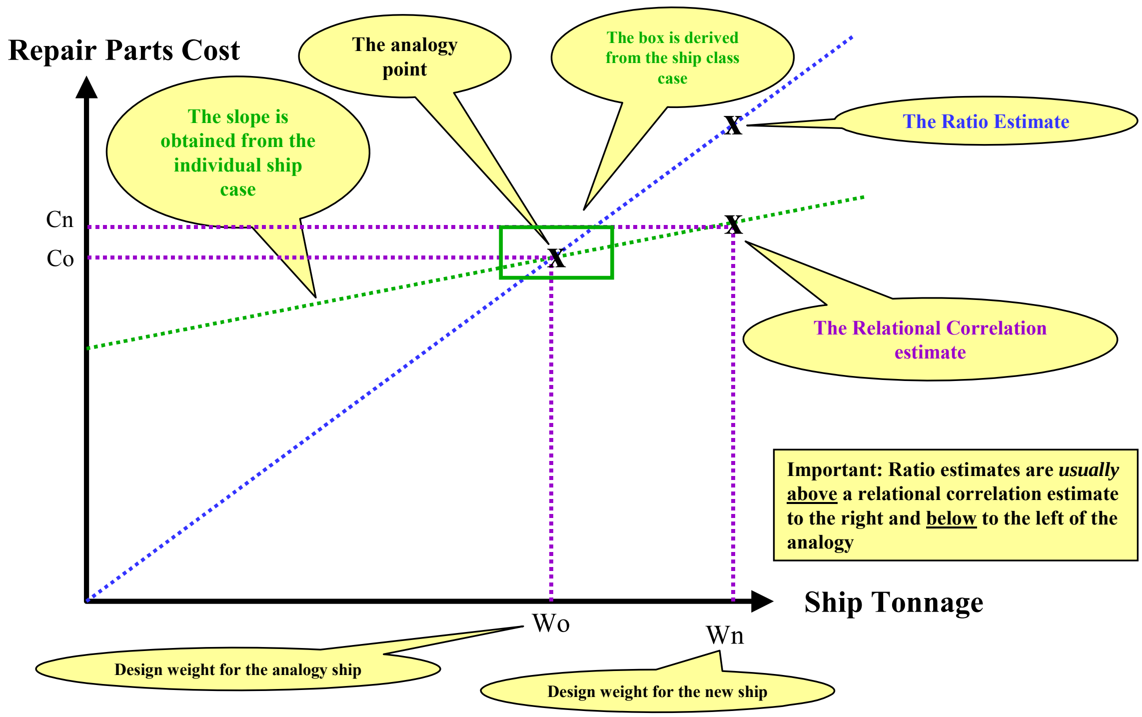

Figure 14: The relational correlation method pictorially

Conclusions

Adjustments of analogies have received too little attention. There is a current method, and two new methods we have proposed.

The Ratio method
Ratio adjustments are the current practice. As we have seen, this method will typically (but not always) overstate above the analogy, and understate below.

The Borrowed slope method
The borrowed slope method needs only a CER to implement it. It is not esoteric and needs little more that the knowledge and beliefs we already possess, and arguably constitutes an improvement in accuracy.

Relational correlation
The relational correlation method is admittedly esoteric, and although based upon earlier work, the authors do not expect it to be used often. *We would most assuredly never use it, and no one should, if the borrowed slope method is available.* Its value is in cases when there is simply no CER at hand. We suppose that it is most likely to be adopted when there is little vulnerability to criticism, and trust predominates. We say this because the method requires considerable judgment and is thus somewhat vulnerable to criticism, even if well done, not to mention hared to understand. That notwithstanding, we believe it to be a well-founded method and we hope it will be adopted for general use!

Its advantage over the borrowed slope method is that it does not need a CER. Its advantage over the ratio method is that it is s better than blind application of a ratio when a CER is not at hand.

Parting words

Hopefully we have convinced you that ratio adjustment is just not good enough! At the very least, we most sincerely hope that the reader will never apply a ratio again without at least understanding the implications of the method.