"To b or Not to b"
The y-intercept in Cost Estimation

Presented at SCEA, June 2007

Authors:
Richard L. Coleman
Jessica R. Summerville
Peter J. Braxton
Bethia L. Cullis
Eric R. Druker
Northrop Grumman Corporation

Background

We have found errors concerning the y-intercept to be a widely-occurring problem. These errors are rampant among engineers, pricers, and others who use factors, rates, and data-based costing techniques; less common (but not non-existent) among cost estimators; more common among engineers who overemphasize engineering or physics as the basis of Cost Estimating Relationships (CERs.) In a prior paper[11] we discussed the dangers posed by the use of simple ratios of parameters in adjusting analogies. In this paper we will discuss the implications of ignoring or suppressing y-intercepts in CERs, in rates, in metrics or thumb rules, and in analogies

Schools of thought concerning the y-intercept

First, let us consider 3 schools of thought on y-Intercepts. We have said that the y-intercept is sometimes missed inadvertently. We find that the y-intercept is a litmus test among cost estimators, and we will now discuss beliefs about the y-intercept. There are about three schools of thought:

1. CERs must pass through the origin
2. CERs which 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.

1. "CERs must pass through the origin." The typical argument is "If I buy no product, I spend no money." The pro is that it sounds good. The con is that it doesn't seem to match the data. E. g., the price of Flash Drives.

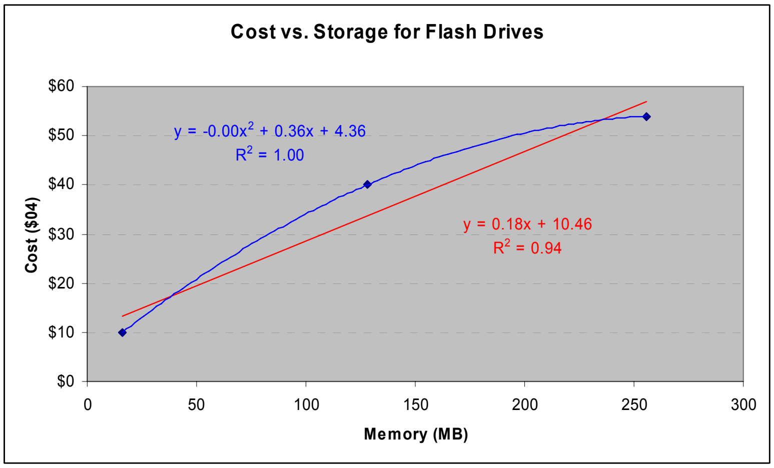


Figure 1: Cost vs. Storage for Flash Drives

[1] *Analogies: Techniques for Adjusting Them*, R. L. Coleman, J. R. Summerville, S. S. Gupta, So. MD SCEA Chapter, Feb 2004, ASC/Industry Cost/Schedule Workshop, Apr 04, SCEA 2004, MORS 2004

This is a simple data set, note that the two most likely curves both have a y-intercept between about $4 and $10, not zero.

2. "Y-Intercepts must make sense." The typical argument for this point of view is "There must be physics-based arguments for CERs." The pro is that it is helpful to think about physics and physical meaning, within reason. The con is that if practiced to the extreme, good CERs can be rejected just because we do not yet understand them, and that Engineers, who hate cost estimation and it's constraints, can usually talk the analyst to a full stop thereby derailing a potentially good CER.

3. "The Y-Intercept is just what it is." Typical arguments for this school of thought are, first, that we are not trying to predict the y-intercept, we are trying to predict the cost of systems of non-zero size, and second that we should take the best advice the data can give us. 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, 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. Finally, this approach is what the practice of statistics currently recommends. It should be noted that this argument gets little traction with engineers, who are trained to believe in literal meaning. The pros are that any existing system (i. e., one of the data points underlying the CER) is well-predicted. The con is that there is no literal meaning to the y-intercept, which is not very satisfying to literalists.

The Y-Intercept in CER Development

If they suspect a power equation may be appropriate, or sometimes to rule one out, analysts usually just take the log of both sides and conduct OLS, but the fit will be poor, and the regression probably will not be significant, as the red curve in the below graphic portrays. In Figure 2, the gray curved oval is meant to convey the region where data lie.

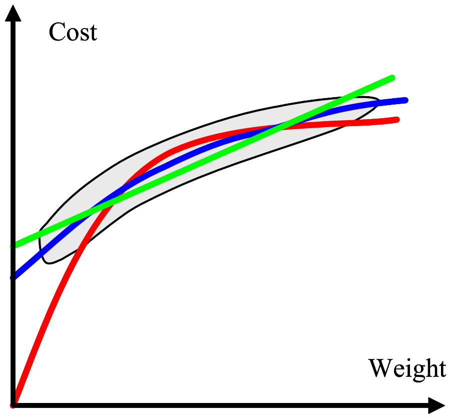


Figure 2: Cost vs. Weight (Conceptual - Curvilinear)

Supposing that the underlying data do support a power form, but that there should be a y-intercept; the higher the would-be y-intercept, the poorer the fit will be if a y-intercept is not specified. The less arced (more linear) the data, the poorer the fit will be. A linear form, y = a * x + b, like the green one above, may fit the data fairly well. With a higher y-intercept, the linear model fits better than the power curve (red). The more arced the data are, the worse a linear form will fit. The best fit would be the blue power curve In an earlier paper[22] we noted that when fitting a power curve, estimators often forget that Ordinary Least Squares (OLS) cannot deal with an equation of the form $y = a * x^c + b$ To perform OLS, the data must be linear, but the log of this equation is not a linear form Not understanding this, analysts proceed with a mechanical approach

There are several ways to fit the blue curve, including Excel Solver. Unfortunately, there is no test of significance, but there is at least one approach to the fitting of confidence and prediction intervals[3].

Sometimes a similar result occurs because the analyst insists that the CER "should" go through the origin ("the cost of nothing should be zero"), and that the y-intercept "must make sense." Refer to the Figure 3 in which the data are more linear in appearance This insistence can lead to rejection of the green line and the blue curve and choice of the red curve or the purple line - the red curve or purple line, though bad fits, are the best choice that go through the origin.

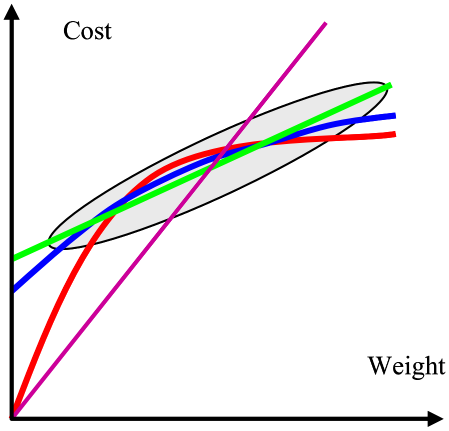


Figure 3: Cost vs. Weight (Conceptual - Linear)

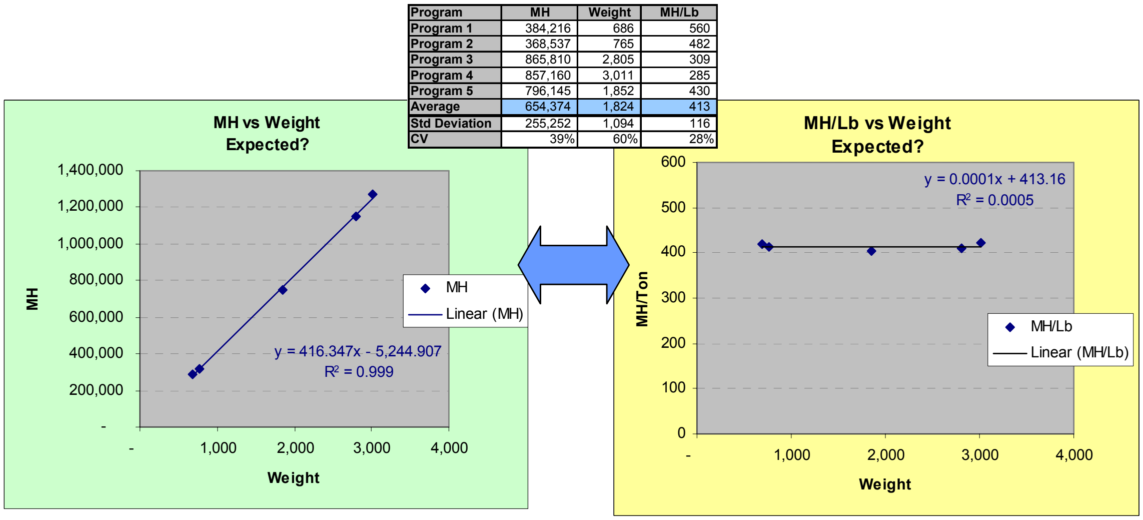
The Y-Intercept and Rates (Also known as Factors)

Suppose we have the below data (the data are real, but are adjusted to avoid exposing their identity. What should we plot? What do we expect? Because we suppose that MH are related to weight, and so MH/Lb might be constant, there might be a good rate to use

[2] *Cost Response Curves - Their Generation, Their Use in IPTs, Analyses of Alternatives, and Budgets*; K. E. Crum, K. L. Allison, R. L. Coleman, R. Klion, 29th ADoDCAS, 1996

[3] *Prediction Bounds for General-Error-Regression CERs,* Stephen A. Book, 39th ADoDCAS

for our estimate. We should look at the data before guessing. We are implicitly expecting data as portrayed below.



Program	MH	Weight	MH/Lb
Program 1	384,216	686	560
Program 2	368,537	765	482
Program 3	865,810	2,805	309
Program 4	857,160	3,011	285
Program 5	796,145	1,852	430
Average	654,374	1,824	413
Std Deviation	255,252	1,094	116
CV	39%	60%	28%

Figure 4: The Expected Relationship between Manhours and Weight

We should always compute the coefficient of variation, and here, we should note that the CV of the supposed rate, shown in Table 1 below, is quite high, not much better than MH were to begin with, so we're not done (our rule: CV should be < 10-15% to use a rate.)

Program	MH	Weight	MH/Lb
Program 1	384,216	686	560
Program 2	368,537	765	482
Program 3	865,810	2,805	309
Program 4	857,160	3,011	285
Program 5	796,145	1,852	430
Average	654,374	1,824	413
Std Deviation	255,252	1,094	116
CV	39%	60%	28%

Table 1: Manhours and Weight

Now, our apologies for misleading you with the graphics we showed before, in Figure 3, those aren't really the graphs of the data, it was just what we expected. We should always look at a scatter plot of the 2 variables in our rate … the real graph. The real data is plotted below:

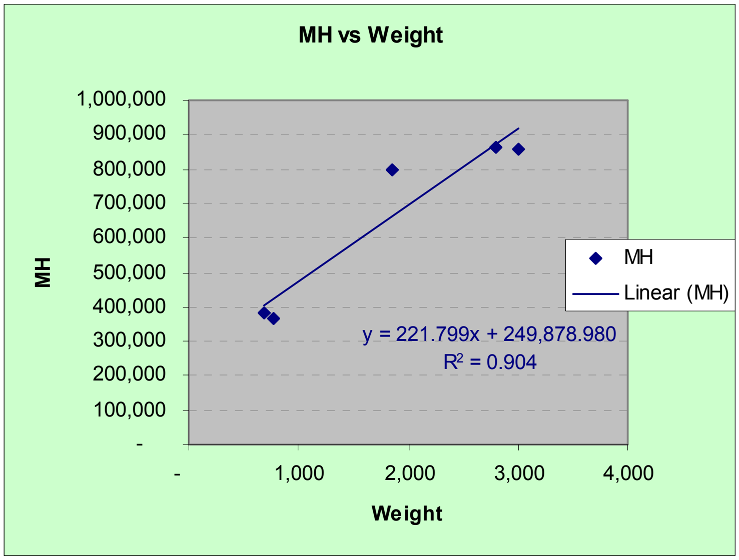


Figure 5: The Real Relationship between Manhours and Weight

What we have is a *failure to graph* - we didn't think the rate was a function of weight but, it was, it is not a rate at all. There is a CER because we overlooked the pesky (large) y-intercept! We now should consider a CER on MH as a function of weight. Given the r^2, we can expect a CV for our CER of 0.904 * 39% = 3.7%.

The Y-Intercept and Metrics (Also known as Ratios or Thumb Rules)

Suppose we have the below data (the data are real, but are adjusted to avoid exposing their identity. What should we plot? What do we expect? We were probably expecting some sort of thumb rule to emerge, and because we suppose that MH/Lb might be constant, there might be a good thumb rule we could use. We were probably expecting the below graphic.

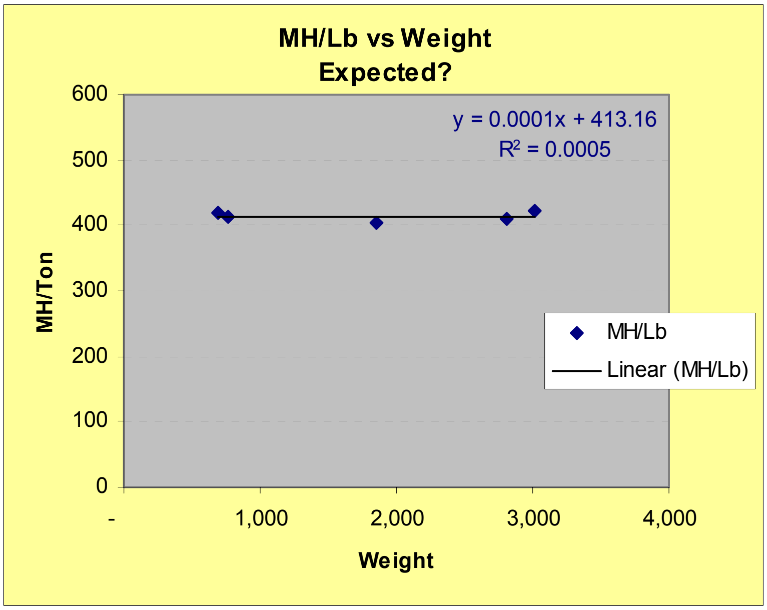


Figure 6: The Expected Relationship between Manhours/Pound and Weight

Now, our apologies for misleading you with the data we showed before, those aren't really the graphs of the data, it was just what we expected. We should always look at a scatter plot of the 2 variables in our rate … the real graph. The real data is plotted below in Figure 7:

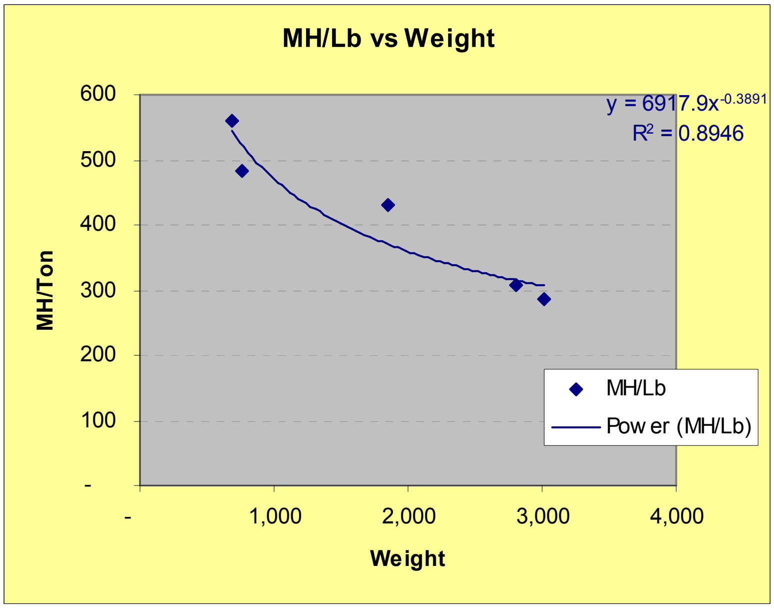


Figure 7: The Actual Relationship between Manhours/Pound and Weight

We actually got a very interesting, even compelling graphic. MH/Lb clearly decreases, and seems to follow a smooth curve. What is happening here? What does this mean? We might form the below hypotheses:

Hypotheses:

1. Larger units are less complex, and so the work is less demanding
2. Larger units have less density, thus are easier to work in/on
3. Larger units have thicker structure, which is easier to work with, being less likely to deform, easier to weld, etc.

Think for a moment … which of these hypotheses is right? Having thought about this for a moment, we should consider that there is a second important graph, Figure .

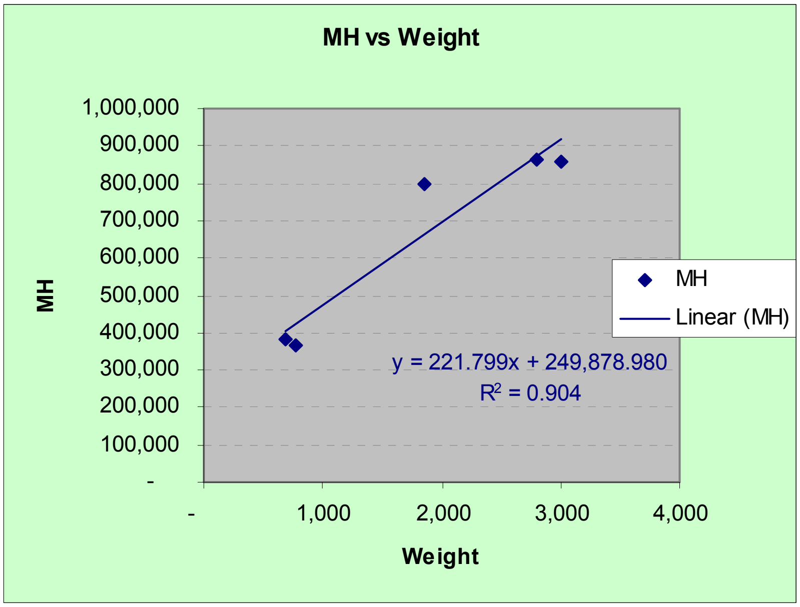


Figure 8: The Actual Relationship between Manhours and Weight

We naturally would plot MH/Lb vs. Weight and MH vs. Weight and regress both. Both regressions seem good, which one do we choose? MH and Weight – Two Choices
But, first, what does it mean to say MH/Lb is a power function of Weight (or Complexity or Density?) Let us see.

Suppose $y/x = a * x^{c}$
Then $y = a * x^{c+1}$

In our equation, y is MH ,x is weight

So, if $MH/Lb = 6918\ Wt^{-.3891}$
Then $MH = MH/Lb *Lb = 6918 *Wt^{-.3891} *Wt = 6918\ Wt^{.6109}$
So, let's plot the linear equation from the last slide and this new equation in figure 9 and see why they both work.

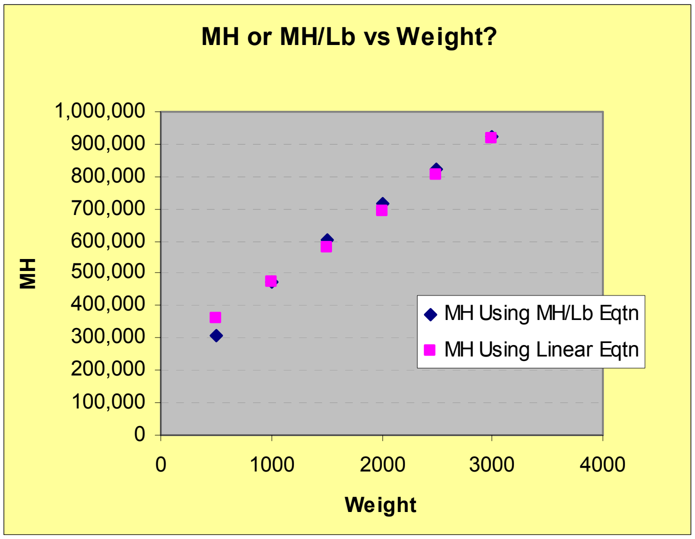


Figure 9: Comparison: Manhours and Weight or Manhours/Lb and Weight

They are almost identical, why choose the more complex power equation? What does each equation say? First, let us consider the linear equation:

Linear: $MH = a * Wt + b$

As weight rises, cost rises, proportionately. Like most CERs, this one does not pass through the origin. As we noted, some analysts say it should, "If there's no weight, there's no cost." But statisticians face this problem all the time. They are prone to dismiss any attempts to attach meaning to the y-intercept and to accept it regardless. Since a zero value of the explanatory variable is not part of the data, we cannot say much about the equation there … requiring any equation to do well at the zero-x value is too restrictive. In other words, the acceptability of this equation depends on one's beliefs. Now let us consider the non-linear equation:

Non-linear: $MH/Lb = a * Wt^{c}$

The usual explanation is that some sort of variable like density or complexity is driving a shift. This equation doesn't pass through the origin either, so the school of thought that demands a zero intercept could be dissatisfied, although we have not often heard this expressed. More importantly, this equation has a hidden trap … if density is a driver, we should plot it, not just assert a smooth "density" variable that moves with size and just wave it away … but density is arcane and cannot be plotted, so proponents of this equation (and here are some) just wave their hands and assert an unseen density, which eludes plotting. In other words, the acceptability of this equation depends on one's beliefs as well.

The real point is that the odd form of the non-linear equation comes about because of the y-intercept … if there were no y-intercept, the graph would be flat. We submit that the non-linear equation is merely an artifact that the search for meaning is a wild goose chase, but, of course, this comports with *our* beliefs

The Y-Intercept and Analogies[4]

Considerable attention is devoted to techniques in the development of CERs for parametric estimating particularly by higher-echelon cost shops and agencies. Considerable expertise is also to be found in buildup techniques, partly because many Original Equipment Manufacturers (OEMs) have large cost shops which practice buildup. Analogy, on the other hand, has been given little attention. This subject was also treated in an earlier paper[4] by the authors.

The current method typically (almost always) involves adjustment by ratio. Adjustments, both in the analogy or the buildup method, typically rely on an "obvious" characteristic. The characteristic used for adjustment is most often weight, Software Lines of Code (SLOC), number of users, linear feet of cable or some such variable. Sometimes weight, or another characteristic of the new system is not known, and so another characteristic is used (often as a proxy for weight), such as bore diameter of a gun. Usually the ratio of the value of the characteristic in the new system to the value in the old system is multiplied by the cost of the old system. In these cases, there may be a presumed relationship to weight, and 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, is cubed.

Let us examine the implications of the current method. An example adjustment by ratio is:

> Suppose the analogy weighs 300 tons and costs $100M
> Suppose the new system weighs 500 tons
> Then we would suppose the new system will cost (500/300) * $100M = $166.67M

This 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 (the reader will, we hope, by now, have noticed that we most strongly advocate scatter plots to understand the messages in one's data!) There is a surprise in the scatter plot for most of us … but first, force yourself to predict what the line between the analogy and the prediction looks like … where does it cross the y-axis? The below graph, Figure 10, shows the previous adjustment.

[4] *Analogies: Techniques for Adjusting Them*, R. L. Coleman, J. R. Summerville, Northrop Grumman; S. S. Gupta, IC CAIG, ASC/Industry Cost & Schedule Workshop Spring 2004, SCEA 2004 72nd MORSS - 2004

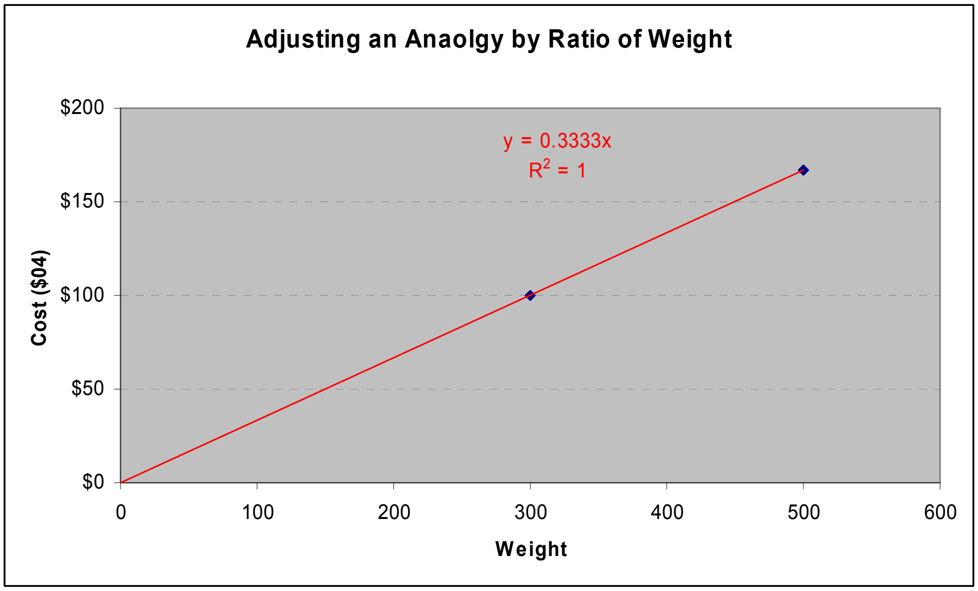


Figure 10: Adjustment of an Analogy by Weight Ratio

Note that the line through the 2 points passes through the origin. We would propose two new methods. The first method, we call "The Borrowed slope Method.[5]" This method is a variant of the methods for calibrating CERs. In this method, we adjust a "trusted analogy" by a "trusted slope." The second method is called "Relational Correlation[6]" and was treated in the paper footnoted earlier, on adjusting analogies. This method takes advantage of the geometry of the bivariate normal and regression. In the "Relational Correlation method, we adjust a "trusted analogy" by a "best guess slope." The Borrowed Slope Method is covered below, but the Relational Correlation Method is considerably more complex, and is beyond the scope of what we intend this paper to cover, so it is in Appendix A.

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.

[5] *Analogies: Techniques for Adjusting Them*, R. L. Coleman, J. R. Summerville, Northrop Grumman; S. S. Gupta, IC CAIG, ASC/Industry Cost & Schedule Workshop Spring 2004, SCEA 2004 72nd MORSS - 2004

[6] *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*

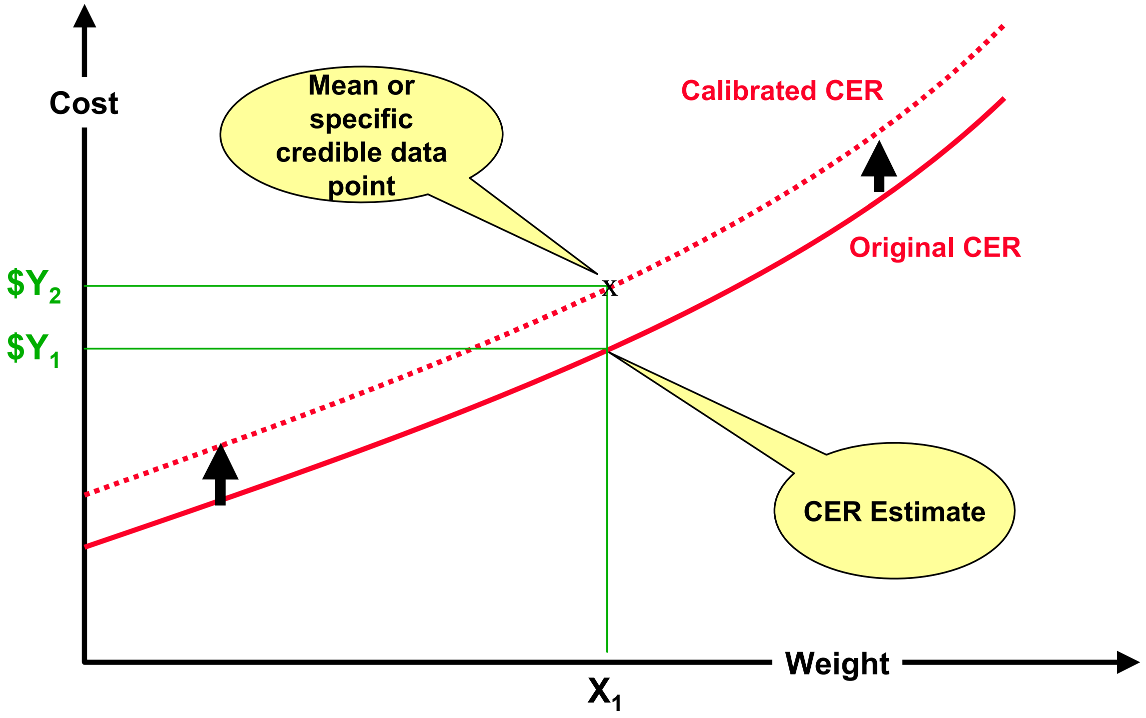


Figure 11: Calibrating a CER

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 Figure 12.

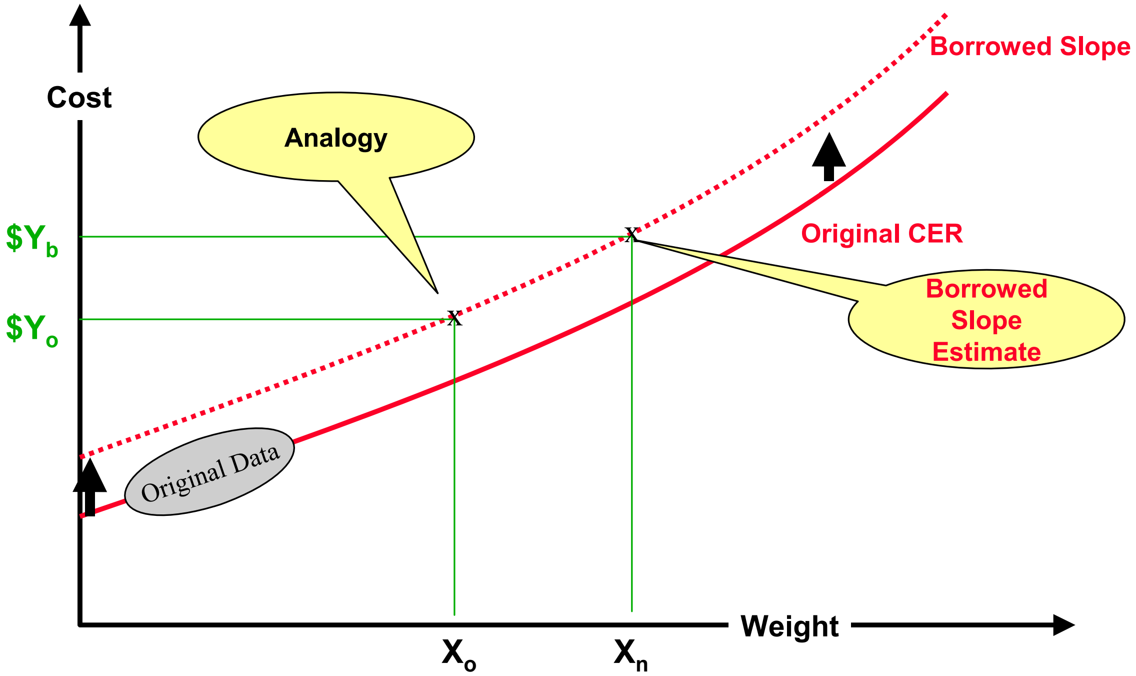
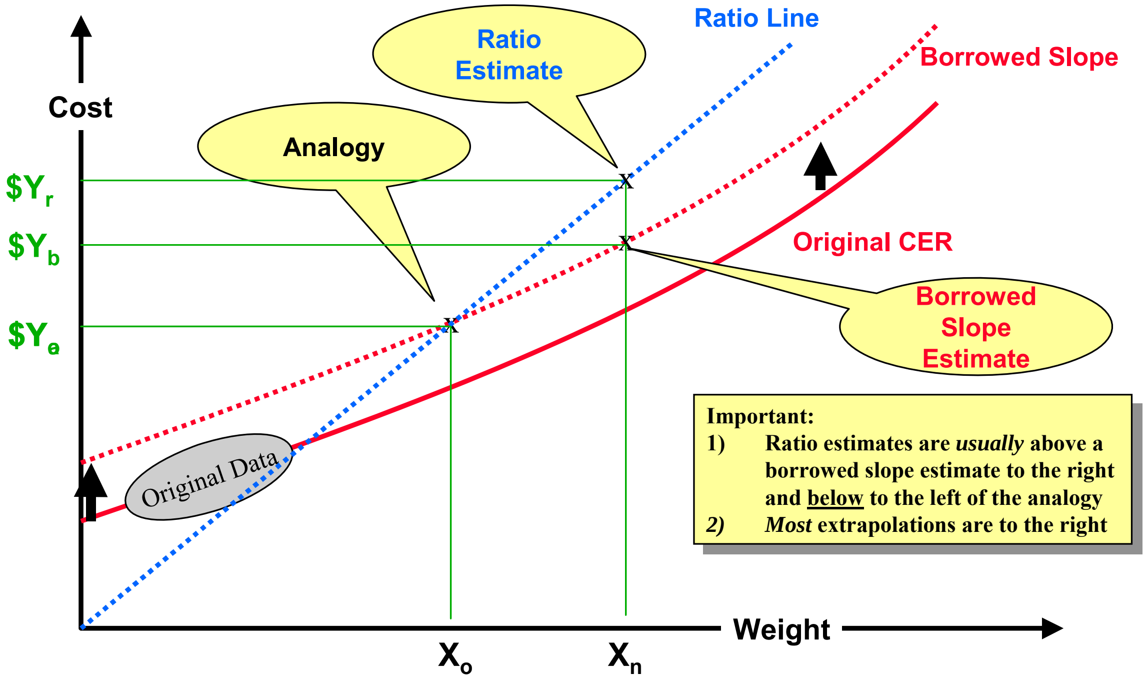


Figure 12: The Borrowed Slope method

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 Figure 13. As can be seen in Figure 13, there can be considerable difference between a borrowed slope adjustment and a ratio adjustment. In general we develop “bigger, faster” and the like, and the values of parameters are usually above those of the analogy, so we tend to over estimate with the ratio method.



Important:
1) Ratio estimates are *usually* above a borrowed slope estimate to the right and below to the left of the analogy
2) *Most* extrapolations are to the right

Figure 13: The borrowed slope method compared to the ratio method

Conclusions

We have come to believe that the y-intercept is a poorly understood part of cost estimation, and have observed a number of significant problems that arose because of failure to consider them. We discussed the 3 schools of thought with regard to y-intercepts. We discussed CERs, rates, metrics (thumb rules), and analogies with regard to the y-intercept. We discussed the implications of ignoring the y-intercept, and of including it in each of these areas, as well as the confusion that the y-intercept can cause in metrics and ratios. We discussed a method of adjusting analogies that takes the y-intercept into account.

We urge you to think about this. We are less offended at your holding beliefs (after all, we do!) than by sailing by this issue all unawares. Of course, if we had our way, we'd hope you believed in y-intercepts[7], but in any event, think about it!

[7] Clap if you believe in y-intercepts: In the second act of Walt Disney's "Peter Pan", Tinkerbell drinks poison that Peter is about to drink in order to save him. Peter turns to the audience and says, *"Tinkerbell is going to die because not enough people believe in fairies. But if all of you clap your hands real hard to show that you do believe in fairies, maybe she won't die."* We all started to clap. I clapped so long and so hard that my palms hurt. Then suddenly the actress playing Peter Pan turned to the audience and she said, *"That wasn't enough. You did not clap hard enough. Tinkerbell is dead."* We all started to cry. The actress stomped off stage and refused to continue the production. They had to lower the curtain. The ushers had to come help us out of the aisles and into the street. You hear that? CLAP LOUDER!

Appendix B

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.

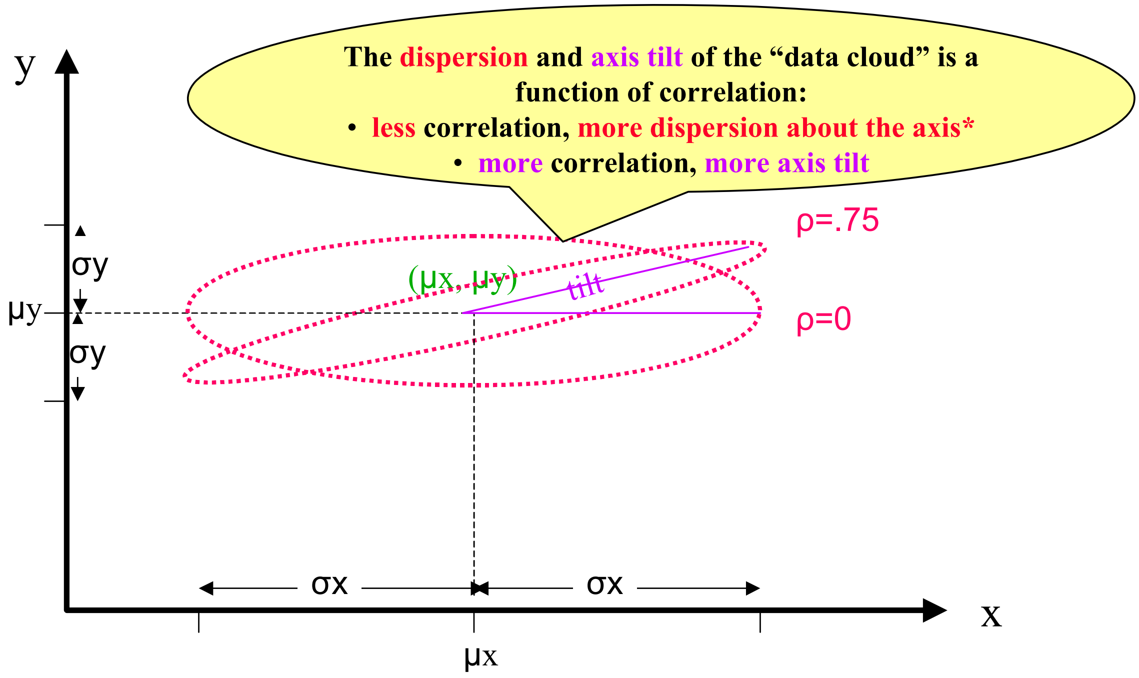
Bivariate Normality

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:



The dispersion and axis tilt of the "data cloud" is a function of correlation:
• less correlation, more dispersion about the axis*
• more correlation, more axis tilt
y
σy
μy
σy
(μx, μy)
tilt
ρ=.75
ρ=0
σx
σx
x
μx

Figure 14: 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:

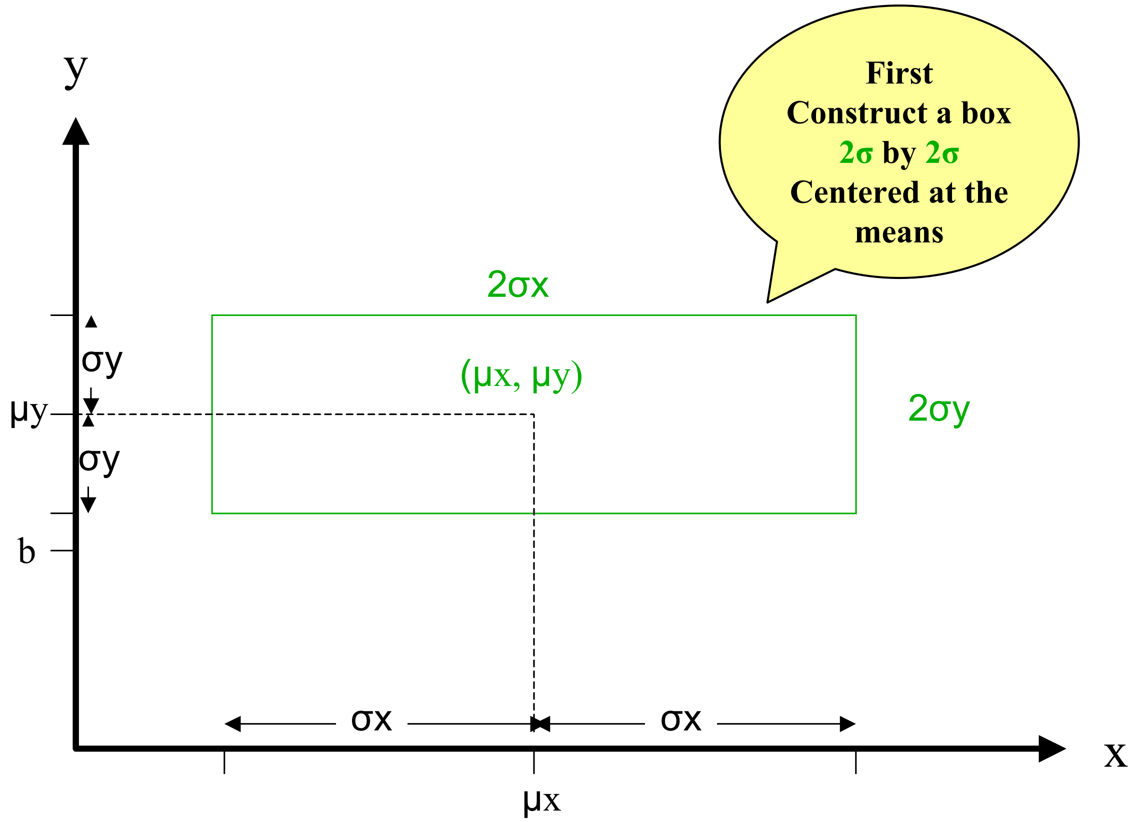


Figure 15: Bivariate normality – constructing the box

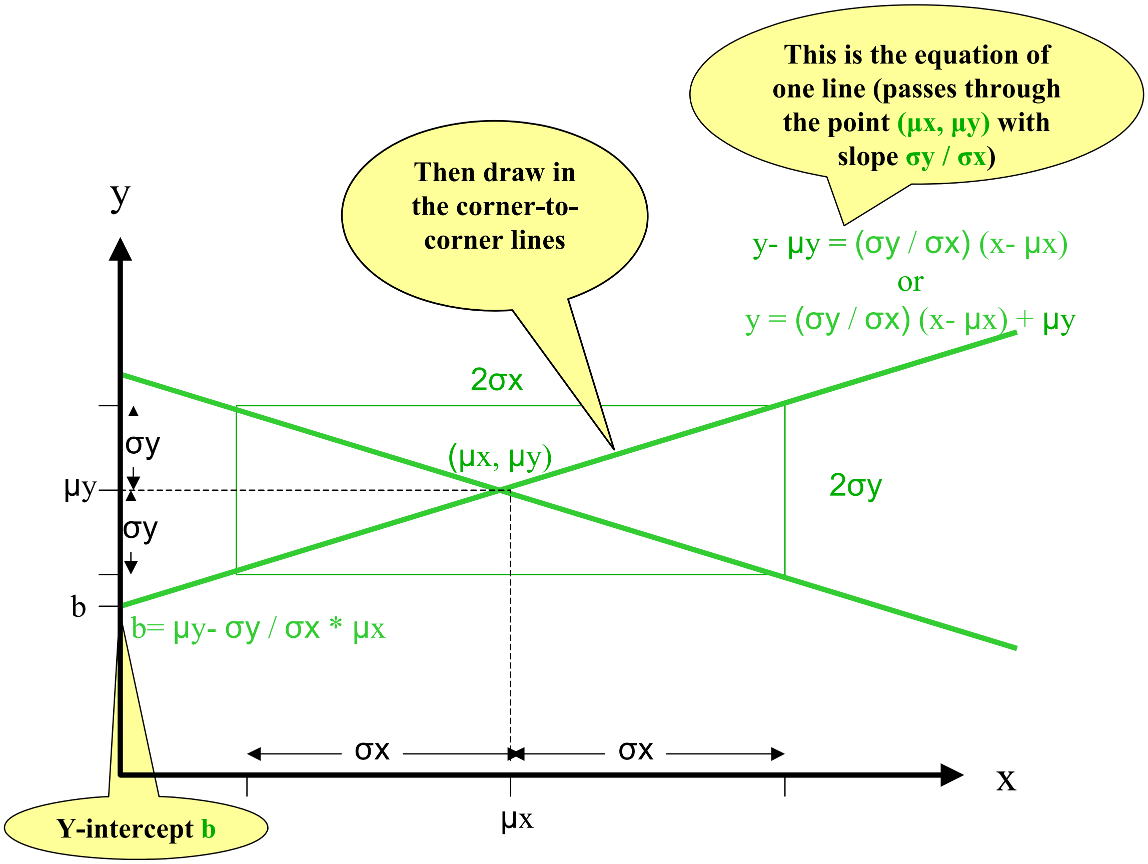

This is the equation of one line (passes through the point (μx, μy) with slope σy / σx)
Then draw in the corner-to-corner lines
y- μy = (σy / σx) (x- μx)
or
y = (σy / σx) (x- μx) + μy
y
2σx
(μx, μy)
2σy
σy
μy
σy
b
b= μy- σy / σx * μx
σx
σx
x
μx
Y-intercept b

Figure 16: Bivariate normality – inserting the diagonals

Now, populate the box with data, shown here as the already illustrated ‘data clouds.”

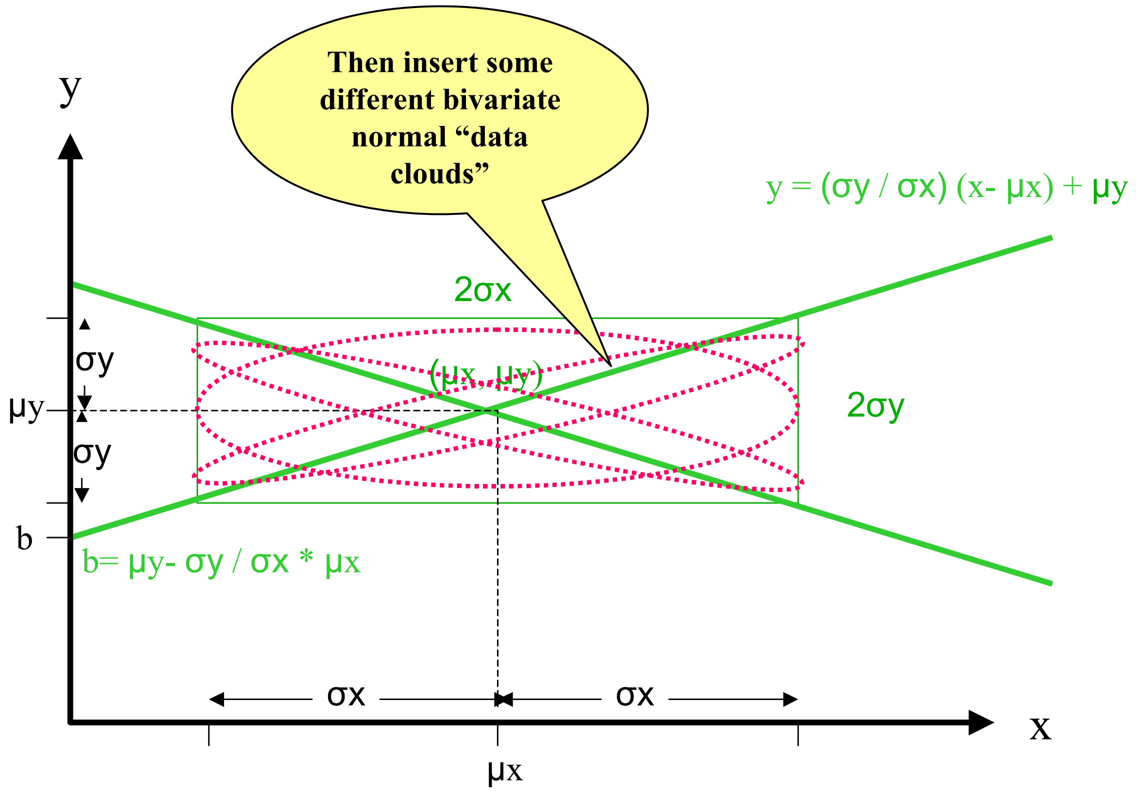

Then insert some different bivariate normal "data clouds"
y
y = (σy / σx) (x- μx) + μy
2σx
σy
(μx, μy)
μy
2σy
σy
b
b= μy- σy / σx * μx
σx
σx
x
μx

Figure 17: 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 Regression

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.

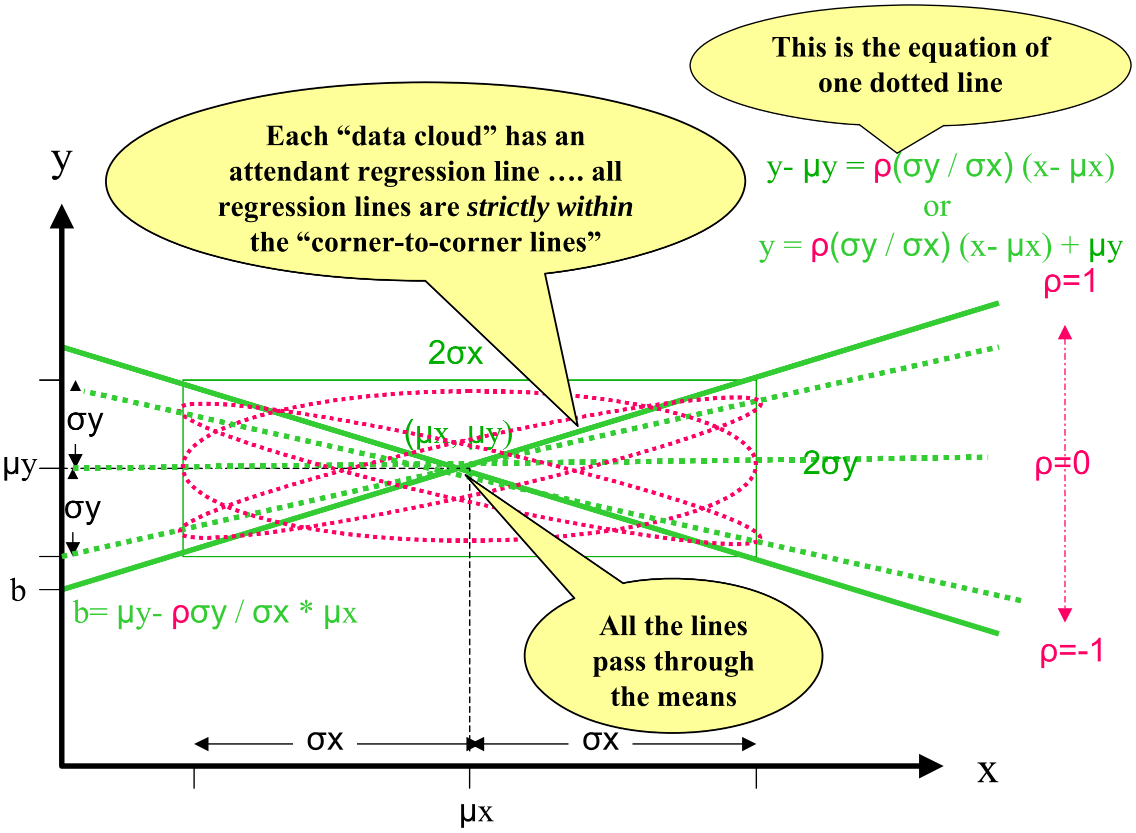

This is the equation of one dotted line
y- μy = ρ(σy / σx) (x- μx)
or
y = ρ(σy / σx) (x- μx) + μy
Each "data cloud" has an attendant regression line …. all regression lines are *strictly within* the "corner-to-corner lines"
y
ρ=1
2σx
σy
(μx, μy)
μy
2σy
ρ=0
σy
b
b= μy- ρσy / σx * μx
All the lines pass through the means
ρ=-1
σx
σx
x
μx

Figure 18: The geometry of regression - the implications

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

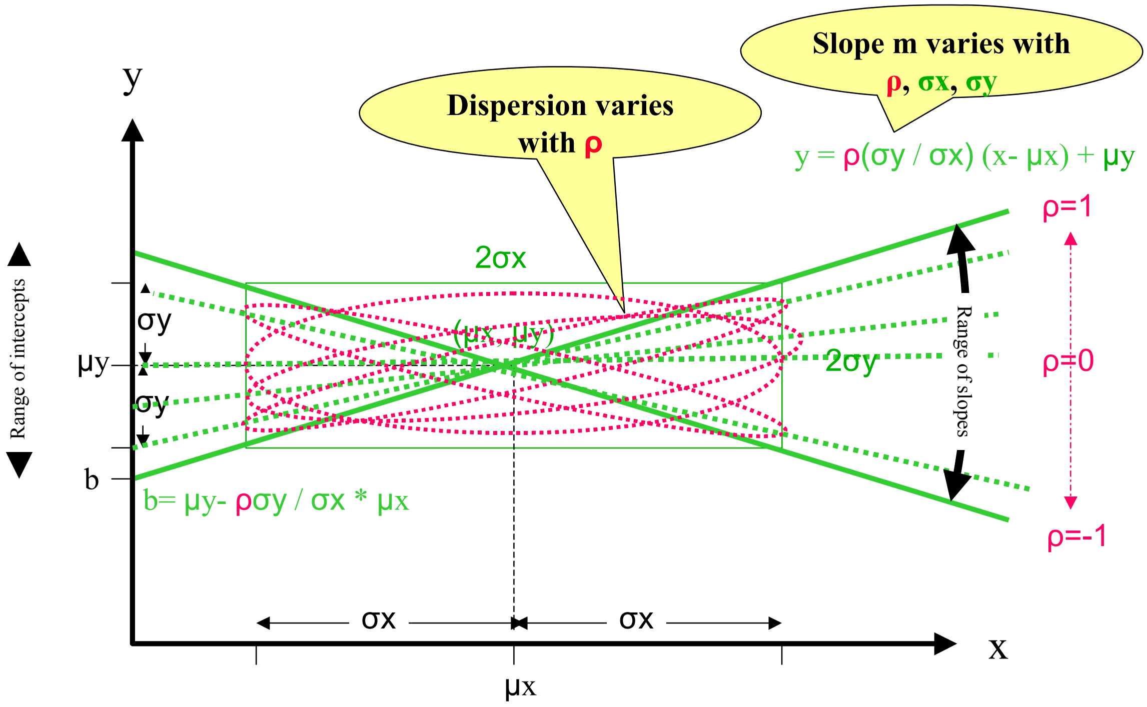


Figure 19: The geometry of regression - the effect of the parameters ρ, σy and σx

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!

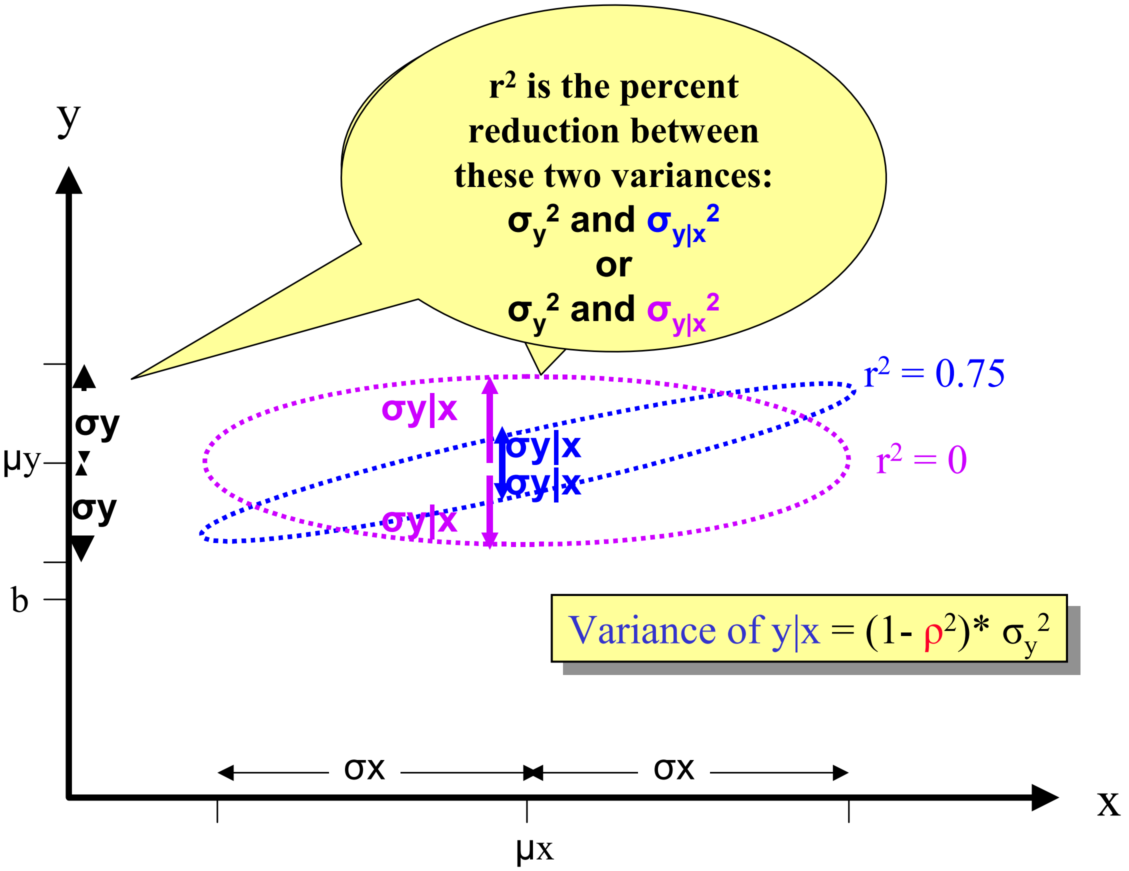

r^2 is the percent reduction between these two variances:
σ_y^2 and $\sigma_{y|x}^2$
or
σ_y^2 and $\sigma_{y|x}^2$
y
x
μy
b
σy
σy
σy|x
σy|x
σy|x
σy|x
$r^2 = 0.75$
$r^2 = 0$
Variance of y|x = $(1-\rho^2)*\sigma_y^2$
σx
σx
μx

Figure 20: The meaning of r2

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.

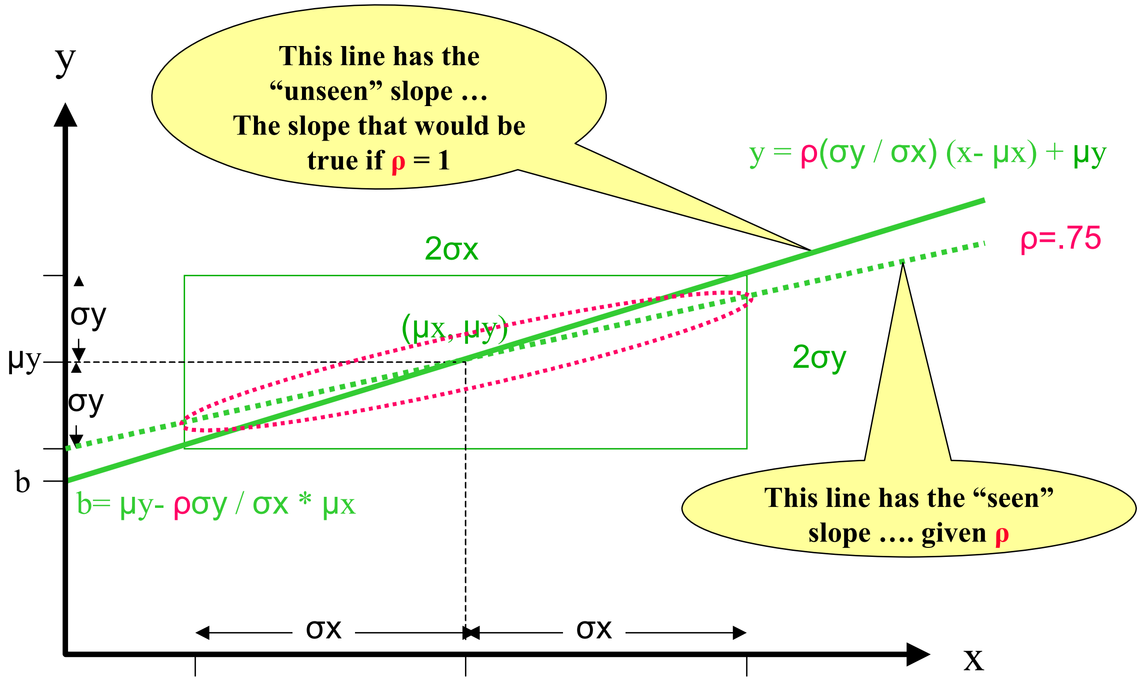

y
This line has the "unseen" slope … The slope that would be true if ρ = 1
y = ρ(σy / σx) (x- μx) + μy
ρ=.75
2σx
σy
μy
σy
(μx, μy)
2σy
b
b= μy- ρσy / σx * μx
This line has the "seen" slope …. given ρ
σx
σx
x

Figure 21: The implications of the geometry of regression

Implementing Relational Correlation for Analogies (and buildups)

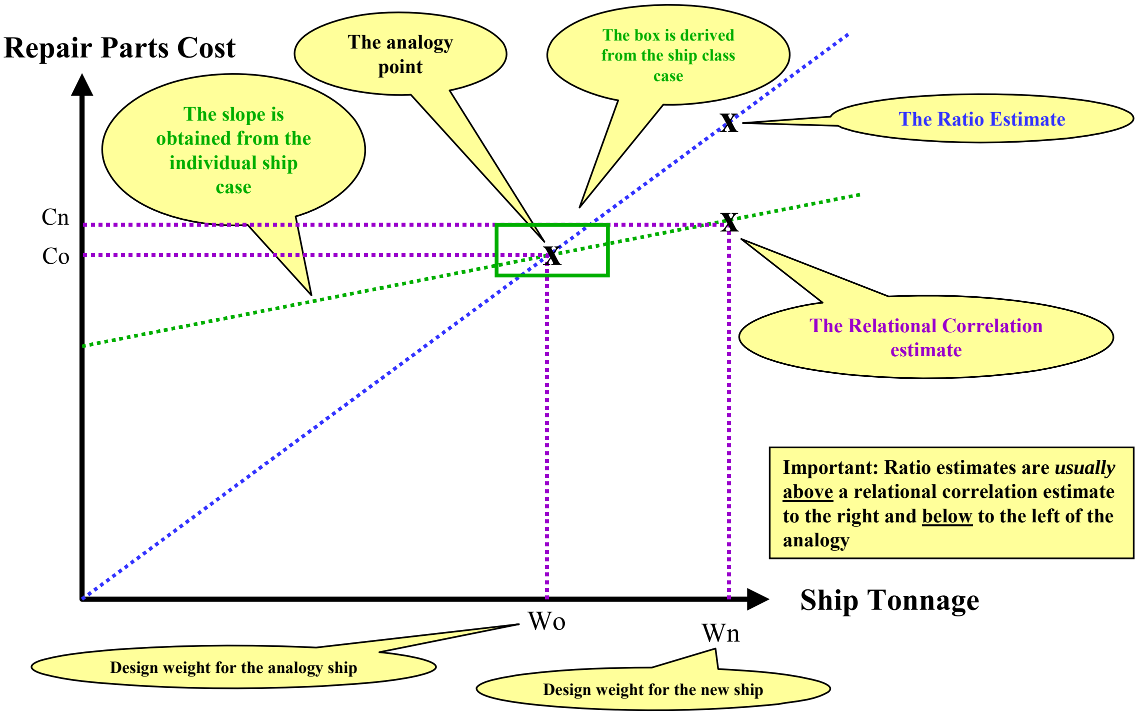
For Single Point Analogies

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.


Repair Parts Cost
The analogy point
The box is derived from the ship class case
The slope is obtained from the individual ship case
The Ratio Estimate
Cn
Co
X
X
X
The Relational Correlation estimate
Ship Tonnage
Wo
Wn
Design weight for the analogy ship
Design weight for the new ship

Important: Ratio estimates are ***usually*** **above** a relational correlation estimate to the right and **below** to the left of the analogy

Figure 22: The relational correlation method pictorially

Appendix B

Definitions

Adjustments: Scaling of a cost by some physical, performance, or other such attribute Scaling is usually (in practice) directly proportional to the attribute. Scaling parameters are usually countable or measurable and intuitively tied to cost.

Analogies: 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” are almost always made

Buildups: Costed-out physical Bill of Materials (BOMs) and CAD-generated material lists and the like We do not mean “buildups” consisting entirely of Staffing levels multiplied times duration. Such estimating techniques are little more than “engineering judgment” in fine detail. Buildups often include “adjustments” to allow for size differences.

Composite methods: A method that involves at least two of the three other types.

Parametric Estimates: Estimates made by developing statistical “Cost Estimating Relationships” (CERs) based on one or more parameter and cost Estimates involving parameters but not based on statistical analysis are more properly called either “adjusted analogies” or “adjusted buildups”