Improving CER building
Getting rid of the R² tyranny

Building a CER with the median

> When you get – in a practical situation – more information than the number of unknowns, the problem of how to combine this information is certainly the most important one you face for applying mathematics to natural philosophy[1].
>
> Karl Friedrich Gauss. 1809.

When the cost estimator possesses costs (or any other attribute, such as the schedules) for a set of products related to the same "family", his main concern is to answer the fundamental question : "how can I use this information to estimate the cost (or the attribute) for any new product belonging to the same family ?".

If we are dealing with cost – as I will do in this paper – I assume that these costs have already been normalized, using all necessary tools.

I also assume that the search for the influential variables[2] (the variables – or "cost drivers" – which really influence the cost) has been done. Several techniques, so such as the Cp statistic or the partial regressions, allow to do so.

We know that products belonging to the same family differ by these variables, such as the size, the material, etc… depending on the products you are interested in. Therefore I intend to use these variables to "explain" the cost differences between the products.

"Using the same information" means I intend to remain at the same level of the WBS as it was for the products of my data base. This means that detailed – or, as the adjective is sometimes used, analytic – cost estimating is not considered here.

Basically there are two ways to answer answering the fundamental question : parametric and non-parametric. These approaches are completely different.

Parametric means : "I want to use a given type of formula to represent the relationship between the cost and the variables". This type of formula can be linear – or additive – such as

$$cost = \alpha + \beta \times x_1 + \gamma \times x_1^{\frac{5}{4}} + \delta \times \ln x_2 + \cdots$$

which includes some parameters (called here α, β, γ, …) which will be used to adjust the of formula to the data. It should be noticed that this type of formula is perfectly valid : "linear" means "linear to the parameters", not to the variables. Another familiar type of formula is multiplicative, such as

$$cost = \alpha \times x_1^{\beta} \times x_2^{\gamma} \times \cdots$$

[11] using the vocabulary of the period.

[2] please do not use the word "parameter" to name these variables. Mathematically speaking – and this is the way this word is used in this paper – a parameter is an auxiliary variable which is used to change the behavior of a function. For instance the coefficients of a linear formula are called "parameters".

Other types of formulae can be – and are – used, these two being the most frequently selected.

Non-parametric means : I do not want to force the relationship to be a selected one, and I don't even want to use any algebraic relationship (and therefore there will be no parameter to compute[3]). A typical example (widely used) is to say : I want all my estimates to be in the vicinity of the current data points ; of course the "vicinity" has to be defined and it can be rather complex !

My purpose here is not to compare these two approaches (that will done later on), but only to investigate the parametric one and how it can be done.

Parametric needs therefore to make a first choice : the type of formula to be used.

The second choice is related to the metric. What does this term mean ? We will compute the values of the formula parameter by adjusting them to the data points we have. Adjusting implies that we are able to measure the "distance" between the formula and the data points. The generally used metric – the one used by Gauss – is the Gaussian metric (let's call it this way), which is the squared difference between the cost value computed – for each data point – by the formula and the cost value stored in the data base.

In his paper Gauss did not explicitly mention this metric. He just said that it is a custom[4] (yes, just a custom !) that, if a quantity is obtained by several observations, the arithmetic average is generally considered as the most likely value of this quantity. And it is not difficult to demonstrate that this implies the use of the Gaussian metric. Suppose we have a distribution of the n measurements (x_1, x_2, …) of a given quantity. We need a value that could replace, for some other computations, this distribution. In order to find it, we would like to obtain a value (generally called the center of the distribution), let's call it μ, which would be as close as possible to all the values of this distribution. To do so, we have to define what we mean by "close" ! Here comes the need for a metric ; using the Gaussian metric we decide that the distance between measurements x_i and x_j will be given by $(x_j - x_i)^2$. Therefore now, the value μ we want to compute will have to minimize the sum

$$\sum_i (\mu - x_i)^2$$

Minimizing the derivative on μ, you will immediately see that $\mu = \frac{1}{n}\sum_i x_i$. This is exactly Gauss' choice.

Using this metric (do not forget it is a choice) Gauss demonstrated that the distribution, which he called $\varphi(\Delta)$, of the errors[5] – between real observations and true values which should be observed – must follow the law

$$\varphi(\Delta) = \frac{h}{\sqrt{\pi}} e^{-h^2 \Delta^2}$$

where h is a "parameter". The "least squares method", so much used nowadays, originated from this discovery. It is important to know where it comes from … It is only the result of the decision to use the Gaussian metric to define the "center" of a distribution !

There are other methods used to combine information and Gauss was perfectly aware of them (he mentioned several times the work made by Laplace, who was not using the same metric). He just added that the least squares method should be preferred for the only reason that it makes computations easier. We must not forget that, at his time, all the computations were made manually. Therefore the

[3] the word "non-parametric" comes from this situation.
[4] he also called it " a simple principle, generally adopted".
[5] he called them "errors" because he was dealing with astronomical observations. For a relationship between variables, the term "residuals" (or "unexplained variations") is certainly better.

question we are trying to answer in this paper is the following one : doesn't the computer allow us to use other, more satisfactory, metrics and computational methods ?

From now on I will concentrate on the problems that the cost estimating community faces. These problems do not deal with the distribution of one variable (except for the persons who prefer to use "specific" costs, such as the cost per kg or pound) but with the search of a relationship between one variable – generally the cost – supposed to not depend on other variables.

The least squares method has, for us, three important drawbacks : what is called the "regression", the sensitivity to outliers and the difficulty of using correlated variables.

The "regression"[6] is very well known (so well known that the least squares method is often called "regression analysis") and I mentioned it in a previous paper. It isn't the case for its consequences. The fact is that, if observations are a little scattered, the least squares algorithm underestimates the cost of large products (products larger than the average) and overestimates the cost of small products, because this algorithm "regresses" the estimated cost towards the average cost ; the more the observations are scattered, the more the estimates regress … As the observed costs are generally rather scattered, this is a very serious drawback as I do not see why cost should regress towards the average cost !

The sensitivity to outliers is also familiar to most cost estimators. Its origin comes from what is called the "breakdown". The breakdown of an estimator is the smallest proportion of the data that, if we change them, can have an arbitrary large effect on the estimated value of this estimator. The Gaussian metric produces values which have a very low breakdown ; for instance the arithmetic average has a breakdown of only *1/n* : the change of only one data in a distribution can have a severe impact on its value. The Gaussian metric is not "robust" at all.

Quite often we have to use several variables in order to "explain" the cost. If these variables are linearly correlated, the result is very poor, which means that the relationship does not really make sense and produces estimates with a very large confidence interval. This comes from the fact that, in order to compute the value of the parameters, we have to find the inverse of a matrix based on the inputs ; it is well known that a matrix which has two proportional columns – which is the case if two variables are linearly correlated – has a determinant equal to 0.

Another drawback of the Gaussian metric (for our studies) is that it greatly favors large cost values. The reason is obvioud : in the cost domain (the one we are interested in here), the precision of our cost measurements is always known as a percentage of the figures (this is not true in other domains[7] where the accuracy of the measurements is a fix value, independent of the size of the measurements) ; it means that the imprecision of large costs is important compared to the ones for small costs. Consequently, as the Gaussian metric uses the squares of the differences, the least squares algorithm pays more attention to large costs than to small costs[8].

For all these reasons it seems reasonable to look for another metric which would not present these inconveniences.

Let's return briefly to the concept of metric. A metric on a space E is a rule $d(x,y)$ of $E \otimes E$ to R that assigns to each couple x,y a value, called the distance $d(x,y)$ between x and y. Such a rule cannot be defined as you want : it has to follow the following properties[9] :

- $\forall x, y \in E, d(x,y) = d(y,x)$

[6] the term was coined by Francis Galton.

[7] such as the ones in which Gauss was interested.

[8] this is why some authors recommend to minimize $(y_i - \hat{y}_i)^2 / y_i^2$ instead of $(y_i - \hat{y}_i)^2$.

[9] for the reader who is not familiar with these notations, the first property must be read : whatever x and y belonging to the set E, then the distance between x and y is equal to the distance between y and x (symmetry).

- $\forall x, y \in E, d(x,y) = 0 \Leftrightarrow x = y$
- $\forall x, y, z \in E, d(x,z) \leq d(x,y) + d(y,z)$

Let's investigate the property of the following metric (although Laplace did not mention it explicitly, I will call it here the "Laplacian metric" because he used it) :

$$d(x_i, x_j) = |x_i - x_j|$$

Using this metric has important two consequences :

- the center of a distribution is not the arithmetic average anymore, but the median, which I called ν,
- the measure of the dispersion around this center is not the standard deviation anymore but the quantity $\tau = \frac{1}{n}\sum_i |x_i - \nu|$.

Let's investigate the median for a distribution, which means for one variable only. Suppose you have a discrete set (of size *n*) of values $x_1, x_2, \ldots, x_i, \ldots, x_n$.

There are two definitions of the median :

- the first one is well known by you : the median is the value which leaves as many points above as below. This definition is valid when *n* is odd ; when *n* is even, one generally takes the average of the two values which are in the middle of the set.
- the second one is probably less popular : it is the value ν which minimizes the sum

$$S_1 = \sum_i |x_i - \nu|$$

Notice that this definition is very similar to the definition of the arithmetic average : the only thing which changes is the metric.
I demonstrated the equivalence of these two definitions of the median.

Another comment about the median when using a continuous variable (which I still call *x*) : you know about the "normal" distribution which plays an important role when the Gaussian metric is used :

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

which has a mean equal to μ and a standard deviation equal to σ. There is a similar law (notice that the metric has changed) to be used when dealing with the median ; I call it $M(x, \nu, \tau)$ and it is given by

$$M(x, \nu, \tau) = \frac{1}{\tau\sqrt{2}} e^{-\frac{|x-\nu|\sqrt{2}}{\tau}}$$

It has a mean equal to ν and a standard deviation equal to τ.

Let's compare these two distributions : figure 1 displays these distributions when their mean and median is 0 and their dispersion 1. Notice they are rather similar, except that the *M* distribution is sharper :

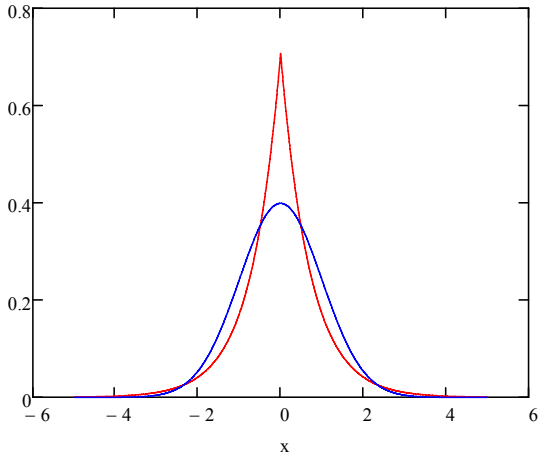

Figure 1

If the median is easy to define for a distribution, what can it be for a set of one "dependent" and several causal variables ? I will present the solution briefly in the case of one variable, but it remains true for several causal variables. To simplify, I will deal a linear (or additive) CER : a straight line ; but it can obviously also be used for a multiplicative one. Therefore we are looking for a straight line $y = a + bx$.

Fortunately enough, the second definition of the median will help us, but the first one will be used to start the discussion.

First of all we apply the first definition in order to find all the straight lines which leave as many data points above as below (figure 2) :

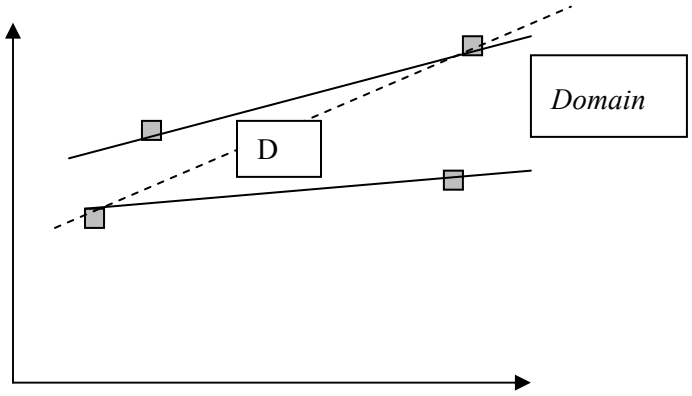

Figure 2

We call a "domain" a part of the plane containing all the straight lines so there are as many points above then as there are below if the number of data points is even, and so there are as many points ± 1 above them as there are below if the number of data points is odd. Now, using the second definition, we look for the line inside that domain which minimizes the sum S_1 of the absolute values of the distances between all the data points. Computations show two results

- as long as the line D remains inside the domain, then the derivative $\frac{\partial S_1}{\partial a}$ does not change,
- but $\frac{\partial S_1}{\partial b}$ changes. Therefore we use this property in order to minimize the sum S_1. Inside a domain there is a unique solution : the line D must "touch" one data which limits the domain upwards and one data which limits it downwards, as line D on Figure 2.

We do that for all the domains and find all the lines which could be used. Among all these lines we select the one for which the sum S_1 is the smallest. If several lines present – inside the limit of the computer – the same sum, we select the one which minimizes the confidence interval of the estimates.

This means that this confidence interval has to be computed, but this is another subject which cannot be dealt with inside this paper.

What is the result of this computation ? Starting with the same data points we first compute the CER given by the median : figure 4.

Then we compute the CER given by the OLS : figure 5.

The improvement between the results is obvious : look more specifically at the low cost figures. In order to quantify the improvement, we compute the sum S_1 for both : the results are given on figure 3 ;

the first line gives its value for the median, whereas the second line gives it for the OLS. We see here that we improve the "fit" with the data by about 5% which is not negligible !

Absolute residual average	938.165
Absolute residual average (Std. reg.)	981.815

F

Figure 3

It should also be noted that the R^2 has strictly no interest here. It is not difficult to prove that maximizing the R^2 is exactly the same as minimizing the sum of the squares $\sum \varepsilon_i^2$. Consequently the R^2 will always be better with the OLS than with the median ; it cannot therefore be used to compare both approaches and the sums S_1 are the only realistic comparisons.

In order to compare various CERs computed with the median (for instance after removing one data point) another attribute – using the same metric as the median – has to be used. It is not difficult to create one.

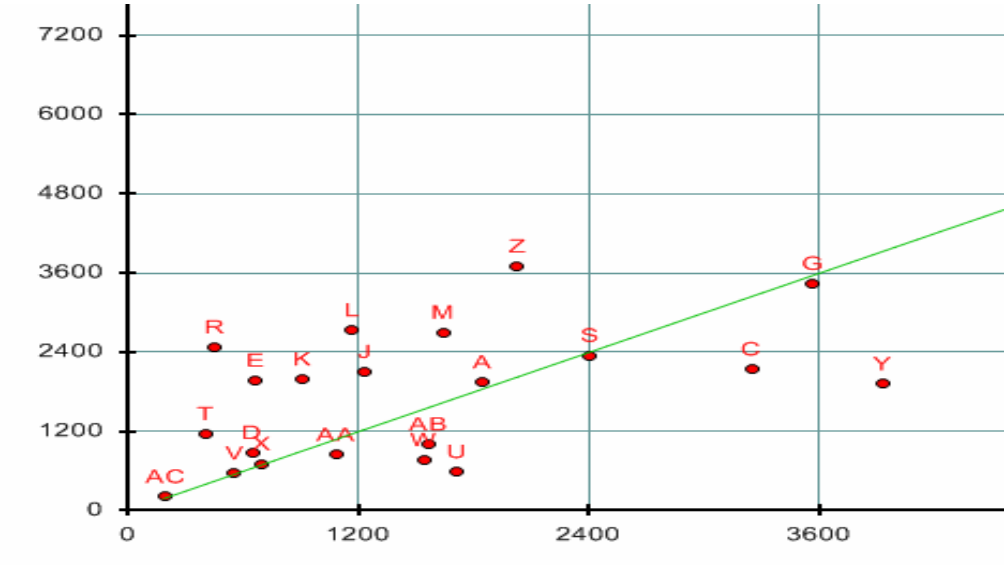

Figure 4. The CER computed by the median

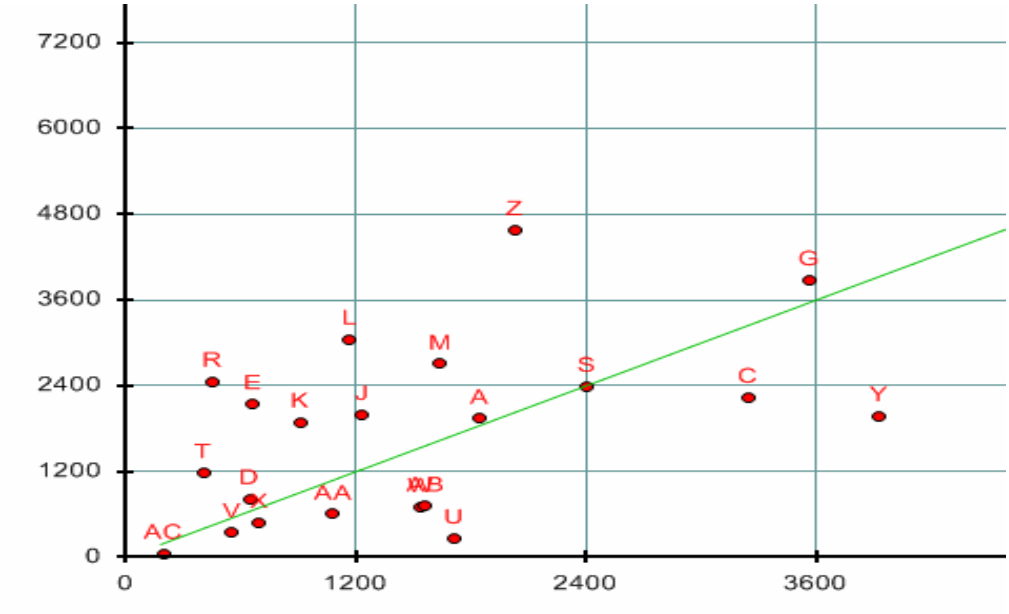

Figure 5. The CER computed by the OLS

As the metric used for the median computations is very robust, the result does not care at all about the outliers : you may have a far away outlier without changing anything. Quite obviously the value of the sum S_1 will change, but the CER will not and the CER will always minimize this sum ; discussing about outliers then becomes obsolete. In order to change the CER, a data point should go over the line ; in such a case the CER is not a median anymore and must be recomputed. But this is a very small drawback compared to the influence of the outliers when using the Gaussian metric !

As there is no longer a matrix to invert, we do not care about using correlated variables. Furthermore there is no "regression" anymore !

Another important comment : it is often said that Gauss demonstrated that his metric was the best … Gauss never said so ! Being a mathematician, Gauss always mentioned his hypothesis. In the present case he said that his metric was the best one if you want the parameters of the CER to be linear functions of the data. Do we really need this hypothesis ? I do not think so …

Quite obviously computing the median by hand would be very time consuming, and this is the reason why Gauss selected his mentioned metric. For this reason, we implemented it in our software : it is as fast as the OLS and … much better !

Pierre Foussier. Paris. March 2010.

Bibliography

1. Karl Friedrich Gauss. *Theoria Motus Corporum cœlestium*. 1825.

2. Pierre Foussier. *From Product Description to Cost. A practical approach*. Springer 2006.

3. Pierrine and Pierre Foussier. *Should we use the median instead of the OLS* ? Parametric World. Fall 2008 Vol 27. No. 4

4. G. Saporta. *Probabilités*, *Analyse des données et Statistique*. Editions Technip. 1990.