Improving CERs building

Getting Rid of the R^2 tyranny

Pierre Foussier pmf@3f-fr.com

ISPA. San Diego. June 2010

Why abandon the OLS ?

- The ordinary least squares (OLS) aims to build a CER by minimizing the sum of the **squares** of the « residuals ».
- It is (with occasionally some added developments) the only procedure which is taught for true « parametric cost estimating ».
- The OLS has however several weak points which are certainly well known by most models builders.

About the OLS weaknesses

- ❶ As soon as data are a bit scattered, the obtained CER is biased
  - this means that the CER underestimates large products and overestimates small products.
  - the name of “regression analysis”, frequently given to this procedure comes from this point : the CER “regresses” towards the average cost !
  - the more the data are scattered (and this is very frequent in the cost domain …), the larger the bias.

About the OLS weaknesses

- ❷ Outliers can seriously damage a CER.
  - this comes from the fact that the "metric" used (I will return to that) is not robust.
  - it means that one of the first tasks of the model builder is to decide which is an outlier and which isn't. This is time consuming, sometimes not obvious (especially when seven variables are considered), and may have an influence on the CER one builds.

About the OLS weaknesses

- ❸ It is also dangerous to use correlated variables (this is due to the mathematical procedures that the OLS needs) :
  - the form of the CER may not be accepted (even if it is correct)
  - more important : the quality of the CER coefficients (their “t”) will be low. This means that the confidence intervals of the estimates will be large
  - N.B. An answer can be found with the Ridge regression.

About the OLS weaknesses

- ❹ In the cost domain, the OLS favors the data points with high cost and does not really care about the low cost items :
  - this comes from the fact that, in our domain, the precision of our data points is a fix percentage of the cost (for instance 5%),
  - consequently high costs have a lower "absolute" precision,
  - as the OLS uses the squares deviations, it favors high costs to the detriment of low costs.

- Is there any theoretical incentive to minimize the **squares** of the deviations ?
- The answer is very clear : **NO**.
- Gauss, who created the method, just said : it is easier … which was true at his time ! Nevertheless he mentioned Laplace’s solution (who did not use the OLS) saying his own solution was nevertheless simpler.

Let’s start from scratch …

- “Parametric cost estimating” is built on two ideas :

① the model builder **decides** what the form of the relationship (between the “dependant” variable and the “causal” ones) will be.
A simple example : the straight line.

② then he/she has to **adjust** it to the data.

To do so …

① he/she adds in the relationship **a few auxiliary variables** (the so-called “parameters”, and the word “parametric” originates from these variables).

② he/she has then to define **the distance** between the data and the relationship.

③ then the distances are added.

And the values of the parameters are then selected to minimize this cumulative distance …

About the “distance”

- The choice of a metric is an extremely important step ; this metric must follow a set of properties (which I won’t discuss here).
- Three different metrics are available
  - the **Gaussian** distance (distance = deviation^2) or OLS
  - the **Euclidian** distance (for two variables only)
  - the **Laplacian** distance (distance = |deviation|) which we will now explore.

A word about the OLS

- In a distribution, the “center” of the distribution is given by **the mean** : it is not difficult to demonstrate that the mean $\overline{y}$ minimizes the sum

$$\sum_i \left(\overline{y} - y_i\right)^2$$

- Everybody knows, after using the mean, that this metric is **not robust at all** (this explains the problem of the outliers …).

The median ?

- **For a distribution**
  - first definition : the median is the value $\tilde{y}$ which leaves as many data points above as below it.
  - second definition : the median minimizes the sum

$$\sum_i \left|\tilde{y} - y_i\right|$$

- We demonstrated that these two definitions are equivalent. It is well known that the median is **very robust**.

Multi-dimensional ?

- Can we expand this idea of the “median” to a multi-dimensional space (we need it to establish a CER) ?
- The answer is **YES**, using the second definition of the median (with the help of the first one !).
- As you may expect, the mathematics used differ completely from the ones used for OLS (playing with “absolute values” is not so easy).

Multi-dimensional ?

- We succeeded to solve the mathematics ; there is no analytical solution to find the parameters, but the CER is expressed the same way (as a formula).
- Several solutions are sometimes possible, giving very close values for the sum of absolute deviations. In such a case, we **select** the one which presents the smallest confidence interval for the estimates.

Consequence of the metric

- Gauss demonstrated that using his distance implies that the probability of a deviation $x-\mu$ is necessarily given by the distribution

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

which is the “Gaussian” distribution (with a standard deviation σ given by $\sqrt{\frac{1}{n}\sum_i(y_i-\mu)^2}$).

Consequence of the metric

- Using the Laplacian distribution implies that the distribution of the deviations is now given by

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

with median ν and τ its measure of dispersion given by

$$\tau = \frac{1}{n}\sum_i |y_i - \nu|$$

- These distributions are obviously consistent with the selected metrics.

Consequence of the metric

- Both distributions are similar, the second one being “sharper” (for the figure we selected $\mu = \nu = 0$ and $\sigma = \tau = 1$).

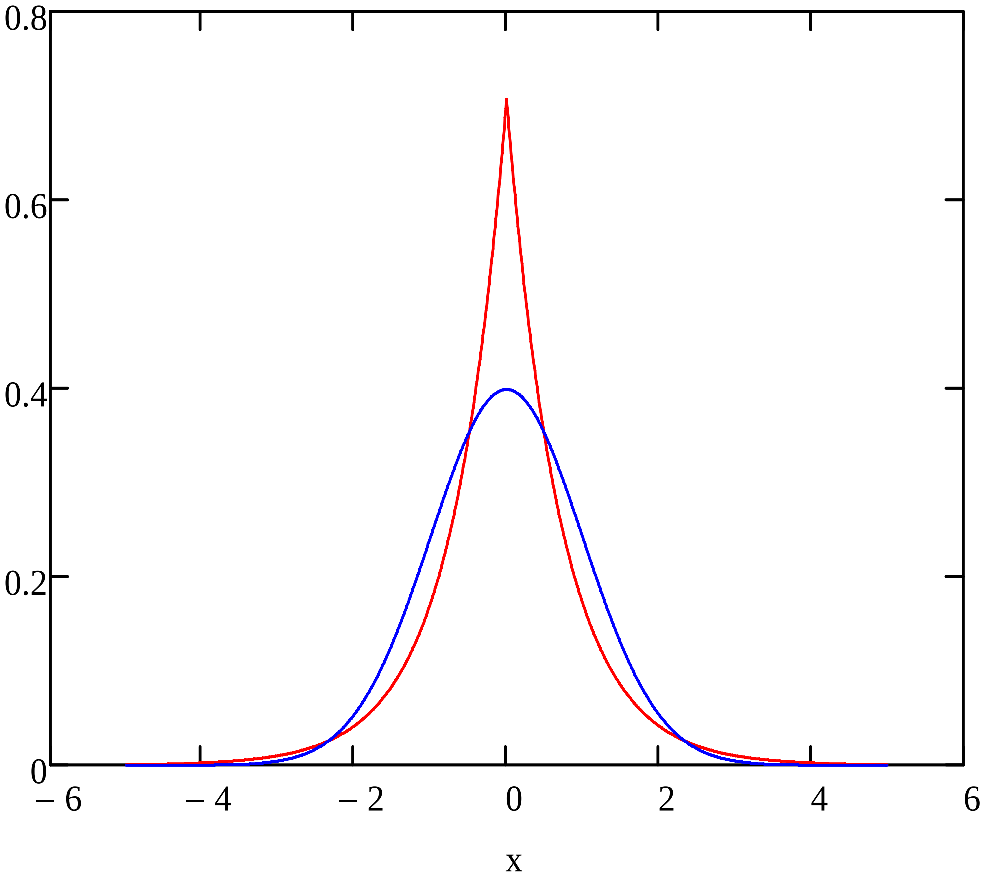
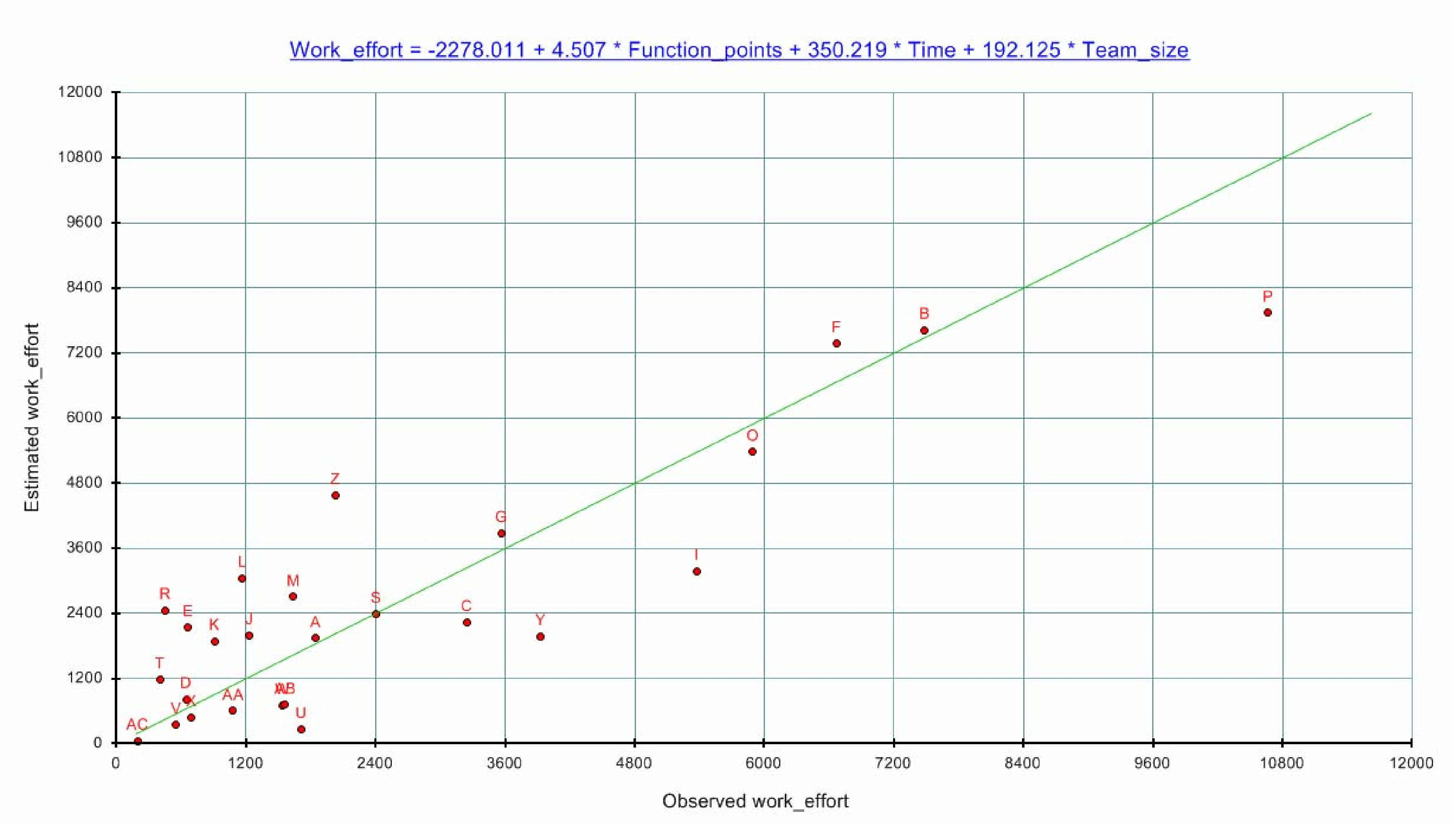
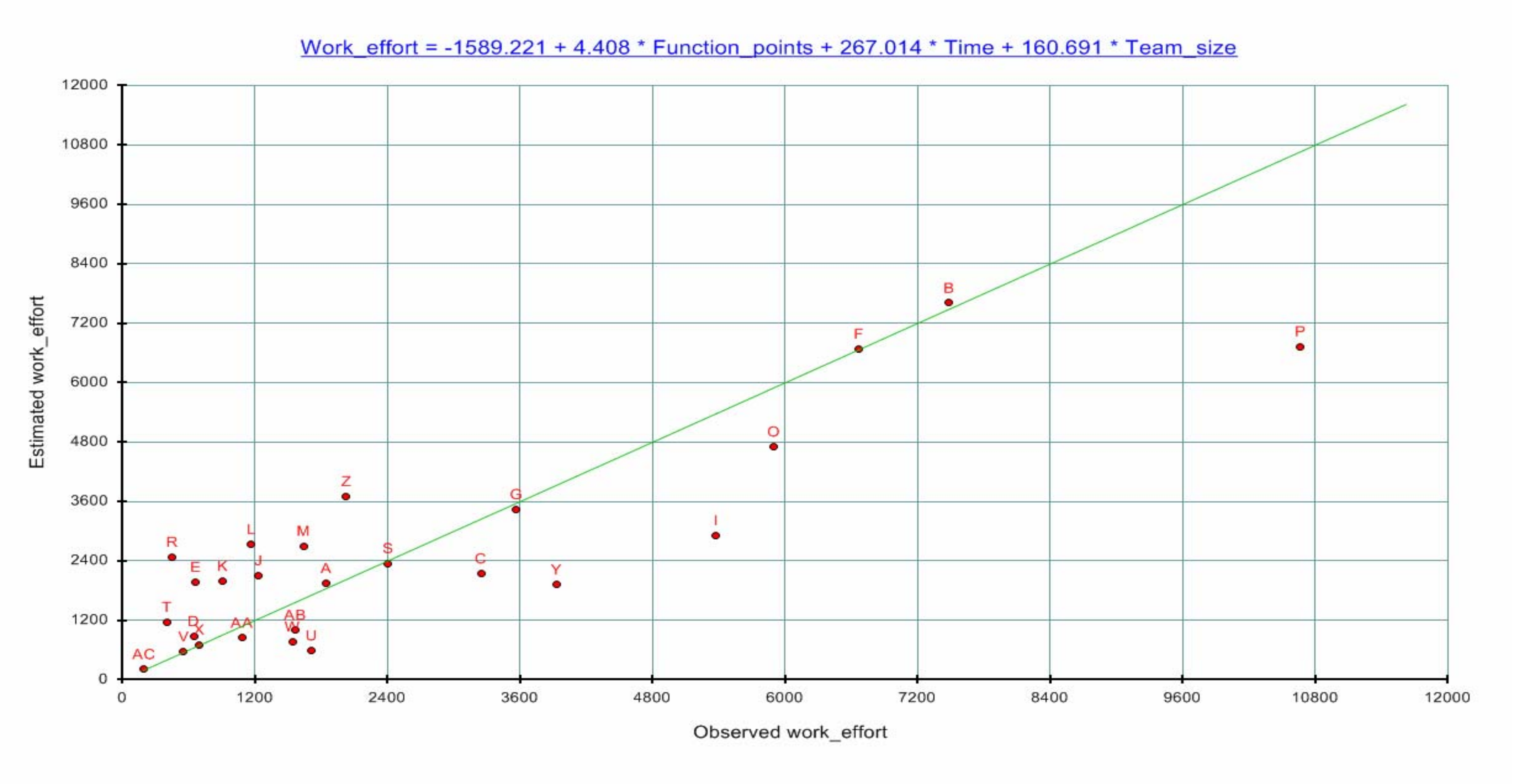


Illustration with OLS (R² = 0.759)



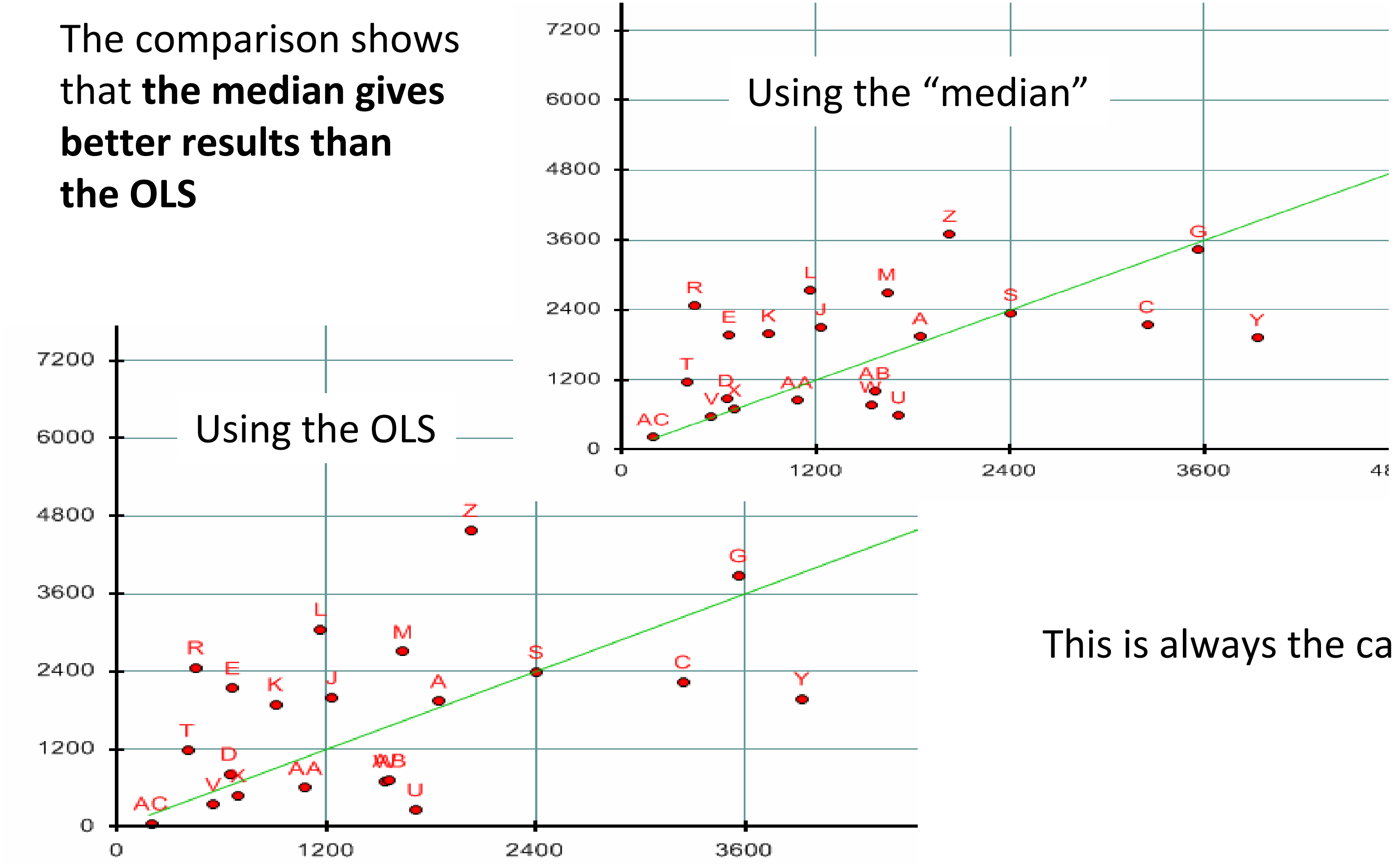
Work_effort = -2278.011 + 4.507 * Function_points + 350.219 * Time + 192.125 * Team_size

Illustration with the "median"


Work_effort = -1589.221 + 4.408 * Function_points + 267.014 * Time + 160.691 * Team_size

Comparison

The comparison shows that **the median gives better results than the OLS**

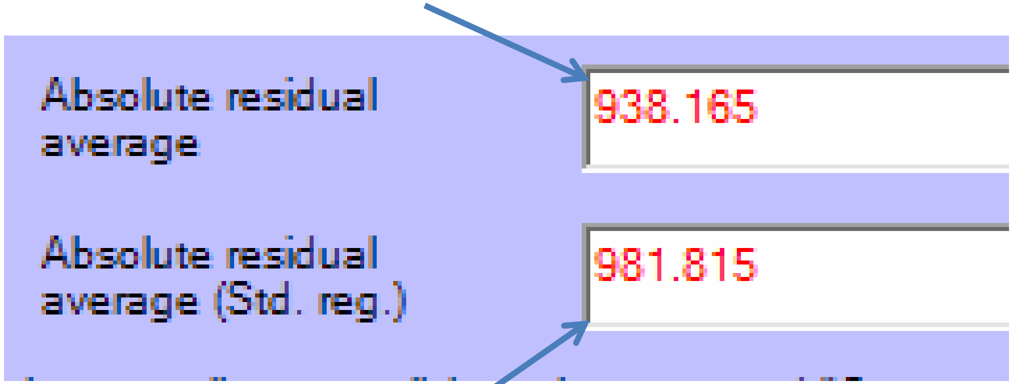


This is always the case

Comparison

- To quantify the comparison, we compute, for both algorithms, the sum of the (absolute) values of the residuals. Here we win **5%** !

computed with the median



Absolute residual average 938.165

Absolute residual average (Std. reg.) 981.815

computed with the OLS

The CER computed with the median is always closer to the data than the one computed with the OLS

What about the R^2 ?

- The CER computed with the median always gives a R^2 lower than the one computed by the OLS.
- This is natural because using the OLS or maximizing the R^2 means exactly the same thing.
- **The R^2 is completely irrelevant here !**

A frequent objection

- I hear sometimes that Gauss proved that the OLS was the best solution to prepare a CER.
- This is not true : he proved that it is the best solution **if** you want the computed parameters to be linear functions of the data.
- This theorem does not apply when the computed parameters are not linear functions of the data.

About the weaknesses of OLS

With the median

- ❶ as there is no “regression”, the CER has no bias.
- ❷ there is obviously no outlier.
- ❸ there is no danger about using correlated variables.
- ❹ the median does not favor high costs.
- The median is much more suited to our needs than the OLS !

Conclusion

- Are you really interested in minimizing the **squares** of the deviations ?

- I guess NO ! Then the “median” is the technique, which is nowadays usable thanks to the computer, to efficiently and safely prepare your CERs.

Thank you for your attention