

# Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE

---

*Source:* <http://sci.tech-archive.net/Archive/sci.stat.math/2006-01/msg00037.html>

---

- *From:* "Reef Fish" <Large\_Nassau\_Grouper@xxxxxxxx>
  - *Date:* 6 Jan 2006 19:45:55 -0800
- 

G Robin Edwards wrote:

> In article <1136440910.097064.23770@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx>,  
> Reef Fish <Large\_Nassau\_Grouper@xxxxxxxx> wrote:  
>  
>> Richard Ulrich's tedious rehash of his blunder was explained in  
>  
>> my January 5 post: <http://tinyurl.com/9bzvu>  
>  
>> The essence of Ulrich's blunder can be summarized here:  
>  
>> ===== excerpt  
>> The independent variable X in a multiple regression CAN be anyone  
>> of these:  
>  
>> 1. An indicator variable with values 0 or 1.  
>  
>> 2. A discrete uniform distribution of ranks.  
>  
>> 3. A distribution that came from Cauchy or other long tail  
>> distributions that would appear to have outliers (compared  
>> to "normal")  
>  
>> 4. The distribution of an observed X can be severely bimodal,  
>> trimodal, left skewed or right skewed ... and it short any  
>> distribution that has ever been observed in the entire history  
>> of statistical distributions that are NOR NORMAL, can be  
>> the distribution of the X used in any multiple regression.  
>  
>> So, why was sehwail and Richard Ulrich want to check the "normality"  
>> or "outliers" of the the data distributions in the INDEPENDENT  
>> variables X?  
>> ===== end excerpt  
>  
>> RU> Normality is only one sort  
>> RU> of baseline; if that wasn't clear immediately, which I thought  
>> RU> it would have been, I made it clear later.  
>

Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE

- >> "Normality" is NEVER a part of the baseline for the independent v
- >> ariables X!
- >
- >> -- Reef Fish Bob.
- >
- > Now I'm only a long retired amateur, but here's what I used to do:--
- >
- > On first receiving data, check visually for absurdities if possible.

Robin, I remember your participation in the re-analysis of the regression data set in the 1975 SPSS Manual.

You were so honest in your individual attempt that you missed the BIG obvious data RECORDING (typo) errors that was missed also by SPSS, while that was one of the "lessons" intended -- to look for obvious keypunch or typing errors.

- > The first computing operation was to generate Box and Whisker diagrams
- > for all variables in the set, with the scales arranged independently so
- > that each BW plot occupies the full screen width. This gives an
- > immediate oversight of the data, showing up any possibly "ridiculous"
- > values and of course "strange" ones if one was anticipating roughly
- > normal data.

If the data is to be examined entirely from an "exploratory" view, those may well be a part of the routine exploration. But if someone has already decided to do a fitting of a particular Y on several X that had already been chosen for the regression task, then your paragraph below is more to the point:

- > However, the points from a designed experiment, such as
- > an RCC design are certainly not going to be normally distributed, which
- > is totally immaterial as far as regression techniques are concerned.

- >
- > The next step was to make the entire set of PLOTS of all the variables
- > two at a time, eg with 6 variables (dependent and independent) there
- > would be 15 plots.

Again, that depends on whether it's an unfocused "exploratory" analysis, or a much more focus multiple regression task of fitting a high dimensional surface.

In the latter case, the "scatter matrix" is what your pairwise scatterplots

Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE

is called, and it is not very useful toward discovering the fitting surface in HIGHER dimensions, just as two-dimensional scatterplots do not reveal the relation in three dimensions.

> It is useful to have them all on one page (or  
> screen). This technique highlights possible "bivariate strangeness"  
> and can provide useful information for the data detective.

But not necessarily for one performing a pre-specified regression with specified variables.

>  
> "Absurd" data highlighted by these preliminaries can probably be  
> discarded but only after consulting the owner of the data.

No, there hasn't been any indication why your "absurd" data should be discarded at all! The owner of the data may be the least qualified person to know that it is a crime to callously throw away data because they appeared unusual to their untrained eye.

>  
> Only then go ahead with fitting the hypothesised model, by linear  
> regression methods. < snip >

This is where the iterative "model building" approach in George Box's JASA paper on "Science and Statistics" takes over.

> Perhaps these days (>20 years on) these notions have been superseded,  
> but they used to work quite well for me :-)) , or so I fondly believed.

I wouldn't put it your way. What used to work well (> 30 years ago -- as when I started teaching them; and even earlier by those before my time) still work well today, and none any better, except for some minor advances in graphical techniques.

The problem is that many folks never knew or learned what worked well and how the model-building process is supposed to work, as is evident from much that has been posted in these sci.stat groups.

-- Bob.

.

---

• *References:*

◆ *Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE*

◇ *From:* G Robin Edwards

Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE

- Prev by Date: [\*Re: Finding Asymptotes from a set of data\*](#)
- Next by Date: [\*Re: 4 nines = 100\*](#)
- Previous by thread: [\*Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE\*](#)
- Next by thread: [\*Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE\*](#)
- Index(es):
  - ◆ [\*Date\*](#)
  - ◆ [\*Thread\*](#)