

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

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2006-01/msg00026.html>

- *From:* G Robin Edwards <robin.edwards@xxxxxxxxxxxxx>
 - *Date:* Thu, 05 Jan 2006 23:27:01 +0000 (GMT)
-

In article <1136440910.097064.23770@xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx>, Reef Fish <Large_Nassau_Grouper@xxxxxxxxx> wrote:

> Richard Ulrich wrote:

>>> Richard, when are you going to LEARN the most BASIC
>>> material in regression? You comment about checking the
>>> independent variables was PRECISELY the result of your
>>> muddle about the standard regression assumptions, what
>>> can be checked and what is ABSOLUTELY UNCESSARILY
>>> to check -- the independent variables X, for outliers or
>>> anything else other than blatant typo errors.

> RU> I've posted elsewhere, giving in full my post of Nov. 7,
> RU> which says LOOK at the data. 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.

> 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,

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

- > trimodal, left skewed or right skewed ... and it short any
- > distribution that has ever seen 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 sehwait and Richard Ulrich want to check the "normality"
- > or "outliers" of the the data distributions in the INDEPENDENT
- > variables X?
- > ===== end excerpt

- > If any and all of those X's are perfectly valid data for the
- > independent
- > variables in a regression problem, why would anyone except those
- > seriously muddled, like Richard Ulrich would make such NONSENSE
- > statements:

- > 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.

- > "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.

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. 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. 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.

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

Only then go ahead with fitting the hypothesised model, by linear regression methods. Be sure to compute the standard regression diagnostics and pay attention to points that produce the largest diagonals of the Hat matrix. They might just be important! Re-check

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

the original data, then report to the owner. I also used to compute jackknife residuals for my own satisfaction, but found them tricky to explain to the typical "bench scientist"! They were more likely to be intrigued by VIFs, which help highlight (multi)collinearity.

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.

Robin

• ***Follow-Ups:***

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

◇ *From:* Reef Fish

• Prev by Date: ***Re: A problem for complete beginners (1/2006)***

• Next by Date: ***To simulate Bivariate Normal samples? Easy.***

• Previous by thread: ***A problem for complete beginners (1/2006)***

• Next by thread: ***Re: Rich Ulrich continues his statistical Muddle, Quackery, and MALPRACTICE***

• Index(es):

◆ ***Date***

◆ ***Thread***