Re: linear regression results approximates to the mean of Y

Source: http://sci.tech−archive.net/Archive/sci.stat.math/2008−06/msg00207.html

- *From*: Richard Ulrich <Rich.Ulrich@xxxxxxxxxxx>
- *Date*: Mon, 09 Jun 2008 19:36:51 −0400

On Fri, 6 Jun 2008 18:49:55 −0700 (PDT), feilian <bslnt@xxxxxxx>
wrote:

> Thank you!
>
> I am not a researcher in mathematics and statistics. In my work,
> Linear Regression is used to estimate human age when given a face
> image.
> That is, the "X" is features extracted from face image , Y is the
> corresponding age.
>
> I am not very sure if there's linear realtion between the X and Y. So
> both linear and quadratic
> regression is test and found that linear is better.
> I have 25 images for each age, and the age range is from 20 to 60.
> When all images are used the result is pleasing, the mean estimated
> age of images at each age exhibits linearity: (in the image below the
> read line is the real age,the blue line is the mean estimated ages:
> http://hiphotos.baidu.com/bslnt/pic/item/1b85bf587ad439ca9d8204d9.jpg
>
> but when images in a smaller age range is used , the means trends to
> near the mean of the age range.(in this image, images aged from 30 to
> 50 is used to train the "local linear regression", its means at each
> age is represented by the blue line, red line is still the real age
> and green is that of the global regression when all images are used.
> http://hi.baidu.com/bslnt/album/item/da168c44dedad396b3b7dcaa.html
>
> I am not very sure what does this result mean. Could it be possible
> that there is no local linearity? Or it means the real distrubution is
> similar to the line below?

I think it is entirely the result of artifact, from reducing
the variance of the outcome.

Whenever you take the correlation over a *restricted* range
of a variable, you get a lower correlation. That is why it is

important to remember that a correlation is a feature of a
sample, in addition to saying something about the two variables.
Some textbooks will include a formula for correcting for
truncation or reduced variance. In this case, you have cut the
range in half, and presumable have cut the variance by a
factor of 4.

This reduces the correlation. Consider the two possible
regressions, X on Y and Y on X, the truncation of one of them
also reduces *one* of the two regression coefficients. I think
that this is all you are looking at. I can't read the language,
so I don't know for sure what is being plotted.

> http://hi.baidu.com/bslnt/album/item/823733178d8632154a90a780.html
>
> And I also want to know how to confirm the linear realationship
> between two variables?

Plotting the means of deciles (using sets that each have
10 percent of the sample) is usually pretty good evidence.

You do not find a really good test without having a prior notion
of what the "nonlinearity" consists of. Looking at polynomial
fits would probably be the default 'variation', or looking
for "basement" or "ceiling" effects at the extremes.

––
Rich Ulrich
http://www.pitt.edu/~wpilib/index.html
.