A note on repeated measures analysis and conditioning on baseline measurement

SDC

September 2013
Version 1.0

Compiled Wednesday 25th September, 2013, 14:48
from: C:/Bendix/Artikler/rep-meas/rm.tex

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://BendixCarstensen.com

Contents

Relationship between approaches

We consider outcome data for individuals i, measured at times 1 and 2, where time 1 is pre-randomization and 2 is post-treatment, so a full model would be:

$$\begin{aligned} y_{it} =& \mu + \beta_t + \delta_g + \gamma_{gt} + \eta + a_i + e_{it}, \quad i = 1, \ldots, I, \quad t = 1, 2 \\ & a_i \sim \mathcal{N}(0, \tau^2), \\ & e_{it} \sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{1}$$

where η represents the effect of possible confounders to be included in the model. For convenience we assume $\beta_1 = \gamma_{1g} = 0$.

Note that in the model (1), we allow different baseline means between the randomization groups $g = g(i)$, in the parameter δ_g. In a randomized study one would expect that $\delta_g = 0$, hence a model without δ_g would be worth considering too.

The usual approach to analysis of repeated measures with a baseline and one follow-up measurement is to use the baseline as covariate [1]. This is a corollary of the basic statistical principle that inference should be made in the conditional distribution given the sufficient statistics for the ancillary parameters, which in this case is the individual-specific values for each person (a_i). The baseline measurement y_{i1} is not the sufficient statistic for this, but it is close and easier to handle.

The model is a 2-dimensional normal distribution of the measurements y_1 and y_2:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

From standard statistical theory we know that under this model, the conditional distribution of y_2 given y_1 is:

$$y_2 | y_1 \sim \mathcal{N}\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(y_1 - \mu_1), \sigma_2^2(1 - \rho^2) \right)$$

Now in the model (1) we have the following values for the parameters μ_1, μ_2, σ_1^2, σ_2^2 and ρ in the 2-dimensional normal model outlined above:

$$\begin{aligned} \mu_1 &= \mu + \delta_g + \eta \\ \mu_2 &= \mu + \delta_g + \beta_2 + \gamma_{g2} + \eta \\ \sigma_1^2 = \sigma_2^2 &= \tau^2 + \sigma^2 \\ \rho &= \frac{\tau^2}{\sigma^2 + \tau^2} \end{aligned}$$

Using this in the formulae for the conditional distribution, gives the conditional distribution of y_2 given y_1 in terms of the model parameters from (1) (well, we maintain ρ):

$$\begin{aligned} y_2 | y_1 &\sim \mathcal{N}\left(\mu + \delta_g + \beta_2 + \gamma_{g2} + \eta + \rho\big(y_1 - (\mu + \delta_g + \eta)\big), (\sigma^2 + \tau^2)(1 - \rho^2) \right) \\ &= \mathcal{N}\left(\big((1-\rho)\mu + \beta_2\big) + \big((1-\rho)\delta_g + \gamma_{g2}\big) + \rho y_1 + (1-\rho)\eta, (\sigma^2 + \tau^2)(1 - \rho^2) \right) \end{aligned}$$

Hence, when fitting the conditional model to data generated by model (1):

- the term $(1-\rho)\mu + \beta_2$ should show up as the intercept,
- the term $(1-\rho)\delta_g + \gamma_{g2}$ as the coefficient to the treatment indicator,
- ρ as the coefficient to the baseline measurement y_1,
- the coefficients to the confounders should appear scaled by $1-\rho$. and
- the residual standard deviation should be $\sqrt{(\sigma^2+\tau^2)(1-\rho^2)}$.

In any practical circumstances, when fitting the two different models (the random effects model and the conditional model) we will find these relationships quite accurately.

Thus it seems that when conditioning on the first measurement y_{i1}, we are implicitly assuming that $\delta_g = 0$ if we interpret the coefficient to the treatment indicator as the treatment effect.

If we want to allow for baseline imbalance, we must fit the random effects model with δ_g. Fitting the random-effects model without δ_g will give a result for the treatment effect γ_{g2} similar to the model conditioning on y_1.

Reporting effects

Usually the treatment effect is reported as the coefficient to the treatment indicator from an analysis with y_1 as covariate.

However very often researchers also wants to report the change in each randomization group separately, and that is usually done by just computing the mean change in each group with the corresponding empirical standard deviation. These are estimates from a model for the changes, allowing for different variance in each group, hence certainly not estimates from any model normally considered.

It could be argued that given that you want to report the within-group changes, it would be more reasonable to fit the random effects model (with or without δ), and report the quantities $\beta_2 + \gamma_{g2}$. Because of the obvious redundancy, these group-specific changes have a difference corresponding to the claimed treatment effect, which would seem an obvious advantage[1].

Thus the random effects model gives the possibility to model baseline imbalance and sensibly report changes observed within groups; the model conditioning on the baseline does not. So the conditional model (ANCOVA) was an excellent practical approach before the advent of software for random effects models. Not so any more.

References

[1] Andrew J Vickers and Douglas G Altman. Analysing controlled trials with baseline and follow-up measurements. *British Medical Journal*, 323:1123–1124, November 2001.

[1] Unless of course you subsribe to the notion that the amount of information obtained is proportional to the number of different models fitted to a given dataset.