BASIC ANALYSIS FOR TRIAL DATA PART FIVE

DAN W. JOYCE

Analysis using ANCOVA

We conclude by examining the analysis method used in (Kane et al., 1988). An *analysis of covariance* – or ANCOVA – takes a slightly different approach to formulating the linear model and requires that the data be organised differently. So far, our data has had one row for each patient, at each time point (baseline and post-treatment). To perform the ANCOVA, we need the data formatted differently, and `Kane-simulated-ANCOVA.csv` contains the same simulated data in a suitable format, a sample of which is reproduced in Table 1.

i (Patient)	Y_{ij} (BPRS post-treatment)	X_{ij} (BPRS pre-treatment)	D_{ij} (Drug)
261	44	50	0
206	48	53	0
138	59	63	0
179	62	66	0
255	65	69	0
39	46	60	1
212	47	53	0
119	69	86	1
189	53	57	0
216	51	57	0

Table 1: Sample of data table arranged for ANCOVA

The data is now organised as follows:

- i still refers to (indexes) the individual patients
- Y_{ij} is the BPRS score at week 6 (post-treatment, corresponding to $T_{it} = 1$) – notice how we have also dropped the Time column (T_{it}) because we know that all values in this column represent *only* post-treatment BPRS scores. This corresponds to extracting Y_{ijt} for each patient where $T_{it} = 1$ in the original table of data (from Part Two)
- D_{ij} as before, refers to the medication group assignment
- X_{ij} is the BPRS score for patient i on medication j *pre-treatment* – which can be obtained simply by extracting Y_{ijt} for each patient where $T_{it} = 0$ in the original table of data (from Part Two)

With our data in this format, the formulation of an ANCOVA (again, as a linear model) is:

$$Y_{ij} = \beta_0 + \beta_1 D_{ij} + \beta_2 X_{ij} + \epsilon_{ij} \tag{1}$$

The important difference between our previous model $Y_{ijt} = \beta_0 + \beta_1 D_{ij} + \beta_2 T_{it} + \beta_3 D_{ij} T_{it} + \epsilon_{ijt}$ and that described by equation 1 is that we have the *covariate* X_{ij} – the BPRS score before treatment. The interpretation of the terms and βs in equation 1 are subtly different:

A **covariate** is a variable that may predict the outcome and is treated similarly to any other predictor variable. This is emphasised by the fact that – like Time and Drug in our previous example – it is just another term in the sum on the right-hand side of the equation for a linear model

- β_0 is still an intercept
- β_1 captures the effect of medication on post-treatment BPRS (Y_{ij}) but is *adjusted* for the difference in group in pre-treatment BPRS (X_{ij})

Importantly, β_2 captures the effect of pre-treatment BPRS on the post-treatment BPRS (Y_{ij}) and the way to understand it's role is to imagine the case where $\beta_2 = 1$ in equation 1 such that

This part can be skipped if the equivalence of all of the linear models discussed is of less interest

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 D_{ij} + X_{ij} + \epsilon_{ij} \\ Y_{ij} - X_{ij} &= \beta_0 + \beta_1 D_{ij} + \epsilon_{ij} \end{aligned} \tag{2}$$

Notice that now, on the left-hand side, we have the *difference* between the post-treatment BPRS and the pre-treatment BPRS with a right-hand side that simply models an intercept with an additional contribution provided by the effect of drug treatment (and an error term, ϵ_{ij}). So an ANCOVA turns out to be the same as a linear model of a dependent variable – the change in BPRS (from pre- to post-treatment) – with independent variables of an intercept and effect of drug.

We prefer ANCOVA when we are sure the design is robustly randomised, with no pre-treatment group difference because it implicitly assumes that there is *no difference* between the chlorpromazine and clozapine group's BPRS scores before treatment, which is a fair assumption given the randomised design implemented in (Kane et al., 1988). If we find that patients assigned to each drug pre-treatment *did* differ on baseline BPRS, then the ANCOVA version of the model would fail to capture this effect and would result in a biased model.

Interpreting the ANCOVA

After fitting the model in equation 1, our statistics package gives the output shown in Table 2

Output	Term	Coefficient	Estimated Beta	95% Conf. Int.
(Intercept)	–	β_0	-4.65^{***}	$[-5.95; -3.35]$
Drug	D_{ij}	β_1	-10.89^{***}	$[-11.30; -10.48]$
BPRS (baseline)	X_{ij}	β_2	0.99^{***}	$[0.97; 1.01]$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Estimated ANCOVA model

We will not spend time methodically reporting these results, because it follows the same approach as for the repeated measures

ANOVA we discussed earlier. Of particular note, to test the effect of medication on the post-treatment BPRS score (ultimately, this is what we are interested in), we only have to inspect the Drug term D_{ij} and the corresponding coefficient β_1. Again, we set up our null hypothesis that there is no effect of drug and β_1 is zero $H_0 : \beta_1 = 0$. And we will make a decision to *reject* the null hypothesis H_0 in favour of H_1 according to:

$$\text{Reject } H_0 \text{ if } \beta_1 \neq 0 \text{ with } p < 0.05 \tag{3}$$

Inspecting the ANCOVA results in Table 2 we find that the mean post-treatment BPRS for clozapine (i.e. when $D_{ij} = 1$) is -10.89 with a 95% confidence interval that tells us the effect could be as much as -11.30 and as little as -10.48. The 95% confidence interval does not include the null hypothesis ($\beta_1 = 0$), and the p-value is < 0.05. We therefore reject the null hypothesis of no effect of Drug on BPRS.

We note that the estimated effect of drug is almost identical to that obtained with the repeated measures ANOVA (from Parts Three and Four) but the ANCOVA model provides narrower confidence intervals.

This results from how β_2 is computed in the ANCOVA formulation – beyond the scope of our discussion – but suffice to say, when ANCOVA is appropriate (i.e. randomisation was robust) it provides more *power* and results in smaller confidence intervals

ANOVA of Change

To complete our discussion of different approaches to modelling, we consider how to use a simple ANOVA on a derived dependent variable, the *change* in BPRS score between pre- and post-treatment. If we are only interested in the change in BPRS, then instead of having to incorporate two variables (pre- and post-treatment) why not simply derive one single variable representing the change over the course of treatment. Further, if we think about the covariate X_{ij} from the ANCOVA formulation – we included this to model how the individual patient's pre-treatment BPRS influences their post-treatment BPRS score. Defining change in symptoms is slightly more complex and subtle; see (Leucht et al., 2009) for details. However, the principles of analysis remain similar. To model the *change* in BPRS score we proceed by calculating, for each patient i, the difference in post-treatment ($T_{i1} = 1$) and pre-treatment ($T_{i0} = 0$) BPRS:

$$\Delta Y_i = (\text{BPRS at } T_{i1}) - (\text{BPRS at } T_{i0})$$

Which in terms of the columns and terms in Table 1 is:

$$\Delta Y_i = Y_{ij} - X_{ij}$$

To make this concrete, take patient 261 (the first row) from Table 1. They were assigned to chlorpromazine ($D_{ij} = 0$), and had a pre-treatment BPRS $X_{ij} = 50$ and a post-treatment BPRS $Y_{ij} = 44$. Then, the new derived change in BPRS value for this patient is:

$$\Delta Y_{261} = 44 - 50 = -6$$

So, patient 261 improves by 6 points – their post-treatment BPRS is 6 points *lower* than their pre-treatment BPRS. Compared to the data from our previous examples, our data will now appear as shown in Table 3. To analyse this data yourself, load the file `Kane-simulated-ANOVA-change.csv`.

i (Patient)	ΔY_i (Change in BPRS post- from pre-treatment)	D_{ij} (Drug)
229	-8	0
68	-19	1
235	-6	0
9	-16	1
80	-16	1
45	-19	1
227	-5	0
238	-6	0
70	-17	1
1	-18	1

Table 3: Sample of data table arranged for ANOVA of change in BPRS

The model we will fit to this data is now very simple:

$$\Delta Y_i = \beta_0 + \beta_1 D_{ij} + \epsilon_i \tag{4}$$

The output from our statistics package gives us Table 4 and the interpretation proceeds as for our previous examples. We want to see if there is an effect of Drug on outcome, which in the ANOVA of change model, is to simply to inspect the estimated coefficient $\beta_1 = -10.90$ with a 95% confidence interval of $[-11.31, -10.49]$. Note the similarity to that obtained from the ANCOVA in Table 2 for the Drug term, D_{ij}, and it's coefficient β_1) as well as (from Part Three) the repeated measures ANOVA where the same effect was given by the interaction term Drug:Time, $D_{ij}T_{it}$, with it's coefficient β_3. The hypothesis testing for the ANOVA of change proceeds in the same way as for the other models, so we will not repeat the statements here.

Output	Term	Coefficient	Estimated Beta	95% Conf. Int.
(Intercept)	–	β_0	-5.03^{***}	$[-5.31; -4.75]$
Drug	D_{ij}	β_1	-10.90^{***}	$[-11.31; -10.49]$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Estimated ANOVA of change model

Conclusion

We have discussed the basics of regression and linear models, first in the continuous-variable case, and then using categorical variables of the kinds you will encounter when comparing treatments. We examined the mechanics of how the components of a linear model behave to describe mean effects of treatments. The take-home message is that – you can use a variety of methods (repeated-measures

ANOVA, ANCOVA and ANOVA of change), and all can be formulated and understood as regression problems (linear models) to be fit with standard statistics software.

You should ensure you are comfortable interpreting the output of a statistics package with reference to the linear model you design to describe the effect of a treatment. This means understanding the hypothesis and its relationship to the estimated coefficients β as well as how to read p values and confidence intervals.

We have not discussed model fitting diagnostics – i.e. deciding if your linear model is really a good description of the data – because our emphasis was on understanding the basic mechanics of linear models applied to clinical trials of treatments.

The most general and flexible model (repeated measures ANOVA) and the ANCOVA and ANOVA of change models were derived from assumptions about baseline differences and how they affect the post-treatment effects. Notably, ANCOVA assumes proper randomisation with no mean difference in BPRS scores before treatment. Kane et al. made this assumption, and using our simulated data, this seems appropriate.

We finish our discussion by repeating advice from (Van Breukelen, 2006) on which of the previous methods to use in different circumstances:

1. if both treatment groups were truly randomised – i.e. patients were equally unwell in each treatment group – then the ANCOVA formulation (equation 1, above) is the most powerful

2. ANCOVA is superior to ANOVA of change for randomised studies

3. if participants drop out (i.e. there is some missing data post-treatment), then the repeated measures ANOVA handles this more robustly

Finally, if you are comfortable with the methods covered here, then you could explore *hierarchical* linear models, which capture difference and change between groups in a more complete way at the expense of the models being somewhat more complex to interpret. A thorough textbook on this topic is Gelman and Hill (2007).

Questions and Exercises

We'll implement the ANCOVA model in SPSS using the *Mixed Models* feature, only this time, we will try to implement equation 1 with similar assumptions – e.g. we have coded the variables so that 0 and 1 are the only allowed numerical values for the categorical Drug and Time. To proceed, you can load `Simulated-kane-ANCOVA.sav` which is identical to `Kane-simulated-ANCOVA.csv` only with the variables (columns) already configured correctly for analysis in SPSS. Note the correspondence of terms in equation 1 with the table

of data in SPSS: the variable named BPRS.T0 is equivalent to X_{ij} in equation 1. Then execute following steps:

1. Select *Analyze*, *Mixed Models* then *Linear*
2. In the dialog box ('Linear Mixed Models: Specify Subjects and Repeated'), add variable 'Patient' to the *Subjects* list and click 'Continue'
3. In the next dialog box ('Linear Mixed Models'), add 'BPRS.T1' – the BPRS score post-treatment – as the *Dependent Variable*, and add both 'BPRS.T0' (pre-treatment, or baseline, BPRS score) and 'Drug' to the *Covariate(s)* list.
4. Click on the *Fixed* button, and then in the dialog, add 'Drug' and 'BPRS.T0' to the *Model* list on the right – ensure that the interaction 'Drug*BPRS.T0' is *not* added by SPSS automatically; if it is, highlight and then hit 'Remove'. Click continue to return to the main dialog box.
5. Now, click *Statistics* and tick the *Parameter estimates* boxes – click continue.
6. Finally, back in 'Linear Mixed Models' dialog, hit *OK* and the model will be estimated

Note, we do not have a Time variable anymore - see discussion earlier in *Analysis using ANCOVA*

As before, we will neglect the diagnostic information. Scroll down to the 'Estimates of Fixed Effects' table, and compare with Table 2.

Next, we'll try implementing the ANOVA of Change model in SPSS : To proceed, you can load `Simulated-kane-ANOVA-change.sav` which is identical to `Kane-simulated-ANOVA-change.csv` only with the variables (columns) already configured correctly for analysis in SPSS. As before, inspecting the data in SPSS, you should see the data corresponds to the sample shown in Table 3 and the model we want to implement is equation 4. Then execute the following steps:

1. Select *Analyze*, *Mixed Models* then *Linear*
2. In the dialog box ('Linear Mixed Models: Specify Subjects and Repeated'), add variable 'Patient' to the *Subjects* list and click 'Continue'
3. In the next dialog box ('Linear Mixed Models'), add 'Delta.BPRS' as the *Dependent Variable*, and add both 'Drug' to the *Covariate(s)* list.
4. Click on the *Fixed* button, and then in the dialog, add 'Drug' to the *Model* list on the right. Click continue to return to the main dialog box.
5. Now, click *Statistics* and tick the *Parameter estimates* boxes – click continue.
6. Finally, back in 'Linear Mixed Models' dialog, hit *OK* and the model will be estimated

Note, we do not have a Time or baseline BPRS (e.g. BPRS.T0) variable anymore - see discussion earlier in *ANOVA of Change*; we have taken care of the effect of Time on BPRS by deriving a single dependent variable – the change in BPRS from pre- to post-treatment – as $\Delta Y_i = Y_{ij} - X_{ij}$ – which is labelled Delta.BPRS in the data table

Scroll down to the 'Estimates of Fixed Effects' table, and compare with Table 4. Notice we are most interested in the effect of Drug (the intercept is less relevant and represents a 'grand mean' which is of little interest).

References

Gelman, A. and Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*. Cambridge University Press: Cambridge, UK.

Kane, J., Honigfeld, G., Singer, J., and Meltzer, H. (1988). Clozapine for the treatment-resistant schizophrenic: A double-blind comparison with chlorpromazine. *Archives of General Psychiatry*, 45(9):789–796.

Leucht, S., Davis, J. M., Engel, R. R., Kissling, W., and Kane, J. M. (2009). Definitions of response and remission in schizophrenia: recommendations for their use and their presentation. *Acta Psychiatrica Scandinavica*, 119(438):7–14.

Van Breukelen, G. J. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of clinical epidemiology*, 59(9):920–925.