BASIC ANALYSIS FOR TRIAL DATA PART TWO

DAN W. JOYCE

Some Initial Results

First, we examine the continuous outcome variable – the change in BPRS scores, reproduced here from (Kane et al., 1988) as Figure 1. Note how from baseline (at week 0) to week 1, the two group's change in BPRS scores quickly diverge, with clozapine appearing to be more effective such that even after 1 week of treatment, the patients on clozapine show an improvement of approximately 5 points over the chlorpromazine group. At 6 weeks, clozapine appears markedly superior. However, a graph alone rarely convinces.

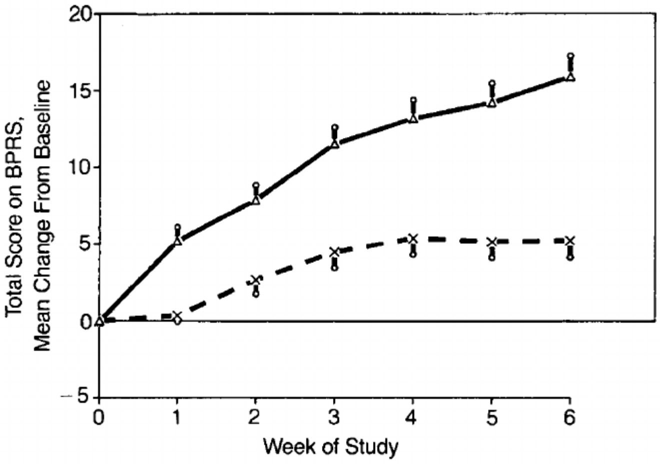

Figure 1: Graph from (Kane et al., 1988) showing change in BPRS over the 6 weeks of treatment. Broken and solid lines show patients assigned to chlorpromazine and clozapine respectively

In their paper, Kane et al. use *analysis of covariance* (ANCOVA) to model their data. To understand this, along with the related *analysis of variance* (ANOVA), it is often useful to frame the model in terms of regression.

There is debate in the statistics literature about whether ANOVA and ANCOVA are really special cases of regression – under the umbrella of **linear models** – but it suffices to note that it is possible to formulate both using regression models (Van Breukelen, 2006)

A Primer on Regression

We begin with a refresher on linear regression. In straight-forward linear regression, we specify that our dependent variable (or outcome) y is related to an independent (predictor) variable x in a linear (straight-line) relationship governed by the simple regression equation (which we call a *model*):

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

It's easier to understand each of the terms in the above equation using a graphical example. In Figure 2, the coefficient β_0 represents the *intercept* of the blue line on the y axis and β_1 is the *gradient* of the blue line. The final term, ϵ, is the *error term* that captures the *variation* in the data that cannot be explained by x in the linear model. This error term essentially captures the 'spread' of the data around the blue regression line, and in most models we require this to be random and normally distributed – that is, the pattern of the data points around the blue line should *not* follow a systematic pattern.

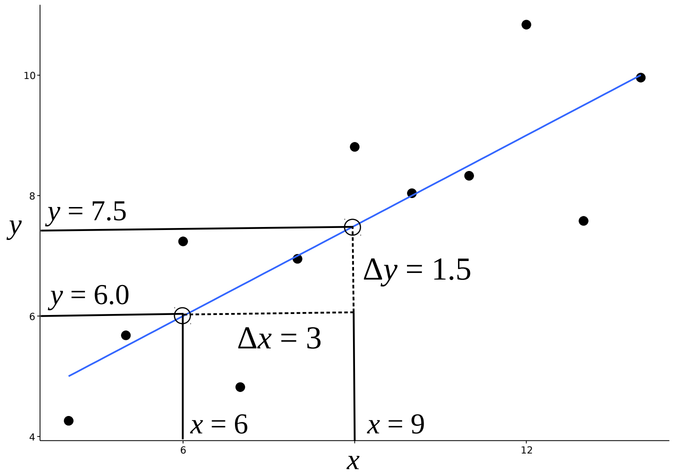

Figure 2: A regression example. Black dots are the data collected from e.g. an experiment. The blue line is the regression equation fitted to the data which exposes the changes in y as x varies

Regression with Continuous Variables

To understand how the coefficient β_1 in equation 1 is estimated, examine Figure 2: if you start at $x = 6$, draw a vertical line up to the blue line, and then horizontally across to the y axis, you arrive at $y = 6$. Do the same at $x = 9$ and you arrive at $y = 7.5$. So, the *change* on the x axis is $\Delta x = 3$, and the corresponding change in the y axis is $\Delta y = 1.5$ resulting in a gradient $\beta_1 = 0.5$. This results in the regression model (ignoring the error term for now):

We use the convention of Δ to denote a **change** in a variable and recall that the gradient is defined as $\Delta y / \Delta x = 1.5/3.0 = 0.5$

$$y = 3.0 + 0.5x \tag{2}$$

The intuition is that, for a one-unit change in x, there is corresponding change of 0.5 units in y, in addition to the baseline (intercept) of 3.0. It should be clear that once we have estimated (fitted) the model given in equation 2 we can then *predict* a value for y given any value of x, even though data for that specific value of x was *not* collected in the original data set. For example, if we want the predicted value of y at $x = 10.5$ (for which there is no corresponding data) we compute y as:

One caveat: while we can predict the expected value of y for a value of x not present in the original data set, we *should not* use values of x that are outside of the original **range** of the data – i.e. we should not expect the regression to **extrapolate**

$$\begin{aligned} y &= 3.0 + 0.5 \times 10.5 \\ y &= 8.25 \end{aligned} \tag{3}$$

Using statistics software to find the β coefficients produces a model of the data that *predicts* the expected outcome (y) in terms of predictor variables (x). Statistics software simply provides a convenient way of doing similar calculations when the model and data are more complex, for example, when we add more terms $(x_1, x_2, \dots)$ and their corresponding coefficients $(\beta_1, \beta_2, \dots)$ when there are more variables needed to explain the changes in y .

We refer to the coefficients β being **estimated** for – or **fitted** to – the data *given* a model such as equation 1. If the model is wrong (i.e. that data do not conform to a straight-line) then the estimated coefficients will be meaningless

Regression with Categorical Variables

In the example shown in Figure 2, both x and y were **continuous** variables. However, there are situations (classically, those analysed using ANOVA) where x is **categorical**. For example, the data in (Kane et al., 1988) requires an x that can assume values representing treatment type (clozapine or chlorpromazine) and time points (week 0, and week 6).

Categorical variables are those that can take on one of a limited number of values that do not necessarily have numerical meaning, such as sex, ethnicity, or treatment group

In this case, it is convention to represent the *discrete* levels of x as 0 or 1 e.g. the control and treatment groups respectively, which in our case will be chlorpromazine or clozapine treatments. The values of x in equation 1 are now either 0 or 1 (rather than continuous values). So, rather than estimating the gradient of a straight line, we end up with the coefficients β representing changes in the *means* of y as we switch the variable x between 0 and 1 (chlorpromazine *vs* clozapine).

Assigning binary codes to discrete, categorical variables is called **dummy coding**

To expand on this method of coding categorical variables, we will set up a tentative model for the data in (Kane et al., 1988) but use more meaningful variable names instead of x_1, x_2 and so on:

$$Y_{ijt} = \beta_0 + \beta_1 D_{ij} + \beta_2 T_{it} + \beta_3 D_{ij} T_{it} + \epsilon_{ijt} \tag{4}$$

This section can be skipped if the terminology is confusing; we won't depend on detailed understanding of equation 4 but will need the variable names, e.g. D, T and Y. The detail of how regression equations are formally equivalent to performing either ANOVA or an ANCOVA – as used in (Kane et al., 1988) – can be found in (Van Breukelen, 2006)

This somewhat confusing array of notation is easier to understand if we look at some data. Table 1 shows a sample of 10 patients arranged in rows, with columns corresponding to the variables in equation 4 as follows:

- i refers to the patient, enumerated $1 \dots N$, with N denoting the total number of patients (i.e. the number of patients assigned to chlorpromazine, plus the number assigned clozapine)
- j describes the *treatment group*, where for chlorpromazine, $j = 0$ and for clozapine, $j = 1$
- t describes the time point; at baseline, pre-treatment (numerical week 0) $t = 0$ and post-treatment $t = 1$ (i.e. at the sixth week)
- Y_{ijt} is the total BPRS score for patient i in group j at time t
- D_{ij} describes the medication assignment correponding to datapoint Y_{ijt} – that is, for patient i in treatment group j at time t – e.g. if patient i is in the chlorpromazine group, then $D_{ij} = 0$ and if in the clozapine group, $D_{ij} = 1$
- for a given datapoint, T_{it} records whether the value of Y_{ijt} represents the BPRS score *before* treatment (week 0) as $T_{it} = 0$ or *after* treatment (week 6) with $T_{it} = 1$

With reference to Figure 1, here we will only concern ourselves with the first and last time points shown in the graph rather than all 5 intervening time points

- the term $D_{ij}T_{it}$ is the *interaction* between the drug treatment for patient i and the time point t and is obtained by multiplying D_{ij} by T_{it}.

i (Patient)	Y_{ijt} (BPRS)	D_{ij} (Drug)	T_{it} (Time)
43	71	0	1
102	59	0	0
81	53	0	1
107	57	0	0
59	62	0	1
103	42	1	1
125	57	0	0
106	78	0	0
104	59	1	0
9	49	1	1

Table 1: Sample of 10 patients illustrating the terms used in equation 4

For example, inspecting row 3; the patient is $i = 81$, assigned to treatment with chlorpromazine ($D_{ij} = 0$), who at time $T_{it} = 1$ (i.e. post-treatment at week 6) had a BPRS of $Y_{ijt} = 53$. Similarly, the sixth row describes patient 103 who post-treatment with clozapine had a BPRS of 42. In contrast, row 2 describes patient 102, who before treatment ($T_{it} = 0$) with chlorpromazine ($D_{ij} = 0$) had a baseline (pre-treatment, week 0) BPRS of 59.

Now, if we want to estimate the average (mean) BPRS scores for *all* patients in the chlorpromazine group at *baseline*, we would select all rows in Table 1 such that

- there is a zero in the column T_{it} (Time)
- and a zero in the column D_{ij} (Drug)
- resulting in patients $i = \{102, 107, 125, 106\}$

Notice that although patient 104 has a 0 in the Time column, there is a 1 in the Drug column and they are therefore excluded from this calculation as they were assigned to clozapine not chlorpromazine

Then, we compute the mean value of Y_{ijt} (BPRS) for just these rows which is equivalent to calculating equation 4 with all patients i where $j = 0$ and $t = 0$

$$Y_{i00} = \beta_0 + \beta_1 D_{i0} + \beta_2 T_{i0} + \beta_3 D_{i0} T_{i0} + \epsilon_{i00} \tag{5}$$

Next we substitute in the corresponding values for the terms $D_{i0} = 0$ and $T_{i0} = 0$ in equation 5 we get:

$$Y_{i00} = \beta_0 + \cancelto{0}{\beta_1 D_{i0}} + \cancelto{0}{\beta_2 T_{i0}} + \cancelto{0}{\beta_3 D_{i0} T_{i0}} + \epsilon_{i00} \tag{6}$$

Which leaves

$$Y_{i00} = \beta_0 + \epsilon_{i00} \tag{7}$$

Remember that for any product of a number of variables e.g. $a \times b$, if any is zero, then the whole term 'cancels' out : shown by $\cancelto{0}{a \times b}$. Similarly, for $a \times b$, if $a = 1$ the term reduces to $1 \times b = b$ and conversely, for $b = 1$, $a \times 1 = a$

Which will be useful to us later because it tells us that the coefficient β_0 computed by our statistics package is really the baseline level of BPRS scores for patients randomised to chlorpromazine.

Similarly, if we want to estimate the baseline BPRS score for those patients treated with clozapine; this corresponds to $D_{i1} = 1$ and $T_{i0} = 0$ and equation 4 then becomes

$$\begin{aligned} Y_{i10} &= \beta_0 + \beta_1 D_{i1} + \beta_2 T_{i0} + \beta_3 D_{i1} T_{i0} + \epsilon_{i10} \\ Y_{i10} &= \beta_0 + \beta_1 \times 1 + \cancelto{0}{\beta_2 T_{i0}} + \cancelto{0}{\beta_3 D_{i1} T_{i0}} + \epsilon_{i10} \\ Y_{i10} &= \beta_0 + \beta_1 + \epsilon_{i11} \end{aligned} \tag{8}$$

So Y_{i10} – the mean BPRS value at baseline of all patients assigned clozapine – corresponds to the *sum* of the mean baseline BPRS scores for patients assigned to chlorpromazine (β_0, equation 7) plus any additional contributions from those patients assigned to clozapine (β_1) . Thinking about this carefully, if the design of the experiment was robust then $\beta_0 + \beta_1$ should be roughly equal to β_0

That is, if the patients were truly randomised then the BPRS scores at baseline for patients assigned to clozapine should be no more (or less) than those assigned chlorpromazine

Questions and Exercises

1. We calculated how the linear model looks for:

   (a) The chlorpromazine group at baseline: $D_{i0} = 0$ and $T_{i0} = 0$

   (b) The clozapine group at baseline: $D_{i1} = 1$ and $T_{i0} = 0$

   Perform the same calculations for:

   (a) The chlorpromazine group at the end of treatment: $D_{i0} = 0$ and $T_{i1} = 1$

   (b) The clozapine group at the end of treatment: $D_{i1} = 1$ and $T_{i1} = 1$

2. Practical problem: We've seen how the linear model described in equation 4 yields values for the mean BPRS in different treatment groups at different times. Try to 'manually' compute the corresponding values for chlorpromazine and clozapine at baseline.

   (a) Load `Kane-simulated.csv` into your preferred statistics package – it should look like Table 1 only more systematically ordered by Drug and Time.

   (b) Select rows where Drug = 0, and Time = 0 (chlorpromazine, and baseline)

   (c) With only these rows selected, compute the mean of the BPRS column

   (d) Repeat this, but where Drug = 1, and Time = 0.

References

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.

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.