Introduction
A common design in empirical research is the two group pre-post design. The general experimental design involves the allocation of participants to two intervention groups (Group A and Group B, commonly “treatment” and “control”) with measures on participants taken pre- (X) and post- intervention (Y). Despite this being a common design, there is no consensus on the most appropriate method for the analysis of the resulting data.

Analysis of covariance (ANCOVA) is one method for analysing the data collected in the two group pre-post design. ANCOVA can be applied to the pre-post design with the post-test data as the dependent variable (Y), treatment group as the independent variable (Z), pre-test data as the covariate (X). This would test whether the treatment has an effect on the outcome measure after taking into consideration the pre-test scores (Jamison, 2004). This method can be extended with the inclusion of a term for an interaction effect between baseline values and the treatment group (X*Z).

Another method of analysing the two group pre-post design difference-in-differences analysis is also known as gain scores analysis where “gain” is defined as the post-test minus the pre-test data (Knapp and Scheler, 2009). The difference between the pre-test and post-test scores (Y – X) can then be analysed using either the independent samples t-test or the unequal variances t-test. Knapp and Scheler (2009) and Warner and Brown (2006) discuss “Lord’s Paradox” (Lord, 1967) where Lord has suggested for naturally occurring and non-randomized groups that ANCOVA finds no treatment effect whereas a treatment effect is found using gain scores analysis.

Another of the two methods proposed for analysing the two group pre-post design in ANCOVA with X as the dependent variable, X*Y as the covariate, and a group dummy variable with or without an interaction effect. These methods, the dependent variable is X (or pre-test minus post-test), the covariate is X*Y (pre-test plus post-test). These models will test the effect of the treatment on the difference between the pre-test and post-test values while adjusting for the total score of both the pre-test and post-test. Oldham (1982) has shown that X,Y and X*Y are uncorrelated in the absence of an effect but otherwise correlated. The methods described above are examples of “mathematical coupling” while one of the terms appears on both sides of the equation and has been criticised for creating phantom effects (Wahl and Lee, 1998).

The analytical techniques outlined above are compared for statistical validity and statistical power via simulation. In the absence of an effect all valid test would have uniformly distributed p-values (Bland, 2013) or meet Bradley's criterion (Bradley, 1978) in which Type I error rate should lie within [0.05, 0.06] when there is no effect. Simulation is often used to investigate and compare the utility of these five different analytical approaches under idealised RCT (randomized controlled trial) conditions.

Simulation Design

Normally distributed, N(0,1) pre-test data for the two groups X1, ..., X5; Y1, ..., Y5 may be generated in computer software (e.g. Minitab, Excel, R). Post-test data for Group A (X) and Group B (X) is then generated as X = X + random error and for each of the 5 test groups the post-test and pre-test, represents the sample size where X1 = X - n using the equation:

\[ x_i = a_0 + a_1 x_{i-1} + a_2 x_{i-2} + a_3 x_{i-3} + \varepsilon_i \]

where \( a_0 \) is the constant, and \( a_1, a_2, a_3 \) are the parameters of the equation.

The parameters of the equation (\( a_0, a_1, a_2, a_3 \)) will have pre-determined values and each model will be simulated using every combination of the parameters for each of the sample sizes \( X_n = n \) \( n = 16, 32, 64 \) for each of the four groups and for each of the correlation coefficients \( (r = 0.3, 0.3, 0.6, 0.7) \). This means that this simulation will be a 2x2x2x5 design for each of the 5 test statistics. The figure below is one of the flowcharts produced to demonstrate how the code for the simulation works for the parallel lines test statistics.

Results

When the null hypothesis is true (i.e. \( a_2 = a_3 = 0 \)) and if the test statistic is valid then the \( p \)-values should follow a uniform distribution [0–1], (Bland, 2013). This will be checked using probability plots (for example, P-P plot and Q-Q plot). The test statistics will also be tested for the percentage of cases where the \( p \)-value is rejecting the null hypothesis. If the \( p \)-values follow a uniform distribution and are working at the 5% level, then the null hypothesis should be rejected for 5% of all cases. Bradley’s (1975) liberal criterion states that if the data is within \( \pm \delta \) of the null then the \( p \)-values are Type I error robust. Power comparisons will also be performed where \( a_2 = 0 \) and \( a_3 = 0 \).

If the null hypothesis is true (i.e. \( a_2 = a_3 = 0 \)) and if the test statistic is valid then the \( p \)-values should follow a uniform distribution [0–1], (Bland, 2013). This will be checked using probability plots (for example, P-P plot and Q-Q plot). The test statistics will also be tested for the percentage of cases where the \( p \)-value is rejecting the null hypothesis. If the \( p \)-values follow a uniform distribution and are working at the 5% level, then the null hypothesis should be rejected for 5% of all cases. Bradley’s (1975) liberal criterion states that if the data is within \( \pm \delta \) of the null then the \( p \)-values are Type I error robust. Power comparisons will also be performed where \( a_2 = 0 \) and \( a_3 = 0 \).

Conclusion

- All five tests are valid as demonstrated by their behavior when the null hypothesis is true.
- For a randomized design ANCOVA is more powerful than the gain score approach.
- For ANCOVA with an interaction term the model with pre-test score is more powerful than the model with X + Y as the covariate.
- In ANCOVA reversing the roles of X and Y might have a power advantage but the advantage is less than that observed in the simulation and models not presented here.
- The results are above for when the correlation coefficient (\( \rho \)) is equal to 0.5 and the sample size (n) is equal to 64. However, the conclusions were the same for the other correlation coefficients and the other sample sizes that were simulated.
- Overall there is a clear winner which is to use ANCOVA with pre- scores as a covariate and to include a covariate by group interaction effect. This finding supports a principled observation given by (Rogosa, 1980). This finding would also be consistent with prior reasoned identification that changes might be dependent on initial starting position.

Literature Cited