A Comparison of Analytical Methods for the Two Group Pre-Post Design

Vanessa Allis and Paul White, Engineering Design and Mathematics,
University of the West of England, Bristol

Introduction

A common design in empirical research is the two group pre-post design. The general structure of this design is random allocation of participants to one of 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-test design with the post-test data as the dependent variable (Y), treatment group as the independent variable (Z), pre-test (baseline) data as the covariate (X). Therefore, this model will test whether the treatment has an effect after taking into consideration the pre-test scores (Jamieson, 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 Schafer, 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 Schafer (2009) and Wainer 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 is ANCOVA with X-Y as the dependent variable, X+Y as the covariate, and a group dummy variable with or without an interaction effect. In these methods, the dependent variable is X-Y (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 (1962) 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' where one of the terms appears on both sides of the equation and has been criticised for creating phantom effects (Walsh and Lee, 1998).

The analytical techniques outlined above are compared for statistical validity and statistical power via simulation. In the absence of an effect a 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 $\alpha \pm \alpha/2$ when there is no effect. Simulation will be used to investigate and compare the utility of these five different analytical approaches under idealised RCT (randomised controlled trial) conditions.

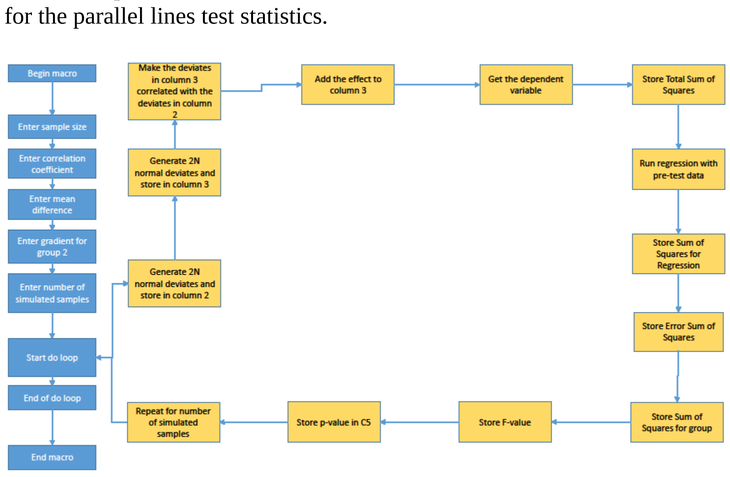
Simulation Design

Normally distributed, N(0, 1) pre-test data for the two groups $x_{1A}, x_{2A}, \dots, x_{nA}$; $x_{1B}, x_{2B}, \dots, x_{nB}$ may be generated in computer software (e.g. Minitab). Post-test data $y_{1A}, y_{2A}, \dots, y_{nA}$; $y_{1B}, y_{2B}, \dots, y_{nB}$ (for both post-test and pre-test, n represents the sample size where $n_a = n_b = n$) using the equation below.

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 x_i z_i + \epsilon_i$$

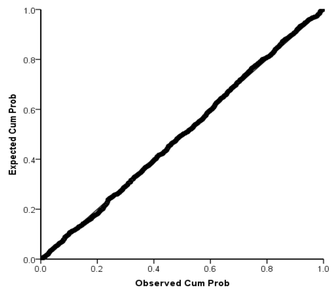
where α_0 is the constant, and α_1, α_2, and α_3 denote the parameters of the equation.

The parameters of the equation (α_1, α_2, and α_3) will have pre-determined values of 0 and 0.5. Each of the ten statistical models will be tested using every combination of the parameters for each of the sample sizes $n_a = n_b = n = 16,\ 32,\ 64$ for each group and for each of the correlation coefficients $\rho = -0.3,\ 0,\ 0.3,\ 0.6,\ 0.9$. This means that this simulation will be a 2x2x3x5 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.

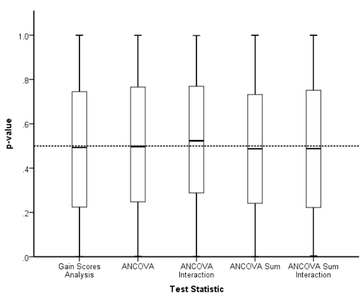


If the null hypothesis is true (i.e. $\alpha_2 = \alpha_3 = 0$) and if the test statistic is valid then the p-values should follow a uniform distribution ($X \sim U(0, 1)$) (Bland, 2013). This will be checked using probability plots (for example, P-P plot and Q-Q plot). The test statistics will also 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. However, Bradley's (1978) liberal criterion states that if the data is within $\alpha \pm \alpha/2$ of α then the p-values are Type I error robust. Power comparisons will also be performed when $\alpha_2 \neq 0$ and $\alpha_3 \neq 0$.

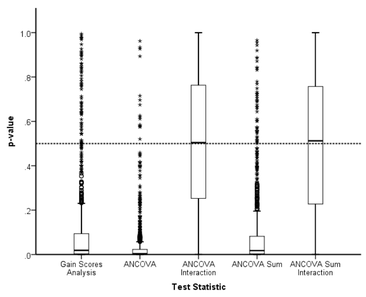
Results



When the null hypothesis is true (i.e. $\alpha_2 = \alpha_3 = 0$) and if the test statistic is valid then the p-values should follow a uniform distribution ($X \sim U(0, 1)$) (Bland, 2013). This was tested through P-P plots where if the p-values are uniformly distributed then the points will follow the line. The p-values were also tested for uniformity using boxplots where the expected mean for a uniform distribution is 0.5 and the expected standard deviation is $\sqrt{(1/12)} = 0.2887$.



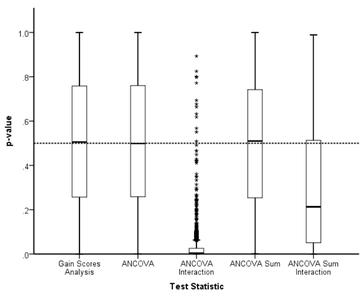
The figure to the left is a P-P plot testing the p-values simulated for ANCOVA for uniformity. The points closely follow the line, therefore, suggesting that the p-values follow a uniform distribution. The boxplots to the right are also consistent with this finding.



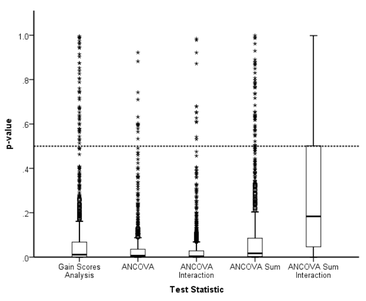
The next set of parameters that was simulated was when there is no change in gradient, but there is a main effect of 0.5 i.e. α_2 = 0.5. The expectation under these parameters is that the gain scores analysis, the ANCOVA, and the ANCOVA with X-Y as the dependent variable and X+Y as covariate would not follow a uniform distribution. This is because these tests are testing for a main effect. Contrary to this, the expectation is that for the ANCOVA with an interaction and ANCOVA with X-Y as the dependent variable and X+Y as the covariate with an interaction the p-values would follow a uniform distribution.

The boxplot to the left is consistent with hypothesised effects indicating the power of ANVOA and the gain score approach in the presence of a main effect and with a power advantage to ANCOVA.

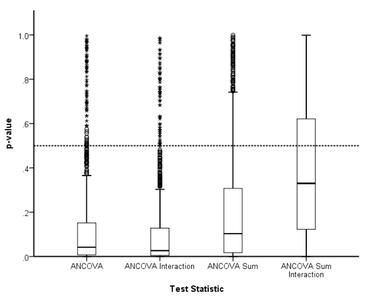
The next set of parameters that was simulated was when there is no main effect, but there is a change in gradient of 0.5 i.e. α_2= 0, α_3= 0.5. The statistical tests that include an interaction term are expected to be more powerful under these parameters. whereas, the statistical tests that do not include an interaction term are not expected to work effectively. The p-values simulated with these parameters are tested using the boxplot to the right.



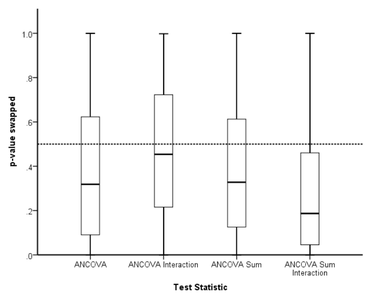
This boxplot to the right is consistent with these hypothesised effects and the ANCOVA with X as the covariate displays a power advantage compared to the ANCOVA with X+Y as the covariate.



When the change in gradient is 0.5 and the main effect is 0.5, the statistical tests should all be working correctly (and, therefore, not following a uniform distribution). This is because the statistical tests all either test for a change in gradient or a main effect. The tests that are the most powerful will be the furthest away from a uniform distribution. The boxplot to the left clearly shows that all four test statistics do not follow a uniform distribution. This means that all four test statistics are working as expected. The test statistic with the lowest p-values is ANCOVA with an interaction followed by ANCOVA. This suggests that ANCOVA with an interaction is the most powerful test statistic and, therefore, is most suited for analysing data with these parameters. The test statistic with the largest p-values is ANCOVA with X-Y as the dependent variable and X+Y as the independent variable with an interaction. This suggests that this test statistic is the least powerful and, thus, is the least suitable for analysing data with these parameters.



The test statistics analysed in the results above are testing which test statistic is most powerful for analysing the two group pre-post design. Gain scores analysis is the only test statistic which is palindromic invariant and, therefore, is not tested further. However, the other four test statistics are not palindromic invariant and are tested further to see if they produce similar conclusions when the roles of the dependent variable and covariate are reversed and whether there is a power advantage to the swapping of the dependent variable and covariate.



The boxplot to the left is for the un-swapped data and the boxplot to the right is for the swapped data. The first boxplot shows that the test statistic with the most power is ANCOVA with an interaction effect. The second boxplot of the swapped data shows that the test statistic with the most power is ANCOVA with X+Y as the covariate and the one with the least power is ANCOVA with an interaction. However, ANCOVA with an interaction testing the un-swapped data is more powerful than ANCOVA with X+Y as the covariate on the swapped data. Thus, this suggests that the test statistics are producing different results when the dependent variable and covariate roles are reversed. Moreover, there is no power advantage to reversing the roles of the dependent variable and covariate.

Conclusions

- 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 scores X as a covariate is more powerful than the model with X + Y as the covariate
- In ANCOVA (with or without an interaction term) there is no power advantage to be had by reversing the roles of X and Y in the model
- In ANCOVA reversing the roles of X+Y and X-Y might have a power advantage but the advantage is less than that observed in the models not using X+Y
- The results above are for when the correlation coefficient (ρ) is equal to 0.3 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

Bland, M. (2013) Do baseline p-values follow a uniform distribution in Randomised trials? Law, M., ed. PLoS ONE. 8 (10), p. e76010.

Bradley, J. (1978) Robustness?. *British Journal of Mathematical and Statistical Psychology*. 31 (2), pp. 144-152. [Accessed 7 March 2017].

Jamieson, J. (2004) Analysis of covariance (ANCOVA) with difference scores. International Journal of Psychophysiology. 52 (3), pp. 277–283.

Knapp, T. and Schafer, W. (2009) From Gain Score t to ANCOVA F (and vice versa). Practical Assessment, Research & Evaluation. 14 (6).

Lord, F.M. (1967) A Paradox in the Interpretation of Group Comparisons. Psychological Bulletin. 68 (5), pp. 304–305.

Oldham, P.D. (1962) A note on the analysis of repeated measurements of the same subjects. Journal of Chronic Diseases. 15 (10), pp. 969–977.

Rogosa, D. (1980) Comparing nonparallel regression lines. *Psychological Bulletin*. 88 (2), pp. 307-321.

Wainer, H. and Brown, L.M. (2006) Three statistical Paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. Handbook of Statistics, pp. 893–918.

Walsh, T.S. and Lee, A. (1998) Mathematical coupling in medical research: Lessons from studies of oxygen kinetics. British Journal of Anaesthesia [online]. 81 (2), pp. 118–120. [Accessed 4 October 2016].