ANALYSIS OF CLUSTERED MEASUREMENTS: A COMPARISON OF THE PERFORMANCE OF FOUNDATION YEAR STUDENTS, 1994 COHORT WITH THOSE OF DIRECT STUDENTS, 1995 COHORT, AT THE UNIVERSITY OF LIMPOPO, SOUTH AFRICA

Maupi E Letsoalo[1] and [2]Maseka Lesaoana
[1]Tshwane University of Technology, South Africa
[2]University of Limpopo, South Africa
masekal@ul.ac.za

In applied sciences, one is often confronted with the collection of correlated data. Dependence between observations from the same study subject renders invalid the usual chi-square tests of independence and inflates the variance of parameter estimates. Disaggregated approaches such as hierarchical linear models which are able to adjust for individual level covariates are suitable in the analysis of such data. We compare the performances of Ex-UNIFY (Foundation Year) students and direct students at first year of Bachelor of Science programme at the University of Limpopo, South Africa. Both aggregated and disaggregated analyses are performed. The results show that Ex-UNIFY students perform better than direct students in their main examination of first year in their Bachelor of Science programme. Disaggregated analysis is able to adjust for individual level covariates, therefore gaining power over aggregated analysis.

INTRODUCTION

Data from intervention studies and sample surveys are often characterized by a hierarchical structure (Ukoumunne, Gulliford & Chinn, 2004). A two-level hierarchy is established when measurements are repeated on the same study subjects. Measurement repetitions or occasions are referred to as level-1 units and subjects as level-2 units (Goldstein & Woodhouse, 2001). Level-2 units are sometimes referred to as clusters, groups or communities (Murray, 1998; Donner & Klar, 2000). In longitudinal studies the outcome variable is for the same individual on several occasions, while in cross-sectional studies each subject is measured only once (Twisk, 2003; Goldstein & Woodhouse, 2001).

Standard regression modeling usually assumes that the errors have zero mean and are mutually independent. However, in clustered data or panel data, it is expected that errors for the same subject are correlated (Rabe-Hesketh & Skrondal, 2005). The distinguishing feature of hierarchical or grouped data is that observations within a cluster may be correlated, and the degree of similarity among responses within a cluster is measured by a parameter called intracluster or intraclass correlation coefficient, or simply ICC (Donner, Piaggio & Villar, 2003). The ICC may be interpreted as the standard Pearson correlation coefficient between any two responses in the same cluster. If we add the assumption that the ICC cannot be negative, then ICC may also be interpreted as the proportion of overall variation in response that can be accounted for by the between-cluster variation (Donner & Klar, 2000). A positive ICC implies that the variation between observations in different clusters exceeds the variation within clusters, hence it can be claimed that the design is characterized by 'between-cluster variation' (Donner & Klar, 2000).

Among others, Aitkin, Anderson and Hinde (1981), demonstrated that when the analysis accounts properly for the grouping, i.e. when ICC is taken into consideration then a more reliable conclusion can be drawn from the analysis. Thus, early work on the analysis of hierarchical data in the context of education was carried out by Aitkin et al. (1981) and later by Aitkin and Longford (1986).

MODELING CLUSTERED DATA

The results obtained from a standard statistical analysis which assumes all observations to be independent in clustered data, may be misleading. Such analysis is referred to as *naïve pooling* (Burton, Gurin & Sly, 1998). Therefore, standard analysis, which ignores an important correlation structure, may well be misleading. One way of solving this problem is to create a single summary statistic such as the mean, for each cluster. This approach, referred to as *data resolution*,

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

automatically avoids any over inflation in the apparent size of the dataset (Burton et al., 1998). The best statistic to use will depend upon the research question being asked.

There are occasions when the only data available for analysis have already been aggregated to a higher level. For example, as stated by Goldstein (1995, p. 63), we may have information on student achievement in terms of the mean achievement for each school, or information on utilization of health services only in terms of total number of episodes for each administrative area.

Aggregated or Clustered Level Analysis

The primary aim of many intervention studies or trials is to compare two groups of subjects with respect to their mean values on outcome variable, assumed to have an approximate normal distribution (Donner & Klar, 2000). The two-sample t-test is employed in the testing of the null hypothesis that the means of the two groups are statistically not different. According to Donner and Klar (2000) the mean of Y_{ijl} (the response of subject l in the j*th* cluster of group *i*), is defined as:

$$\overline{Y}_i = \frac{\sum_{j=1}^{k_i} m_{ij}\overline{Y}_{ij}}{\sum_{j=1}^{k_i} m_{ij}}, \quad i = 1, 2 \tag{1}$$

and the sample variance of the observed score in the subgroups is defined as:

$$S_i^2 = \sum_{j=1}^{k_i}\sum_{l=1}^{m_{ij}} \frac{\left(Y_{ijl} - \overline{Y}_i\right)^2}{M_i - 1}, \quad i = 1, 2 \tag{2}$$

The standard two-sample t-statistic under the null hypothesis of no difference between the true score values, is given by

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\left(\frac{\sum_{i=1}^{2}(M_i - 1)S_i^2}{M - 1}\left(\sum_{i=1}^{2}\frac{1}{M_i}\right)\right)}} \tag{3}$$

where

$\overline{Y}_1 - \overline{Y}_2$ estimates the population mean difference $\mu_1 - \mu_2$,

M_i is the total number of observations (or individuals) in group *i*,

M is the total number of individuals in the study, and

m_{ij} is the total number of subjects in a cluster.

Disaggregated Analysis

A hierarchy consists of units grouped at different levels. Thus, observations may be level-1 units in a 2-level structure, where the level-2 units are the subjects. In multilevel research, variables can be defined at any level of hierarchy, and some of these variables may be measured directly at their natural level (Hox, 1995). A more general way to look at multilevel data is to investigate a cross level hypothesis, or multilevel problem which concerns the relationships between variables that are measured at a number of hierarchical levels (Hox, 1995).

The main statistical model of multilevel analysis is the hierarchical linear model - an extension of the multiple linear regression model to a model that includes nested random coefficients. Hierarchical linear models are sometimes called multi-level linear models, nested models, mixed linear models or covariance-component models, and they have historically been used in researches where hierarchies occur naturally (Sullivan, Dukes & Losina, 1999). A full

multilevel regression model assumes that there is a hierarchical dataset with one single dependent variable that is measured at the lowest level (or level-1) and explanatory variables at all existing levels (Hox, 1995).

Analysis of continuous outcome data from clustered data or group randomized trials can often be accomplished using mixed-effect linear models, as fitted by generalized least squares (Donner, 1985). The mixed-effect linear models are used to estimate the effect of treatment, to test if the observed effect is due to chance, and to adjust for imbalance on baseline risk factors, and it is given by:

$$y_{ijst} = \beta X_{ijst} + C_{(ij)s} + \varepsilon_{(ijs)t} \quad (4)$$

where

y_{ijst} is the score or observation of the tth subject from cluster s, treatment j and strata i. The vector β represents the fixed effect of treatment, strata and baseline risk factors. $C_{(ij)s}$ and $\varepsilon_{(ijs)t}$ denote the respective independent random effects of clusters nested in treatment and stratum, assumed to be *iid* $N\left(0,\sigma_c^{\,2}\right)$, and subjects nested in the cluster, assumed to be *iid* $N\left(0,\sigma^2\right)$, (Donner & Klar, 1994).

A special case of a mixed effect linear regression model is an approach which is based on two-stage nested analysis for testing the null hypothesis, $H_0 : \mu_o = \mu_1$ against the alternative hypothesis $H_1 : \mu_0 \neq \mu_1$, where μ_0 and μ_1 are the underlying means of the outcome variable y. Ukoumunne et al. (2004) provide the following model:

$$y_{ijk} = \mu + \beta_1 x_{1i} + \mu_{ij} + \varepsilon_{ijk} \quad (5)$$

where μ is the true mean response, and β_1 is a regression constant representing the fixed effect of the intervention group i $(i = 0, 1)$, such that

$$x_{1i} = \begin{cases} \text{Active, if } i = 1 \\ \text{Control, if } i = 0 \end{cases} \quad (6)$$

μ_{ij} denotes a random cluster effect, assumed to be $N\left(0, \sigma_A^2\right)$.

ε_{ijk} denotes the error term, assumed to be $N\left(0, \sigma_w^2\right)$.

DESCRIPTION OF THE DATASET

The dataset used in this article was supplied by the University of Limpopo's IT Section. The University of Limpopo is a result of a merger in January 2005, between the former University of the North (UNIN) and MEDUNSA (Medical University of South Africa). University of the North Foundation Year programme (UNIFY) is a one year programme that prepares a group of selected students to enter into a science-based degree programme.

The dataset is about Ex-UNIFY students, those students who prior to enrolling for the Bachelor of Science degree in 1995 studied year-long courses offered through UNIFY, and Direct students, students who were admitted directly into the Bachelor of Science degree in 1995, without having to go through UNIFY. Both these cohorts were registered in the Faculty of Mathematics and Natural Sciences of the then University of the North. The performance of the two groups is compared in their first year of a Bachelor of Science degree programme.

The UNIFY programme is aimed at:

i. increasing the number of science graduates among the disadvantaged groups, students who are educationally at risk of not continuing with studies at tertiary level in their field of interest, or inadequately prepared students;
ii. recruiting and increasing suitable female students in University of Limpopo science related studies;
iii. increasing the number of students in the University of Limpopo science based faculties; and
iv. improving the quality of students in science faculties.

Students admitted in UNIFY are those who do not fully qualify for direct admission into degree programme, because they do not meet the minimum admission requirements set forth by science faculties (Mabila, Malatje, Addo-Bediako, Kazeni & Mathabatha, 2006). Applicants must have passed the matriculation examination (have at least school leaving certificate) and must show evidence that mathematics and one of the science subjects were taken at this level, but not necessarily passed (Mabila et al., 2006).

UNIFY offers courses in Biology, Chemistry, English and Study skills, Mathematics and Physics. Students at the University of Limpopo are evaluated at two stages in each academic year. Students accumulate a year mark by passing assignments and tests during the academic year, and a student qualifies to write end-of-year examination called main examination if his/her year-mark is at least 40%. A student passes a course if the final mark, computed as a weighted average of year mark and examination mark, is at least 50%.

The final mark for the two groups, Ex-UNIFY students and Direct students, are compared at the end of their first year using random-intercept model, which calculates the maximum likelihood estimates of the regression coefficient of *group membership*, whether a student was Ex-UNIFY or Direct student.

RESULTS

The 1995 first year Bachelor of Science study population consisted of 117 Ex-UNIFY students (21 female students and 96 male students) and 457 direct students (187 female students and 270 male students). This study population was randomly chosen since the work in this project concerns the statistical modeling issues not the impact of foundation year programmes.

Application of two-independent-samples t-test: Naïve Pooling and Aggregated Analysis

The independent two-sample t-test was employed to compare the average performance between the two groups (Naïve pooling). Table 1 indicates that Ex-UNIFY students had significantly higher averages than direct students, 52.55095 and 47.19009, respectively ($p < 0.001$); and the 95% CI for the difference is (-6.74; -3.97). Therefore, of the students who enrolled for courses offered in the Faculty of Mathematics and Natural Sciences in 1995, Ex-UNIFY students had scored significantly higher marks than direct students in the Bachelor of Science programmes.

Table 1. Naïve Pooling

```
                              Two-sample t test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Direct   |    1857    47.19009    .3493222    15.05332    46.50499    47.8752
Ex-UNIFY |     579    52.55095    .5912885    14.22783    51.38961    53.71229
---------+--------------------------------------------------------------------
combined |    2436    48.46429    .3045772    15.03267    47.86703    49.06154
---------+--------------------------------------------------------------------
    diff |           -5.360858    .7073825               -6.747992   -3.973724
------------------------------------------------------------------------------
    diff = mean(Direct) - mean(Ex-UNIFY)                          t =  -7.5784
Ho: diff = 0                                     degrees of freedom = 2434

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Table 2 indicates the results of the independent two-sample t-test which was employed in the aggregated analysis or data resolution to compare the average performances between the two groups. The results indicate that Ex-UNIFY students had significantly higher averages than direct

students, 52.502 and 45.422, respectively ($p < 0.001$); and the 95% CI for the difference is (-9.77; -4.39).

Table 2. Data Resolution

```
                              Two-sample t test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Direct   |     454    45.42198    .6379812    13.59364    44.16821    46.67575
Ex-UNIFY |     115    52.50216    1.024168    10.98298     50.4733    54.53103
---------+--------------------------------------------------------------------
combined |     569    46.85295    .5619487    13.40457     45.7492     47.9567
---------+--------------------------------------------------------------------
    diff |           -7.080185    1.368676               -9.76848   -4.391891
------------------------------------------------------------------------------
    diff = mean(Direct) - mean(Ex-UNIFY)                          t =  -5.1730
Ho: diff = 0                                     degrees of freedom =   567

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Results for the Disaggregated Analysis

The ICC, ρ was found to be about 0.3997 for unadjusted model and 0.3992 for adjusted model, implying that the designs are characterized by 'between-cluster variation'.

The unadjusted model indicates that Ex-UNIFY students are expected to significantly score an average final mark of about 6.49 (95% CI: 4.1657 – 8.8275) points higher than Direct students (Table 3).

Table 3. The hierarchical models

	Unadjusted Model	**Adjusted Model**
	Final-Mark	Final-Mark
Ex-UNIFY	6.497	6.389
	(5.46)**	(5.27)**
Male		0.463
		(0.44)
Constant	46.208	45.934
	(82.20)**	(54.72)**
Observations	2436	2436
Number of student	569	569

Absolute value of z statistics in parentheses

***** significant at 1%***

Adjusting for sex, whether a student was a male or a female, Ex-UNIFY students are expected to significantly score an average final mark of about 6.389 (95% CI: 4.0119 – 8.7669) more than Direct students.

DISCUSSION

In this article, we have demonstrated the applications of two approaches to analyse clustered data. The first approach assumes independence of observations and the other approach recognises dependency of observation to one another. Strictly speaking the researchers have taken into account the intraclass correlation coefficient when dealing with clustered data. Failure to take

account of the ICC does not bias the point estimate of the intervention effect, but leads to falsely narrow confidence intervals.

The modeling approach mostly used for regression analysis when the outcome is continuous in clustered data is generalized linear mixed models (GLMM), which may be characterized as 'population averaged' in that it measures the expected change in a response as the value of covariate increases by a unit.

The advantage of using disaggregated approaches is their ability to adjust for individual level covariates, and thus gain more power. However, there is nothing wrong with aggregation in cases where the researcher is only interested in macro-level propositions, although it should be borne in mind that the reliability of an aggregated variable depends, among other things, on the number of micro-level units in a macro-level unit.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Nana Selota (IT Section, University of Limpopo, South Africa) for making data available. Our gratitude also goes to Makgathe Ronny Sekgobela (Department of Education, University of Limpopo, South Africa) for managing the data used in this research.

REFERENCES

Aitkin, M., Anderson D., & Hinde J. (1981). Statistical modeling of data on teaching styles (with discussion). *J. R. Stat. Soc. A148*, 144-161.

Aitkin, M., & Longford N. (1986). Statistical modeling in school effectiveness studies (with discussion). *J. R. Stat. Soc. A149*, 1-43.

Burton P., Gurin L., & Sly P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine, 17,* 1261-1291.

Donner A. (1985). A regression approach to the data arising from cluster randomization. *International Journal of Epidemiology, 2*, 322-326.

Donner A., & Klar N. (2000). *Design and analysis of cluster randimization trials in health research.* London: Arnold.

Donner A., & Klar N. (1994). Cluster randomization Trials in Epidemiology: Theory and Application. *Journal of Statistical Planning and Inference, 42*, 37-56.

Donner A., Piaggio G., & Villar J. (2003). Meta-analyses of cluster randomization trials: Power considerations. *Evaluation & the Health Professions, Vol. 26(3),* 340-351.

Goldstein H. (1995). *Multilevel statistical models* (2nd Edition). Arnold:London.

Goldstein H., & Woodhouse G. (2001). *Modeling repeated measurements in multilevel modeling of health statistics*. In A. H. Leyland & H. Goldstein (Eds.). John Wiley & Sons, Ltd.

Hox J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties.

Mabila T. E., Malatje S. E., Addo-Bediako A., Kazeni M. M. M., & Mathabatha S. S. (2006). The role of foundation year programmes in science education: The UNIFY Programme at the University of Limpopo, South Africa. *International Journal of Educational Development, 26(3),* 295-304.

Murray D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press, Inc.

Rabe-Hesketh S., & Skrondal A. (2005). Multilevel and longitudinal modeling using Stata. StataCorp LP.

Sullivan, L. M., Dukes, K. A., & Losina, E. (1999). An introduction to hierarchical linear modeling. *Statistics in Medicine, 18,* 855-888.

Twisk, J. W. R. (2003). Applied longitudinal data analysis for Epidemiology: A Practical Guide. Cambridge University Press.

Ukoumunne, O. C., Gulliford, M. C., & Chinn, S. (2004). On the distribution of random effects in a population-based multi-stage cluster sample survey. *Journal of Official Statistics, 20*(3), 481-493.