

# A LONGITUDINAL STUDY OF FACTORS AFFECTING CHILDREN'S BEHAVIOUR

Ivan Carrillo, Christina Chu, Wanhua Su, Xinlei Xie<sup>1</sup>

## ABSTRACT

In this paper we present a longitudinal data analysis that examines children's behavioural development over time. We consider model-based analysis and model assisted design-based analysis (i.e. joint randomization analysis) of the data collected in the National Longitudinal Study on Children and Youth (NLSCY). The primary objective of this study is to investigate the relationship between a child's aggressive behaviour and some explanatory variables of interest such as age, gender, and family socio-economic conditions. The two approaches found age to be an important factor that affects a child's aggressive behaviour in a non-linear way. Furthermore, conclusions drawn from the model-based analysis would appear misleading when the design features were not taken into account.

**KEY WORDS:** Available case, Complete case, Generalized estimating equation (GEE), Joint randomization estimator, Model-based estimator.

## RÉSUMÉ

Dans cet article, nous présentons une analyse de données longitudinales qui étudie le développement comportemental des enfants dans le temps. Nous considérons une analyse fondée sur un modèle ainsi qu'une analyse fondée sur le plan mais assistée d'un modèle (i.e. une analyse de randomisation conjointe) des données recueillies par l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ). L'objectif premier de cette analyse est d'étudier la relation entre le comportement agressif d'un enfant et quelques variables explicatrices telles que l'âge, le sexe et les conditions socio-économiques de la famille. Les deux approches ont déterminé que l'âge est un facteur important qui affecte, de façon non linéaire, le comportement agressif de l'enfant. De plus, les conclusions tirées de l'analyse fondée sur le modèle semblent induire en erreur lorsque les composantes du plan de sondage ne sont pas prises en compte.

**MOTS CLÉS :** Équations d'estimation généralisées; estimateur selon modèle; estimateur selon randomisation conjointe; un cas complet; un cas disponible.

## 1. INTRODUCTION

The data set used in this study is a subset of the National Longitudinal Study on Children and Youth (NLSCY) data, which comprises repeated observations from cycle 1 to cycle 4. The selected subjects satisfy the following criteria: they must be 2 to 5 years old and have a non-zero longitudinal weight at cycle 1; dropout is the only possible missing pattern from cycle 1 through cycle 4; they did not change their province of residence across the survey cycles. As a result, there are all together 1033 children selected in the data set.

In this analysis, there are two response variables, namely emotional disorder anxiety (EDA) score and physical aggression (PA) score. A higher EDA score indicates a stronger tendency of behaviours associated with anxiety and emotional disorders, whereas a higher PA score indicates a bigger trend of hostile behaviours and physical aggression. For children who are 2 to 3 years old, EDA score is scaled from 0 to 12 based on six questions; while for children who are 4 to 11 years old, the score is scaled from 0 to 14 based on seven questions. Similarly, PA score is scaled from 0 to 16 based on eight questions for children who are 2 to 3 years old, and is scaled from 0 to 12 based on six questions for children who are 4 to 11 years old. At cycle 1, 65 EDA scores for children who are 2 to 3 years old are based on the questionnaire that is designed for children who are 4 to 11 years old; on the other hand, 70 scores for children who are 4 to 5 years old are obtained by using the questionnaire that is designed for children who are 2 to 3 years old. For the results to be comparable

---

<sup>1</sup> Ivan Carrillo, Christina Chu, Wanhua Su, and Xinlei Xie: Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1. Email address: iacarril@uwaterloo.ca, cychu@uwaterloo.ca, wsu@uwaterloo.ca, and x2xie@uwaterloo.ca.

across different age groups, EDA scores are unified to a scale of 0 to 14, and PA scores to a scale of 0 to 12. Although both EDA and PA scores are ordinal data, we treated them as continuous variables during data analysis.

The explanatory variables include number of hours in daycare per week; family status (whom a child lives with); number of siblings that live in the same household, including full, half, step, adopted and foster siblings but excluding him/herself; depression score of the PMK (the person who is most knowledgeable about the child)---a score ranging from 0 to 36, a high score indicates the presence of depression symptoms; the highest education of the PMK at cycle 1; urban/rural (a categorical variable indicating the range of the population size of the area where the child lives according 1996 census); province of residence; current working status of PMK; age and gender. Each record also contains the information of child identifier, which is unique for each child. Some explanatory variables have been recoded, see appendix for detail.

The objective of this study is to examine the relationship between EDA (or PA) score and the explanatory variables. In the model-based analysis, marginal generalized linear models (MGLMs) are fit and generalized estimating equations (GEEs) approach is applied to make inference on the regression coefficients. As for the joint randomization analysis, characteristics of the survey such as weighting, stratification and clustering are incorporated into the model-based approach. Hence, results obtained from the model-based analysis and the joint randomization analysis are compared to investigate which approach is better to analyze the NLSCY data.

This paper is organized as follows. Section 2 outlines the model-based estimator obtained by the GEE approach and the joint randomization estimator obtained by incorporating model and design features. Section 3 compares the results obtained by these two methods. Finally, conclusions and comments are given in section 4.

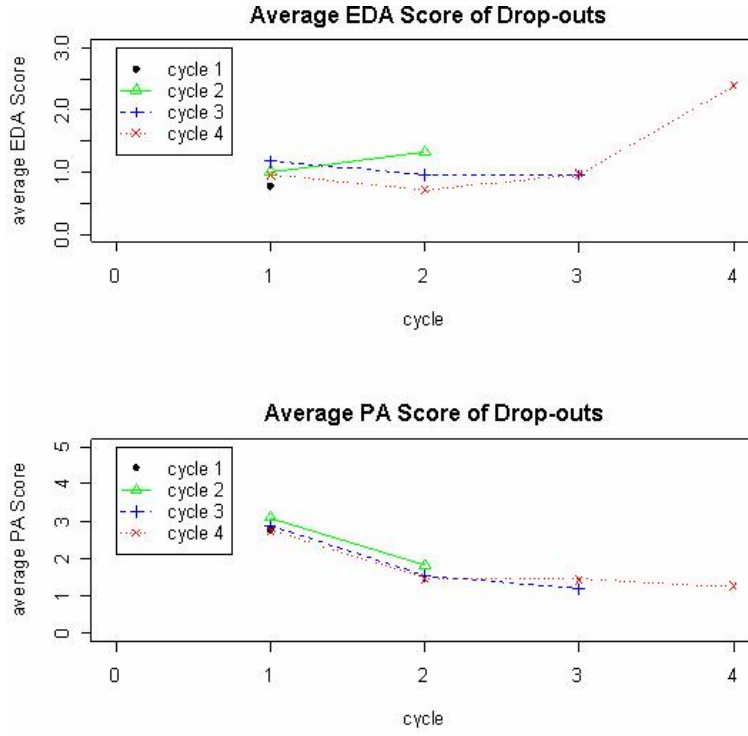
## 2. METHODOLOGY

### 2.1 Understanding Dropouts

Just as other longitudinal studies, the NLSCY survey has the missing data problem. The main reason for incomplete data in the study is due to attrition; in other words, subjects drop out in subsequent cycles prior to the end of the study. Before analyzing the data, we have to decide which subset of the data to use. One option is to use the available data, which includes all cases where the observations of the variables of interest are present. An alternative option is to use the complete data, which includes subjects who have completed the study and do not have missing covariates.

*Figure 1* plots the average EDA/PA scores for those children who discontinued the study at each cycle. This is a descriptive way to visualize the dependence of the dropout process with the response. The graphs show that the dropout pattern at each cycle for both EDA and PA Scores are consistent, so the dropout is not informative with respect to the change of response variables. No significantly distinguishable patterns across dropout groups were found; therefore, it is not necessary to model the dropout process. In the data set, there are 810 complete cases (out of 1033), and this sample size is large enough so that available case and complete case analyses give similar results. Thus, complete-case data is used for the analysis in subsequent sections.

*Figure 1 Average scores of drop-outs at four cycles*



## 2.2 Model-based Analysis

For model-based analysis, the model is  $Y_i = X_i' \beta + \varepsilon_i$ , where  $X_i$  is a  $4 \times (p+1)$  matrix, with rows being the values of the explanatory variables (including the intercept term) for child  $i$  in four cycles,  $Y_i$  is a  $4 \times 1$  column vector containing the EDA (or PA) scores for child  $i$  across 4 cycles, and  $\varepsilon_i$  represents the random error. Furthermore,  $\hat{R}$  is a  $4 \times 4$  matrix which estimates the within subject correlation for any pair of cycles, and  $\beta$  is the coefficient vector of the explanatory variables, which can be estimated by weighted least square

$$\hat{\beta} = \left[ \sum_{i=1}^n X_i' \hat{R}^{-1} X_i \right]^{-1} \left[ \sum_{i=1}^n X_i' \hat{R}^{-1} Y_i \right]. \quad (1)$$

In this study, we assume that the correlation structure  $R$  is the same for all subjects. The model variance of  $\hat{\beta}$  is estimated by

$$\hat{Var}_{\xi}(\hat{\beta}) = \left[ \sum_{i=1}^n X_i' \hat{R}^{-1} X_i \right]^{-1} \left[ \sum_{i=1}^n X_i' \hat{R}^{-1} \hat{Var}(Y_i) \hat{R}^{-1} X_i \right] \left[ \sum_{i=1}^n X_i' \hat{R}^{-1} X_i \right]^{-1}, \text{ where } \hat{Var}(Y_i) = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{Y}_i)(Y_i - \hat{Y}_i)'$$

The subscript  $\xi$  indicates the super-population model. The estimates are obtained by using the `proc genmod` in SAS or `GEE` function in R with an unspecified working correlation structure. Both SAS and R give identical results. Furthermore, step-wise backward selection is applied to choose which explanatory variable should be included in the model. Since the underlying distributions of EDA and PA scores are not normal, the estimates given in (1) are weighted least squares estimates rather than maximum likelihood estimates (MLE).

## 2.3 Joint Randomization Analysis

### 2.3.1. Motivation and Model

Liang and Zeger (1986) proposed a model-based approach to account for the correlation within subjects in longitudinal data analysis under the assumption that subjects are randomly sampled from the study population. In a complex survey such as the NLSCY, however, this assumption may be violated in that subjects are recruited from a particular finite population through a complex sampling design with different inclusion probabilities, stratification or clustering. In other words, subjects surveyed over time cannot be naïvely assumed to be independent from each other (due to clustering) or identically distributed (due to stratification and differential weights).

Binder and Roberts (2003) argue that if the assumed model is incorrect, model-based analysis will lead to biased estimates, and underestimated variances of the estimators. On the other hand, a joint randomization approach is robust against informative sampling scheme where the assumed model for the finite population is no longer valid for the sample data. Using design weights in an appropriate way in analysis will provide valid inference in that situation. For complex survey data, there are two sources of randomization: one is the randomization of the “super-population” model generating the finite population; and the other one is the randomization of the sampling design generating the sample from the finite population. The main difference between joint randomization approach and model-based approach lies in that the former method not only accounts for the variability due to the model but also takes into account the variability contributed by the design.

### 2.3.2. Joint Randomization Estimates

In the joint randomization analysis, the estimated coefficients under the sampling design,  $\hat{\beta}_p$ , are given by

$$\hat{\beta}_p = \left[ \sum_{i=1}^n w_i X_i' \hat{R}^{-1} X_i \right]^{-1} \left[ \sum_{i=1}^n w_i X_i' \hat{R}^{-1} Y_i \right], \quad (2)$$

where the subscript  $p$  indicates estimations under the sampling design, and  $w_i$  is the “funnel weight” for child  $i$ . Equation (2) looks similar to (1), except that weights are incorporated. Again, estimates can be obtained by using `proc genmod` in SAS with the `weight` statement. The variance of the estimate is decomposed into two parts—variation coming from the model randomization and variation owing to the design randomization. The joint variance of  $\hat{\beta}_p$ ,  $Var_{\xi p}(\hat{\beta}_p)$ , is estimated as

$$\hat{Var}_{\xi p}(\hat{\beta}_p) = \hat{Var}_{\xi}(\beta_N) + \hat{Var}_p(\hat{\beta}_p), \quad (3)$$

where the subscript  $\xi p$  indicates estimations under the joint randomization approach, and  $\beta_N$  is the unknown population parameter vector that could be found by fitting an unweighted GEE model if the finite population is known. The model variance is given by

$$\hat{Var}_{\xi}(\beta_N) = \left[ \sum_{i=1}^n w_i X_i' \hat{R}^{-1} X_i \right]^{-1} \left[ \sum_{i=1}^n w_i X_i' \hat{R}^{-1} \hat{Var}(Y_i) \hat{R}^{-1} X_i \right] \left[ \sum_{i=1}^n w_i X_i' \hat{R}^{-1} X_i \right]^{-1}, \quad (4)$$

where  $\hat{Var}(Y_i) = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{Y}_i)(Y_i - \hat{Y}_i)'$ , and the fitted value  $\hat{Y}_i$  is calculated using  $\hat{\beta}_p$  in (2). Furthermore,  $\hat{Var}_p(\hat{\beta}_p)$  is obtained by the Bootstrap method. Let  $\hat{\beta}_p^{(j)}$  be the estimated coefficients using the  $j$ th bootstrap replicate funnel weights, and then  $\hat{Var}_p(\hat{\beta}_p)$  is given by  $\hat{Var}_p(\hat{\beta}_p) = \frac{1}{B-1} \sum_{j=1}^B (\hat{\beta}_p^{(j)} - \hat{\beta}_p)(\hat{\beta}_p^{(j)} - \hat{\beta}_p)'$ , where  $B$  is the bootstrap sample size. Under the joint randomization approach, the theoretical variance is given by  $Var_{\xi p}(\hat{\beta}_p) = Var_{\xi} [E_p(\hat{\beta}_p)] + E_{\xi} [Var_p(\hat{\beta}_p)]$ . To estimate the model variance component,  $Var_{\xi} [E_p(\hat{\beta}_p)]$ , one relies on the assumption that the model is appropriate.

## 3. RESULTS

Table 3.1 compares the results obtained by both model-based approach and joint randomization approach for the EDA score; and Table 3.2 compares the results obtained by both model-based approach and joint randomization approach for the PA score.

Table 3.1 Estimated coefficients and results for EDA score under both model-based and joint randomization approaches

FACTOR	(CATEGORY)	Model Based			Joint Randomization		
		Estimate	SE	p-value	Estimate	SE	p-value
Intercept		1.960	0.30	<.0001	1.653	0.47	<.0001
Family Stat	(51-61)	-0.022	0.16	0.890	-0.078	0.21	0.714
Family Stat	(21-22)	0.187	0.20	0.339	0.117	0.25	0.639
Education PMK	(Other)	<b>-0.245</b>	<b>0.12</b>	<b>0.042</b>	-0.343	0.19	0.072
Education PMK	(Bachelor's+)	<b>-0.393</b>	<b>0.15</b>	<b>0.009</b>	-0.342	0.21	0.103
Education PMK	(Diploma)	-0.103	0.14	0.463	-0.207	0.25	0.406
Hours Daycare	(>50)	0.281	0.31	0.372	0.189	0.41	0.643
Hours Daycare	(1-50)	0.086	0.09	0.343	0.077	0.14	0.572
Working Status	(Not in past 12)	0.104	0.11	0.337	0.280	0.20	0.171

Working Status	(Not but had in)	-0.142	0.14	0.318	-0.046	0.25	0.853
Gender	(Female)	0.016	0.09	0.866	0.135	0.15	0.362
Age		<b>-0.408</b>	<b>0.09</b>	<b>&lt;.0001</b>	-0.276	0.15	0.060
Age^2		<b>0.046</b>	<b>0.01</b>	<b>&lt;.0001</b>	<b>0.038</b>	<b>0.01</b>	<b>0.001</b>
Depress PMK		-0.002	0.01	0.784	-0.009	0.01	0.528
Number of Sibb		<b>-0.119</b>	<b>0.05</b>	<b>0.014</b>	<b>-0.239</b>	<b>0.08</b>	<b>0.004</b>

Table 3.2 Estimated coefficients and results for PA score under both model-based and joint randomization approaches

FACTOR	(CATEGORY)	Model Based			Joint Randomization		
		Estimate	SE	p-value	Estimate	SE	p-value
Intercept		4.863	0.32	<.0001	4.844	0.57	<.0001
Family Stat	(51-61)	<b>0.433</b>	<b>0.19</b>	<b>0.024</b>	0.192	0.22	0.378
Family Stat	(21-22)	-0.307	0.19	0.115	-0.491	0.30	0.098
Education PMK	(Other)	0.007	0.12	0.955	0.023	0.18	0.897
Education PMK	(Bachelor's+)	-0.213	0.13	0.108	-0.277	0.19	0.140
Education PMK	(Diploma)	0.018	0.14	0.899	0.039	0.23	0.866
Hours Daycare	(>50)	0.720	0.48	0.130	<b>1.232</b>	<b>0.63</b>	<b>0.050</b>
Hours Daycare	(1-50)	-0.016	0.09	0.863	0.092	0.16	0.573
Working Status	(Not in past 12)	0.003	0.11	0.979	-0.015	0.21	0.942
Working Status	(Not but had in)	-0.284	0.15	0.066	-0.361	0.23	0.122
Gender	(Female)	<b>-0.269</b>	<b>0.09</b>	<b>0.004</b>	-0.046	0.13	0.725
Age		<b>-0.844</b>	<b>0.09</b>	<b>&lt;.0001</b>	<b>-0.856</b>	<b>0.16</b>	<b>&lt;.0001</b>
Age^2		<b>0.048</b>	<b>0.01</b>	<b>&lt;.0001</b>	<b>0.050</b>	<b>0.01</b>	<b>&lt;.0001</b>
Depress PMK		0.010	0.01	0.223	0.022	0.01	0.120
Number of Sibb		0.052	0.06	0.406	-0.045	0.11	0.674

Table 3.1 shows that under the model-based approach, children with corresponding PMK who have a bachelor degree or above have a comparatively lower EDA score than those with corresponding PMK who have diploma or some college level of education. Under model-based and joint randomization approaches, EDA score decreases as children's age increases for those children between 2 to 4 years old; however, the EDA score increases as age increases for children between 4 to 11 years old. Table 3.1 also shows that under both approaches, children with more siblings tend to have significantly (5%) lower EPA score. An interesting phenomenon occurs with the explanatory variable number of siblings. Although the estimated standard error does indeed increase here, this variable remains significant, and its p-value decreases as well. This is due to the fact that the absolute value of the point estimate itself doubled when changing from model-based to joint randomization approach. This indicates that, at least for the EDA score, the design is informative and should not be ignored.

Table 3.2 shows that under both approaches, the PA score decreases as children's age increases for those children between 2 to 9 years old; however, the PA score increases as age increases for children between 9 to 11 years old. The variables age and age<sup>2</sup> are both significant (5%) under both methods of analysis. In the model-based analysis, children who live with one biological parent have a significantly (5%) higher PA score than those who live with both biological parents. However, in the joint randomization analysis this difference is not significant. Under model-based approach, girls have significantly lower PA score than boys; yet, this difference is not significant under joint randomization analysis. It is interesting to note that the point estimate for females is about 6 times lower in the joint randomization analysis than in the model-based analysis. The magnitude of the coefficient for category 50 or more hours in daycare roughly doubled when switching from model-based to joint randomization analysis. The categories become significantly different from the baseline (less than 50 hours) in this approach. Thus, the design for the PA score is also informative and should be taken into account for inference.

Note that the standard error of the joint randomization estimate is always larger than that of model-based estimate. This leads to different conclusions under model-based and joint-randomization analyses, which further confirms that the design is informative and should not be ignored.

#### 4. CONCLUDING REMARKS

We found that the point estimates are generally different, and sometimes remarkably different, under the two approaches. Additionally, under complex survey, the variance obtained by joint randomization approach is usually larger than the one derived from the model-based approach, which is what we have observed in this study. According to Chen *et.al* (2004), the joint randomization variance is dominated by the design variance (of order  $O(N^2/n)$ ) over the model variance (of order  $O(N)$ ), where  $N$  indicates the population size, and  $n$  represents the sample size. Therefore, inferences based on only the

design component will generally remain valid. Joint randomization approach is preferred over the model-based approach in the analysis of complex survey data such as data from NLSCY.

An alternative approach is to model subject-specific effects by fitting linear mixed-effects models. Results obtained from the linear mixed model are similar to those produced under the GEE model based approach. However, the within subject correlation in this data is low (less than 0.02), and consequently leads to interpretation difficulties.

The analysis in this study is based solely on the complete cases. Although available-case analysis is more efficient than complete-case analysis, when the correlations among explanatory variables are low, available-case and complete-case studies will give similar results due to large sample size (Little and Rubin, 2002). In this study, we further assume that data are missing at random (MAR) and claimed that modeling of the dropout process is not necessary. However, the reasons for dropout are varied in practical settings, and MAR assumption is difficult to justify. A sensitivity analysis should be conducted to explore the impact of deviations of MAR assumption to not missing at random assumption (Molenberghs *et al.*, 2004). For example, logistic regression model may be fit to examine whether modeling the dropout process is necessary.

A limitation of our results is that our model depends on the specification of the correct mean structure. However, both EDA and PA scores appear to be zero-inflated. As a result, the mean structure we used may be extended to accommodate the mass of zero observations. One possible solution is to assume a Tobit model for both the EDA and PA scores.

### ACKNOWLEDGEMENTS

This paper is written based on our presentation at the 2005 SSC annual meeting for the case study on NLSCY survey data. We sincerely thank Dr. Peter Song and Dr. Changbao Wu for their inspiring guidance and financial support. We would like to express our gratitude to Dr. Peggy Ng for organizing the case study.

### APPENDIX

Some explanatory variables are recoded as follows:

Variable Names	Recoded Level	Categories of the Original Variables
Family status	1	11 (both biological parents)
	2	51-61 (single parent)
	3	21-22 (1 biological and 1 step-parent)
Education of PMK	1	Some trade, technical, vocational school, business college, community college, CFGEP, nursing school or university.
	2	Diploma or certificate
	3	Bachelor degree or above
	4	Others or NA
Hours of Daycare	1	Less than 1 hour per week
	2	1-50 hours per week
	3	More than 50 hours per week
Working Status	1	Currently working
	2	Not currently working but had at least one job in the last 12 months
	3	Did not work in the past 12 months

### REFERENCES

Binder, D.A. and Roberts, G.R. (2003). Design-Based and Model-Based Methods for Estimating Model Parameters, In *Analysis of Survey Data* (Chambers, R.L. and Skinner, C.J., eds), Ch. 3. Chichester: Wiley.

Jiahua Chen, Mary E. Thompson and Changbao Wu (2004). Estimation of Fish Abundance Indices Based on Scientific Research Trawl Surveys. *Biometrics*, **60**, 116-123.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Molenberghs, G., Thijs, H., Jansen, I., Beunkens, C., Kenward, M. G., Mallinkrodt, C. and Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**; 445-464.