

ANALYSIS OF THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH DATA: DESIGN-BASED APPROACHES

Yuanyuan Liang¹ and Qiaohao Zhu²

ABSTRACT

The paper is motivated by a case study of the National Longitudinal Survey of Children and Youth data presented at the Statistical Society of Canada's annual meeting in 2005. In longitudinal surveys, a certain level of non-responses is virtually certain to occur. In this paper, we will focus on the design-based approach to study the children's behavior and care over time and deal with the missing data problem by using imputation, bootstrapping, and adjustment longitudinal weights approaches.

KEY WORDS: Bootstrap Percentile Confidence Interval, Bootstrap weights, Design-based estimation, Imputation, Longitudinal weights, Missing data.

RÉSUMÉ

Cet article est motivé par une étude de cas utilisant les données de l'enquête longitudinale sur les enfants et les jeunes présentée au congrès annuel de la société statistique du Canada en 2005. Dans les enquêtes longitudinales, il existe toujours un certain niveau de non réponse. Dans cet article, on se concentrera sur l'approche selon le plan pour étudier le comportement et la protection des enfants au fil du temps et tenir compte du problème des données manquantes par les approches d'imputation, de réplification et d'ajustement des poids longitudinaux.

MOTS CLÉS : Données manquantes; estimation selon le plan; intervalle de confiance Bootstrap du percentile; poids Bootstrap; poids longitudinaux.

1. INTRODUCTION

The National Longitudinal Survey of Children and Youth (NLSCY), launched by the federal government, has been following a representative sample of children from Canada once every two years since 1994. It provides a unique opportunity to study the progress of children from infancy to adulthood. In this case study, a sub-sample of the synthetic data file from the NLSCY is used. The children in the study (1) were 2-5 years old during cycle 1; (2) had a non-zero longitudinal weight during cycle 1; (3) had a monotonic response pattern from cycle 1 through cycle 4; and (4) did not change their province of residence for cycles where they were respondents.

All together, the synthetic database has 1033 records, one record per child. Data are collected during each of the four cycles. Variables of interest are the Anxiety Score, Aggression Score, and Number of Hours in Daycare. Several covariates are also collected. Some of them change over time, for example, Age and Family Status. Some of them are constant over time, like Gender and Province of Residence. The two types of non-responses are total or unit non-responses (when no information is collected on a sampled unit) and partial or item non-responses (when the absence of information is limited to only some variables). The objective of this study is to study the children's behavior and care over time while dealing with the missing data problem.

This paper is organized into three sections. The first part presents a profile of the variables in the dataset. The next part turns to the design-based approach to study the children's development over time and to link the selected environmental variables such as the Numbers of Hours in Daycare, Family Status and Number of Siblings to child

¹ Yuanyuan Liang, Department of the Mathematical and Statistical Sciences, University of Alberta, Canada, T6G 2G1, yliang@ualberta.ca

² Qiaohao Zhu, Department of the Mathematical and Statistical Sciences, University of Alberta, Canada, T6G 2G1, qzhu@stat.ualberta.ca

outcomes such as Anxiety Score and Aggression Score. In the last section, we summarize our conclusions and make suggestions for future studies.

2. DATA TRANSFORMATION AND CLEANING

¹ Yuanyuan Liang, Department of the Mathematical and Statistical Sciences, University of Alberta, Canada, T6G 2G1, yliang@ualberta.ca

² Qiaohao Zhu, Department of the Mathematical and Statistical Sciences, University of Alberta, Canada, T6G 2G1, qzhu@stat.ualberta.ca

In the study, child behaviors were measured by Anxiety Score and Aggression Score during each of the four cycles. The Anxiety Score is measured on an ordinal scale. For 2-3-year-old children, it ranges from 0-12, while, for 4-11-year-olds, it ranges from 0-14, because in the questionnaire for this age group has one more question. For 2-3-year-old, the Anxiety Score was available only during cycle 1 due to the longitudinal nature of the survey, while for 4-11-year-olds, the Anxiety Score was taken during each of the four cycles.

In our dataset, we have variables called ABECs03 and Aemot411. They are both the Anxiety Score during cycle 1 for 2-3-year-old and 4-11-year-olds, respectively. We notice that there are many “not applicable” (coded as 96) for these two variables in the dataset. In fact, at the time when a child first entered the study, he/she belonged to one of the two age groups and could have only one measurement of either ABECs03 or Aemot411. If ABECs03=1, then the corresponding Aemot411 must be 96 (not applicable). Therefore, it is best to create a new variable to represent the Anxiety Score during cycle 1 by combining these two variables. We also notice that the Anxiety Score has different scales for the two age groups. In order to combine them, we need to standardize them first. Therefore, we divide the score by its maximum scale. Finally, the new Anxiety Score at cycle 1 is defined as follows:

$$Aemot1 = \begin{cases} ABECs03/12 & \text{if } Aemot411 = 96 \\ Aemot411/14 & \text{if } Aemot411 \neq 96 \end{cases}$$

The Anxiety Score for other cycles will also be rescaled to 0-1. Table 1 shows an example of how to combine Anxiety Scores during cycle 1. Two major advantages of combining two cycle1 data are (1) replacing the “not applicable” responses with meaningful values and (2) defining a uniform baseline measurement for all children in the study.

Table 1 – Example of combining Anxiety Score during cycle 1

Case Number	PERSRUK	AMMCQ01	ABECs03	Aemot411	Aemot1
3	01000000017105	2	1	96	1/12
17	01000000020903	4	96	1	1/14

A similar transformation is applied for the Aggression Score. We combine the cycle 1 data for two age groups and rescale the new Aggression Score to 0-1.

The Number of Hours in Daycare is a continuous variable, with a range of 1-168, and 996 (did not go to daycare). We define a dummy variable indicates whether a child went to daycare or not.

In our dataset, most covariates are categorical variables and could take several possible values. For simplicity, we combine certain groups together and reduce the total number of categories. In addition, we avoid the possible problem of no or few observations in a certain category caused by scarce observations. For Family Status, instead of considering 21 categories, we classify them into three groups: living with both parents, single parent, or without parents. For Number of Siblings, we define two groups: with siblings and no siblings. The Education of PMK is divided into 3 groups: undergraduate, postgraduate and other education.

For the Urban/Rural and current Working Status variables, the codes for cycle 1 are different from those for other cycles. Fortunately, after comparing the codes carefully, we are able to recode the data for cycle 1 and make them consistent through out all four cycles. It reminds us that data clearing plays a very important role and is crucial in the survey data analysis.

3. DESIGN-BASED DATA ANALYSIS

In this section, we will find design-based estimates for the three response variables: Aggression Score, Anxiety Score and Number of Daycare Hours, and examine the effects of covariates on these three variables. First, we have to deal with missing values.

3.1 Imputation and Weight Adjustment

The two types of non-responses are total or unit non-responses (when no information is collected on a sampled unit) and partial or item non-responses (when the absence of information is limited to only some variables). For the first type, we use the hot-deck imputation method to impute the missing response. Since every subject has response values for gender, age and province, we match the subject with a missing response item to the subjects that have the same gender, age and live in the same province and have a response for the item. Those matched subjects are called “donors”. We then randomly select one donor and replace the missing response with the response from the selected donor. For the second type, we use the weight adjustment method to modify the sampling weights for all subjects. We use the following formula (Lohr, 1999):

$$Adjusted\ weight = \frac{\sum_{\text{Respondents+non-respondents}} Weights}{\sum_{\text{Respondents}} Weights} \bullet sampling\ weight\ at\ cycle\ 1$$

3.2 Design-Based Estimates of the Anxiety Score, Aggression Score and Number of Hours in Daycare

Suppose y_i is the response value from the i^{th} subject, w_i is the sampling weight, $\hat{N} = \sum_{i \in S} w_i$, then, the design-based estimator for the mean value is $\hat{\mu} = \frac{1}{\hat{N}} \sum_{i \in S} w_i y_i$, and the variance estimate can be computed by using bootstrap method.

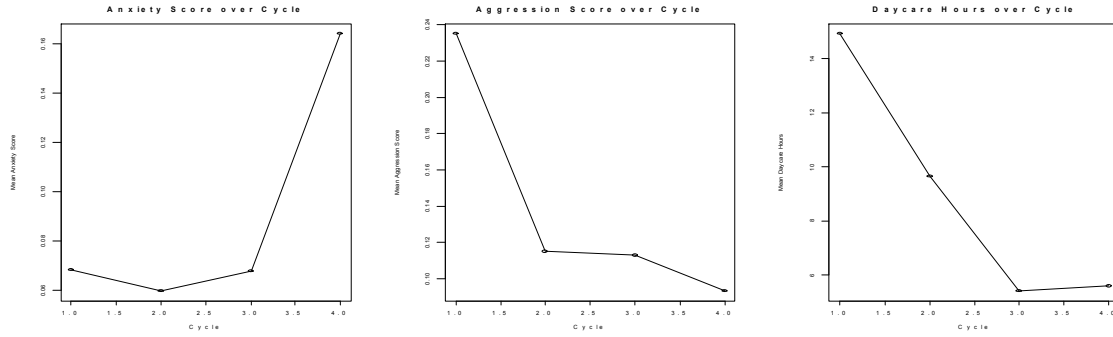
Using the above estimator, we find the means and variances for the Anxiety Score, Aggression Score and the Number of Daycare Hours. The estimates are listed in Table 2:

Table 2 - Estimates of Anxiety Score, Aggression Score and Hours in Daycare

Response Variables	Statistics	Cycle 1	Cycle 2	Cycle 3	Cycle 4
Anxiety Score	Mean	0.0683	0.0598	0.0679	0.1640
	Variance	2e-5	3e-5	4e-5	8e-5
	Count	1033	970	911	810
Aggression Score	Mean	0.2350	0.1148	0.1129	0.0932
	Variance	1.1e-4	5e-5	6e-5	4e-5
	Count	1033	970	911	810
Hours in Daycare	Mean	14.8903	9.6508	5.4266	5.6008
	Variance	0.9389	0.6458	0.3460	1.1340
	Count	1033	970	911	810

To see the overall trend from cycle 1 to cycle 4, we plot the mean estimates in the following graph (Figure 1). We see that as the children become older, their Anxiety Score increases, but their Aggression Score decreases, and the Number of Daycare Hours becomes less.

Figure 1: Mean Anxiety Score, Mean Aggression Score, and Mean Hours in Daycare for 4 Cycles



3.3 Effects of Covariates on Anxiety Score, Aggression Score and Number of Daycare Hours.

Since all the covariates are categorical or ordinal variables, it would be difficult (in terms of methodology and computation) to use a complicated model to test the effects of covariates, especially under the design-based approach. In this section, we propose two different methods to examine the marginal effects of each covariate on the three response variables under the design-based framework. The first method is a parametric method, under the normality assumption, and the second method is a non-parametric method carried out by applying bootstrap confident interval. The testing procedure is as follows:

- Classify the whole sample into several groups according to the values of the covariate.
- Use the design-based approach to estimate the mean and variance of each of the three response variables for each group.
- For each response variable, test the difference of the group means by using the parametric and non-parametric method.
- If there is any significance in the group means, then we can conclude that the covariate has a significant effect on the response variable.

For the parametric test, we use the following model assumptions:

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where y_{ij} is the j^{th} subject in the i^{th} group, μ_i is the mean of the i^{th} group, and ε_{ij} is the residual, and we assume $\varepsilon_{ij} \sim N(0, \sigma^2)$, and all the ε_{ij} are independent.

Under the above model assumptions, we test H_0 : all μ_i are equal.

For this hypothesis, we use the following test statistics:

$$(1) \quad F = \frac{\sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 w_{ij} / df_1}{\sum_i \sum_j (\bar{y}_{ij} - \bar{y}_{i.})^2 w_{ij} / df_2}$$

where

$$\bar{y}_{i.} = \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}}, \quad \bar{y}_{..} = \frac{\sum_i \sum_j w_{ij} y_{ij}}{\sum_i \sum_j w_{ij}},$$

and $df_1 = a - c$, $df_2 = n - a$, with

$$a = \left(\frac{\sum_j w_{ij}^2}{\sum_i \sum_j w_{ij}} \right) \cdot \frac{n}{\hat{N}}, \quad c = \left(\frac{\sum_i \sum_j w_{ij}^2}{\sum_i \sum_j w_{ij}} \right) \cdot \frac{n}{\hat{N}}, \quad \text{and } \hat{N} = \sum_i \sum_j w_{ij},$$

and w_{ij} are the sampling weights, n is the total sample size. It is easy to show that under H_0 , $F \sim F(df_1, df_2)$. It is obvious when all the w_{ij} 's are equal, the above test reduces to the ordinary ANOVA F-test.

For the non-parametric testing method, we also use the model assumptions, but release the normality assumption. We then use the bootstrap percentile confidence interval method. For each sampled unit, we have the longitudinal weight and 1000 bootstrap weights. We compute the observed F-value from formula (1) by using the longitudinal weights. Denote this observed F-value as F_{obs} . Then, we compute the F-values using the same formula, but replace the longitudinal weights with the bootstrap weights. For 1000 bootstrap weights, we can obtain 1000 F-values. Then, we rank these F-values in an ascending order, and obtain the $100(1 - \alpha)$ percentile of these F-values. We denote this F-value as F^* . We then compare F^* with F_{obs} , if $F_{obs} > F^*$; then, we reject the null hypothesis H_0 .

The test results from cycle 1 are listed in Table 3 (results from cycle 2 to 4 can be obtained from the authors). For the parametric method, we compute the F_{obs} , df_1 , df_2 and the p value by using w_{ij} , the sampling weights. We also compute these values by ignoring the sampling weights, i.e., by letting all sampling weights $w_{ij} = \hat{N}/n$. By doing so, we can compare the effects of using and not using the sampling weights. For this testing method, we make our conclusions based on p value. For the BPCI test, we compute F^* and F_{obs} , and then make our conclusion by comparing these two values.

4. CONCLUSIONS AND FUTURE STUDY

Table 3 shows the testing result for cycle 1. Similar analyses have been done for other cycles (not shown). Table 3 shows that the BPCI is more conservative than the parametric F-test. All the significances under the BPCI test are also significant under the F-test, both using and not using the sampling weights, but significance under the F-test may not be significant under the BPCI test. For the F-test, there are different results from using and ignoring sampling weights; this suggests that ignoring the sampling weights would lead to incorrect conclusions.

We also found that from the F-test with sampling weights, Gender has a significant effect on the Aggression Score from cycles 2 to 4, but at cycle 1, this effect is insignificant. The estimated mean Aggression Scores for male children are higher than female children from all four cycles, but during cycle 1, the two mean scores are very close. This suggests that at a younger age, male and female children tend to have similar Aggression Scores, but when they become older, male children tend to be more aggressive than female children.

For other covariates, the testing results are mixed, none of them are found to be significant through out all four cycles, and would be hard to draw any interesting conclusions. This may be due to our approach of examining each covariate separately. It would be more appropriate to analysis the data by linking the response variables with more than one covariate, and also incorporating the sampling weights. This new approach will be given in our future study.

Table 3 – Testing Results for Cycle 1

Response Variable	Covariates	Cycle 1	
		Using Sampling Weights	Equal Sampling Weights

		F_{obs}	F^*	df_1	df_2	p	F_{obs}	p	df_1	df_2
Anxiety Score	daycare:	0.00	1.16	2.2	1028.6	1.00	0.19	0.66	1	1031
	Family status:	0.19	1.13	2.5	1028.3	0.87	2.52	0.11	1	1031
	number of siblings:	0.52	1.39	2.3	1028.6	0.61	4.77	0.03	1	1031
	PMK education:	0.66	1.77	7.0	1023.8	0.71	0.05	0.99	3	1029
	PMK working status:	0.38	1.45	4.6	1026.2	0.85	0.41	0.66	2	1030
	provinces:	0.62	1.30	11.0	1019.8	0.81	1.32	0.22	9	1023
	For age group:	12.79	9.35	4.5	1026.3	0.00	26.93	0.00	2	1030
	Gender:	0.00	1.08	2.2	1028.6	1.00	0.01	0.92	1	1031
rural/urban areas:	0.55	1.42	5.9	1025.0	0.77	0.77	0.55	4	1028	
Aggression Score	daycare:	1.56	3.14	2.2	1028.6	0.21	3.07	0.08	1	1031
	Family status:	2.94	4.73	2.5	1028.3	0.04	4.96	0.03	1	1031
	number of siblings:	0.06	1.21	2.3	1028.6	0.96	3.45	0.06	1	1031
	PMK education:	1.30	2.03	7.0	1023.8	0.25	0.77	0.51	3	1029
	PMK working status:	0.21	1.65	4.6	1026.2	0.95	0.16	0.86	2	1030
	provinces:	0.55	1.33	11.0	1019.8	0.87	1.37	0.20	9	1023
	For age group:	44.78	31.31	4.5	1026.3	0.00	101.79	0.00	2	1030
	Gender:	0.04	1.39	2.2	1028.6	0.97	3.99	0.05	1	1031
rural/urban areas:	1.59	2.15	5.9	1025.0	0.15	1.00	0.41	4	1028	
Hours in Daycare	daycare:	687.45	400.09	2.2	1028.6	0.00	1590.57	0.00	1	1031
	Family status:	2.67	3.89	2.5	1028.3	0.06	5.67	0.02	1	1031
	number of siblings:	0.33	1.81	2.3	1028.6	0.74	0.87	0.35	1	1031
	PMK education:	0.58	1.72	7.0	1023.8	0.77	0.43	0.73	3	1029
	PMK working status:	2.90	3.20	4.6	1026.2	0.02	7.92	0.00	2	1030
	provinces:	2.33	2.92	11.0	1019.8	0.01	2.27	0.02	9	1023
	For age group:	2.10	2.77	4.5	1026.3	0.07	0.22	0.81	2	1030
	Gender:	0.10	1.41	2.2	1028.6	0.92	3.13	0.08	1	1031
rural/urban areas:	1.51	2.08	5.9	1025.0	0.17	1.72	0.14	4	1028	

5. ACKNOWLEDGEMENTS

Special thanks to Dr. N.G.N. Prasad for his invaluable guidance and continuous encouragement.

6. REFERENCES

- Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer-Verlag.