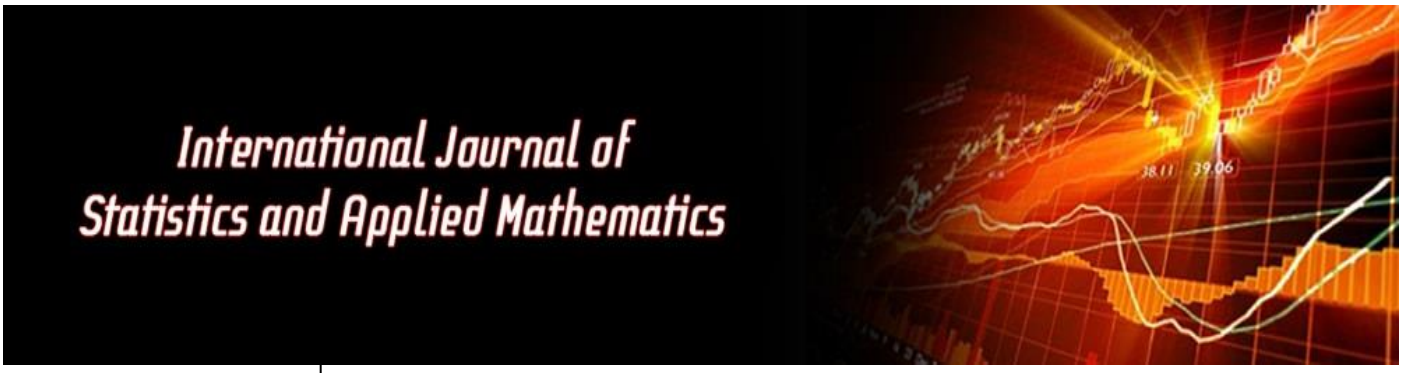
International Journal of Statistics and Applied Mathematics 2024; 9(2): 200-203



ISSN: 2456-1452
Maths 2024; 9(2): 200-203

www.mathsjournal.com
Received: 18-01-2024
Accepted: 22-02-2024

Rahul Banerjee
Division of Sample Surveys, ICAR-IASRI, Library Avenue, New Delhi, India

Pankaj Das
Division of Sample Surveys, ICAR-IASRI, Library Avenue, New Delhi, India

Bharti
Division of Sample Surveys, ICAR-IASRI, Library Avenue, New Delhi, India

Smriti Bansal
Department of Mathematics and Scientific Computing, National Institute of Technology, Hamirpur, Himachal Pradesh, India

Ankita
Birsa Agricultural University, Kanke, Ranchi, Jharkhand, India

Sarita Devi
Department of Basic Sciences, Dr. YSP UHF, CoHF Thunag, Himachal Pradesh, India

Sanghamitra Pal
Department of Statistics, West Bengal State University, Berunanpukhuria, West Bengal, India

Tauqueer Ahmad
Division of Sample Surveys, ICAR-IASRI, Library Avenue, New Delhil, India

Corresponding Author:
Bharti
Division of Sample Surveys, ICAR-IASRI, Library Avenue, New Delhi, India

Prediction approach in repeated measurement surveys: A methodological exploration

Rahul Banerjee, Pankaj Das, Bharti, Smriti Bansal, Ankita, Sarita Devi, Sanghamitra Pal and Tauqueer Ahmad

DOI: https://dx.doi.org/10.22271/maths.2024.v9.i2c.1715

Abstract
In biological and life sciences, including fields like agriculture and medicine, we frequently encounter data with repeated measures. Repeated measures indicate that measurements have been taken on the same individual unit multiple times, either over time or across space. If a population contains repeated measures, there will necessarily be correlation within that population. Analyzing data with a repeated measures structure requires special consideration because it can invalidate standard analysis of variance techniques. This project investigates a prediction approach that has not been previously explored in the presence of intraclass correlation within the population. In this study, we attempt to predict the population total by drawing samples from a repeated measures population using Probability Proportional to Size with Replacement (PPSWR). The prediction approach outlined by Brewer (1963) and Royall (1970) is employed. The estimates of variance (σ^2) and intraclass correlation coefficient (ρ) are obtained through analysis of variance (ANOVA) by fitting a one-way random effects model and equating the mean squares (MS) to the expected mean squares (EMS).

Keywords: Repeated measures, prediction approach, intraclass correlation, analysis of variance (ANOVA), Proportional to Size with Replacement (PPSWR), Random effects model.

Introduction
Agricultural research heavily relies on data with repeated measures (Liu *et al*., 2019) [4]. This refers to situations where multiple measurements are collected on the same unit. The term "unit" can encompass various entities depending on the study design, such as an experimental unit in a designed experiment, a sampling unit in a survey, or a subject in a retrospective study (Field *et al*., 2012) [2]. Repeated measures are typically collected over time to track changes, but spatial measurements can also be included. In this paper, we will use "point" to refer to either a point in time or space.

The ubiquity of repeated measures data in agriculture is evident across various fields. A common example involves growth measurements of plants or animals monitored over time (Carmer *et al*., 1989) [1]. Other examples of repeated measures over time include: Crop yields from repeated harvests on experimental plots (Littell *et al*., 2009) [5]. Daily milk production from individual cows (Wolfinger, 1998) [6]. Weekly livestock prices at specific auction markets. Spatial repeated measures can include: Soil moisture content at various depths in core samples. Pollutant concentrations measured at multiple points on a line transect. Spray deposition amounts at various locations within citrus trees. For consistency throughout this paper, we will primarily use terminology associated with repeated measures in time. Additionally, "unit" will refer specifically to the sampling unit.

Analyzing repeated measures data requires special consideration due to the inherent correlation structure within the data for each unit. This correlation structure can invalidate standard analysis of variance (ANOVA) techniques (Field *et al*., 2012) [2]. For balanced data (meaning all units have complete data at identical time points), multivariate ANOVA methods can be employed. Alternatively, adjustments can be applied to univariate methods to account for the correlation structure (Kutner *et al*., 2005) [3].

Balanced repeated measures data are traditionally analyzed using the split-plot in time ANOVA, also known as the univariate repeated measures ANOVA (Littell *et al*., 2009) [5].

Data Description

This study examines the exercise component of a larger investigation into the effects of nutrition and exercise on physical strength in geriatric citizens. Three weight training programs were implemented, with subjects randomly assigned to each program. Subject strength, measured by the amount of weight lifted, was assessed every other day for two weeks. The first program served as a control group (CONT) with no weight training. The second program (R1) utilized a weight training system where the number of exercise repetitions progressively increased over time. In the third program (W1), the weight lifted was incrementally increased throughout the study. Systolic blood pressure was also measured repeatedly for each subject. Since a significant correlation exists between blood pressure and subject weight (the study variable), blood pressure can be considered an auxiliary variable in the analysis. Subject height represents the size measure variable. This information will be utilized for drawing samples using Probability Proportional to Size with Replacement (PPSWR) from the CONT, R1, and W1 populations, respectively.

Methodology

This exercise therapy study exemplifies a common type of research design known as a repeated measures design. In such designs, subjects are randomly assigned to different "treatment" groups. A response variable of interest is then measured repeatedly over time for each subject. A model for data from this type of study is.

$$y_{ijk} = \mu_{ik} + \epsilon_{ijk}$$

where, y_{ijk} is the response of the j^{th} subject in the i^{th} treatment at the k^{th} time, and μ_{ik} is the population mean for treatment i at time k. The errors ε_{ijk} are assumed normally distributed with mean zero and $V(\varepsilon_{ijk})=\mathbf{V}$, where $\varepsilon_{ijk} = [\varepsilon_{ij1} \quad \varepsilon_{ij2} \quad \cdots \quad \varepsilon_{ijt}]'$. A key issue with repeated measures data is the structure of the covariance matrix V. Here, are five particular structures in terms of mathematical conditions on $\sigma_{kk'}$ the element in row k, column k', of **V**, that play a role in repeated measures data.

Table 1: Show structure, restriction and mathematical condition

S. No.	Structure	Restriction	Mathematical Condition
1.	Unstructured	No Restrictions	$\mathbf{V} = (\sigma_{kk'})$; Positive Definite
2.	Spherical	Equal variances; Zero Covariances	$\sigma_{kk} = \sigma^2$; $\sigma_{kk'} = 0$
3.	Compound symmetric	Equal Variances; Equal Covariances	$\sigma_{kk} = \sigma^2$; $\sigma_{kk'} = \delta\sigma^2$
4.	Huynh-Feldt	Unrestricted Variances; Restricted Covariance	$\sigma_{kk} = 2\tau_k + \phi$; $\sigma_{kk'} = \tau_k + \tau_{k'}$
5.	Autoregressive	Covariance Function of Time Interval between Repeated Measures	$\sigma_{kk'} = \Theta_{\lvert k-k' \rvert}$

Let, the three Populations *viz*. CONTD, R1 and W1 be modelled as.

$$Y_i = \beta X_i + \varepsilon_i; i = 1(1)N$$

Where, X_i's are >0 values and are non-stochastic and ε_i's are random variables such that, $E_m(\varepsilon_i) = 0, Var_m(\varepsilon_i) = \sigma_i^2, Cov_m(\varepsilon_i, \varepsilon_j) = \rho\sigma_i\sigma_j, \forall i = 1,2,\ldots,N; and\ i \neq j = 1,2,\ldots,N.$
Here, E_m, Var_m and Cov_m are model dependent Expectation, Variance and Covariance Operators respectively. The Population total $Y = \sum_{i-1}^{N} Y_i$ can be bifurcated into two components *viz.*

$$Y = \sum_{i=1}^{N} Y_i = \sum_{s} Y_i + \sum_{r} Y_i$$

where, $\sum_s Y_i$ is the part contained in the sample and $\sum_r Y_i$ is the remaining component not covered in the sample and is the value of a random variable i.e. needs to be predicted.

Results

For Control (CNTD) Population Size, N=140, Sample Drawn by (PPSWR) Size n=50, the sample obtained is as follows:

Table 2: Sample Obtained for CNTD Population

Ys	Xs	Ys/Xs	Ys	Xs	Ys/Xs
80	93	0.860215	83	130	0.638462
78	113	0.690265	81	103	0.786408
84	125	0.672000	79	88	0.897727
80	120	0.666667	84	118	0.711864
76	132	0.575758	79	89	0.887640
79	85	0.929412	82	122	0.672131
76	118	0.644068	82	114	0.719298
77	133	0.578947	81	109	0.743119
79	102	0.774510	87	91	0.956044
77	123	0.626016	76	121	0.628099
82	87	0.942529	83	124	0.669355
84	107	0.785047	79	89	0.887640
83	94	0.882979	77	105	0.733333
78	127	0.614173	81	86	0.941860
80	101	0.792079	86	102	0.843137
78	120	0.650000	79	107	0.738318
85	89	0.955056	76	87	0.873563
77	118	0.652542	84	107	0.785047
82	123	0.666667	77	134	0.574627
78	105	0.742857	74	121	0.611570
80	95	0.842105	81	122	0.663934
80	93	0.860215	78	100	0.780000
80	106	0.754717	83	95	0.873684
78	110	0.709091	79	85	0.929412
80	85	0.941176	82	120	0.683333

Based on this sample the values of t_0 and M_0 for Contd Population are as follows.

$$\hat{\beta} = \frac{1}{n}\sum_{s}\frac{Y_s}{X_s} = 0.76$$

$$t_0 = \sum_{s} Y_i + \hat{\beta}\sum_{r} X_r = 4004 + (0.76 * 9661) = 11346.36$$

$$M_0 = Var_m(t_0 - Y)^2 = E_m(t_0 - Y)^2 = \frac{N^2}{n}(1-f)\frac{\bar{X}\overline{x_r}}{\bar{x}}\sigma^2 = 213600.57$$

For R1 Population Size, N=112, Sample Drawn by (PPSWR) Size n=35, the sample obtained is as follows:

Table 3: Sample Obtained for R1 Population

Ys	Xs	Ys/Xs	Ys	Xs	Ys/Xs
79	122	0.64754	80	87	0.91954
83	89	0.93258	79	101	0.78218
80	130	0.61539	82	85	0.96471
76	115	0.66087	82	91	0.90110
77	122	0.63115	76	95	0.80000
84	85	0.98824	79	116	0.68103
76	97	0.78351	87	110	0.79091
79	123	0.64228	78	132	0.59091
78	118	0.66102	86	120	0.71667
79	126	0.62698	82	109	0.75229
83	97	0.85567	82	119	0.68908
83	120	0.69167	78	88	0.88636
84	111	0.75676	86	113	0.76106
78	112	0.69643	88	120	0.73333
75	91	0.82418	86	130	0.66154
77	85	0.90588	75	111	0.67568
80	120	0.66667	75	113	0.66372
78	126	0.61905			

Based on this sample the values of t_0 and M_0 for R1 Population are as follows.

$$\hat{\beta} = \frac{1}{n}\sum_{s}\frac{Y_s}{X_s} = 0.75$$

$$t_0 = \sum_{s} Y_i + \hat{\beta}\sum_{r} X_r = 2810 + (0.75 * 8376) = 9092$$

$$M_0 = Var_m(t_0 - Y)^2 = E_m(t_0 - Y)^2 = \frac{N^2}{n}(1-f)\frac{\bar{X}\overline{x_r}}{\bar{x}}\sigma^2 = 366381.03$$

For W1 Population Size, N=91, Sample Drawn by (PPSWR) Size n=30, the sample obtained is as follows:

Table 4: Sample Obtained for W1 Population

Ys	Xs	Ys/Xs	Ys	Xs	Ys/Xs
84	117	0.71795	75	116	0.64655
74	120	0.61667	81	97	0.83505
83	87	0.95402	82	125	0.65600
82	122	0.67213	80	89	0.89888
81	121	0.66942	81	105	0.77143
83	133	0.62406	83	116	0.71552
85	112	0.75893	76	122	0.62295
75	114	0.6579	81	128	0.63281
87	111	0.78378	87	90	0.96667
80	127	0.62992	83	96	0.86458
79	95	0.83158	75	128	0.58594
89	86	1.03488	83	134	0.61940
81	125	0.64800	83	87	0.95402
82	103	0.79612	76	120	0.63333
80	118	0.67797	76	120	0.63333

Based on this sample the values of t_0 and M_0 for W1 Population are as follows.

$$\hat{\beta} = \frac{1}{n}\sum_{s}\frac{Y_s}{X_s} = 0.74$$

$$t_0 = \sum_{s} Y_i + \hat{\beta}\sum_{r} X_r = 2427 + (0.74 * 6538) = 7265.12$$

$$M_0 = Var_m(t_0 - Y)^2 = E_m(t_0 - Y)^2 = \frac{N^2}{n}(1-f)\frac{\bar{X}\overline{x_r}}{\bar{x}}\sigma^2 = 259863.088$$

To estimate the variance σ^2 and the intraclass correlation coefficient ρ, ANOVA estimators were applied to all three populations. The results are outlined below:

Estimation of σ and ρ by ANOVA estimators

For Control Population, NM=140, 20 clusters each of size 7 we draw a sample of 10 clusters each of size 7 by PPSWR.

Table 5: Show mean squares and expected mean square

Source	Mean Squares	Expected Mean Square
Between Cluster	$B = \frac{1}{n-1}\sum_{i\epsilon s} m(\overline{Y_{si}} - \bar{Y})^2 = 60.66$	$\sigma^2[1 + (m-1)\rho] = 60.314$
Within Cluster	$W = \frac{1}{n}\sum_{i\epsilon s}\sum_{j\epsilon s_i}\frac{1}{m-1}(Y_{ij} - \overline{Y_{si}})^2 = 0.80$	$\sigma^2(1-\rho) = 0.799$

Thus, $\rho = 0.914$, $\sigma^2 = 9.302$

For R1 Population, NM=112, 16 clusters each of size 7 we draw a sample of 8 clusters each of size 7 by PPSWR.

Table 6: Show between cluster and within cluster

Source	Mean Squares	Expected Mean Square
Between Cluster	91.71	91.224
Within Cluster	22.89	22.88

Thus, $\rho = 0.299$, $\sigma^2 = 32.65$

For W1 Population, NM=91, 13 clusters each of size 7 we draw a sample of 6 clusters each of size 7 by PPSWR.

Table 7: Show mean squares and expected mean square NM=91, 13 clusters each

Source	Mean Squares	Expected Mean Square
Between Cluster	181.64	150.61
Within Cluster	1.873	1.872

Thus, $\rho = 0.919$, $\sigma^2 = 23.12$

Conclusion

This study investigated the application of a prediction approach for data with repeated measures. We demonstrated that this method provides a viable strategy for estimating population parameters, including population means and totals. Additionally, the approach allows for the estimation of key statistical measures like standard deviation (σ) and intraclass correlation coefficient (ρ) through analysis of variance (ANOVA) estimators. Our findings highlight the potential benefits of this method for analyzing data with correlated populations. Compared to traditional approaches that may be rendered invalid by the presence of correlation, this prediction approach offers a reliable and potentially more accurate alternative.

Acknowledgement

The authors are thankful to the ICAR-IASRI and Department of Statistics, West Bengal State University for providing facilities for carrying out the present research. The authors are also thankful to editor of journal and the anonymous reviewers for their constructing insights in uplifting the quality of the study.

References

1. Carmer SG, Swanson MR. Balanced fixed effects and mixed effects models in environmental studies. Statistics in Ecology & Environmental Science; c1989.
2. Field A, Miles J, Glover N. Discovering Statistics Using R. Sage Publications Ltd; c2012.
3. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied Linear Statistical Models. McGraw-Hill; c2005.
4. Liu S, Zhang H, Xu J. Repeated measures data analysis: a tutorial. Journal of Educational and Behavioral Statistics. 2019;44(2):125-153.
5. Littell RC, Henry PR, Ammann CB. Stats4: An Introduction to Statistical Analysis and Design. SAS Institute; c2009.
6. Wolfinger RD. Approaches to analyzing longitudinal data. Annual Review of Sociology. 1998;24(1):681-706.