Longitudinal Data and Correlated Measures Bias: The Alternative of Mixed Models

Johnnatas Mikael Lopes,[1] *Marcello Barbosa O.G. Guedes,[2]* *Rafael Limeira Cavalcanti,[3]* *Clecio Gabriel de Souza[2]*

Universidade Federal do Vale do São Francisco (UNIVASF) - Colegiado de Medicina,[1] Paulo Afonso, BA – Brazil
Universidade Federal do Rio Grande do Norte (UFRN) – Fisioterapia,[2] Natal, RN – Brazil
Centro Universitário Maurício de Nassau (UNINASSAU) – Fisioterapia,[3] Natal, RN – Brazil

Longitudinal studies have two important data typologies: single outcomes or repeated measures.[1] Single outcome, such as death or disease onset, should have a different data treatment than those studies with repeated measures outcome. But, they have in common the detection of changes over time and the contributing factors for this change. Cohort differs from cross-sectional studies that desire only variables relationship, without causal effect.

Fernandes et al.[2] wrote an article entitled *The Relationship between Lifestyle and Costs Related to Medicine Use in Adults*, published in this journal, volume 112, number 6, 2019, and they used behavioral independent variables to estimate their effects on drug costs outcome, collected as repeated measurements in a prospective cohort design.

The aim of this exposition is to show that, probably, there was a mistake in the Fernandes et al.[2] data analysis, which compromises the causality inferences due to the great possibility of the estimates' accuracy to be mistaken.

Let's get to the facts. Considering the prospective cohort design with repeated measures, there is a hierarchical structure in the outcome data due to their clustering in the same participant after various measures. Data cluster leave to the model error, that is the difference between what was predicted by the model and the actual measurement, of the same participant, at different times, to be correlated.[3] This is a condition for not using multiple linear regression (MLR) which assumes the independence of the model error given by the assumption that the distribution of each participant is equal. MLR does not extract from the data which is variability within the individual from variability between individuals (population).[3]

Keywords

Cohort Studies; Longitudinal Studies; Cross-Section,Studies; Epidemiology; Biostatistics.

Using RLM in repeated measures generates regression coefficients with standard errors biased. This requires covariance matrix application that will produce more reliable estimates, in others words, narrower confidence intervals from Mixed Effects Models.[4] This is the best alternative to verify changes over time or the conditioners effects on repeated measures outcomes in longitudinal studies, controlling for individual effects.

There is greater variability between individuals than within individuals, mainly due to biological and social conditioning differences, it's observed that drug costs will be more correlated over time in the same individual than among participants. To think that this distribution is the same among the participants ignores theoretical assumption in the social determination on people's behavior.[5]

Build distinct MLRs (A, B, C and D), see Fernandes et al.[2], does not control this covariance effect, and therefore may be producing coefficients with confidence intervals biased in independent variables and can not detect the rate of change from basal either.[3] In addition, with mixed models it would also be possible to take advantage of measurements that were measured on lost participants, increasing modeling sensitivity.[4]

From another perspective, the objective of the research being to estimate the interrelation of drug cost and behavioral habits, without establishing causality, would only require a cross-sectional design of the participants with the collection of outcome data and independent variables at a single moment. Thus, the basal regression model would be sufficient to estimate gross and adjusted associations.[1]

Thus, the use of RLM should be restricted to cross-sectional research designs and longitudinal studies with repeated measures outcomes need to differentiate the individual effect of the population effect in the identification of temporal changes and their conditioning. Possibly, the findings of Fernandes el al.[2] should be based regarding their conclusions about the inverse relationship between alcohol use and drug costs or the statistically non-significant relationships with body fat, gender and smoking status that have great impact on other health situations, especially chronic diseases.

References

1. Medronho RA, Bloch KV, Raggio LR, Werneck GL. Epidemiologia. 2ª ed. São Paulo: Atheneu; 2009. 685 p.

2. Fernandes RA, Mantovani AM, Sanches Codogno J, Camilo Turi-Lynch B, Pokhrel S, Anokye N. Relação entre Estilo de Vida e Custos Relacionados ao Uso de Medicamentos em Adultos. Arq Bras Cardiol. 2018;11(2):749–55.

3. Fausto MA, Carneiro M, Antunes CMDF, Pinto JA, Colosimo EA. O modelo de regressão linear misto para dados longitudinais: Uma aplicação na análise de dados antropométricos desbalanceados. Cad Saude Publica. 2008;24(3):513–24.

4. Helena Constantino Spyrides M, José Struchiner C, Tereza Serrano Barbosa Gilberto Kac M. Análise de Dados com Medidas Repetidas. In: Kac G, Sichieri R, Gigante D Epidemiologia nutricional. Rio de Janeiro: Editora FIOCRUZ/Atheneu; 2007.p. 245–60.

5. Garbois JA, Sodre F, Dalbello-Araujo M. Da noção de determinação social à de determinantes sociais da saúde. Saúde Debate. 2017;41(112):63–76.

Reply

I appreciate the opportunity to answer the questions concerning our manuscript recently published in Arquivos Brasileiros de Cardiologia.[1] Academic discussion is always healthy and welcome.

Firstly, thank you for your interest in our study. The question raised refers to the use of linear regression in the treatment of data from a prospective cohort with repeated measures, which is believed to have caused mistaken estimates (mixed linear regression is suggested instead). Linear regression is debated as it fails to detect intra-individual variability properly, as it focuses on variability between individuals. From a theoretical point of view, this statement is correct, but it does not reflect the way the data were analyzed in the study.

The dependent variable of the study was defined as "drug spending over 12 months." In the study, we did not analyze the history of drug spending over the year[2] (and how this history would be affected by behavioral variables), nor did we seek to identify the relationship between changes compared to baseline (for dependent and independent variables). We did try to analyze the relationship of behavioral variables with the final amount spent over the year.

In fact, this dependent variable is unusual in its construction, as it was longitudinally designed (expenditures on drugs computed over 12 months), but treated cross-sectionally (total amount spent over 12 months). The total amount of drug spending reflects a cross-sectional construct, although its construction considers the 12 months of follow-up. This particularity of the dependent variable, added to the fact that the behavioral variables were collected at only two moments (baseline and at the end of 12 months), led us to create the four models proposed in the study, which characterize a cross-sectional view of the problem (especially models A [baseline data] and B [at the end of 12 months]). Unfortunately, the monthly assessment of behavioral variables was not an available methodological option.

In an ideal model, the dependent variable and the independent variables should be collected monthly, allowing to identify the impact of changes on behavioral variables on changes in drug spending history over the year. However, I repeat, this was not the purpose of the study.[1] For this type of analysis, specific structural equation modelling (latent growth curve analysis) would be more suitable (even more so than mixed linear regression), as they would make it possible to analyze the direct impact of changes on independent variable (slope)over changes observed on dependent variable (slope).[3] The "impact" measures generated by the model are easily interpreted, as they can be expressed as correlation coefficients, which additionally provide effect-size measurements.[4]

Additionally, the dependent variable as it was presented (cross-sectionally, with spending accruing over follow-up time) was necessary due to the particularities observed in its structure. Unlike other variables usually measured in different areas of health sciences (height, blood pressure, lipid profile components), which do not have zero value, drug spending occurs irregularly, reflecting the high occurrence of zero values (that is, spending can be reported in the first month of collection, then no spending can be reported over the subsequent months). Against this background, analyses considering the month-to-month variable would be problematic. Likewise, the issue of intra-individual variability needs to be considered with caution in this study because drug spending in the previous month does not recur in the following month, unlike what was observed for variables like height[5] which, even without any gain, the amount of the previous month will repeat in the following month.

Finally, the absence of significant relationships for obesity and smoking is not surprising in this study, as the sample is relatively young, without the presence of chronic diseases and low occurrence of smoking.

Rômulo Araújo Fernandes

References

1. Fernandes RA, Mantovani AM, Sanches Codogno J, Camilo Turi-Lynch B, Pokhrel S, Anokye N. Relação entre Estilo de Vida e Custos Relacionados ao Uso de Medicamentos em Adultos. Arq Bras Cardiol. 2018;11(2):749–55.

2. Lemes ÍR, Fernandes RA, Turi-Lynch BC, Codogno JS, de Morais LC, Koyama KAK, Monteiro HL. Metabolic Syndrome, Physical Activity, and Medication-Related Expenditures: A Longitudinal Analysis. J Phys Act Health. 2019;16(10):830-5.

3. Brailean A, Aartsen MJ, Muniz-Terrera G, Prince M, Prina AM, Comijs HC, Huisman M, Beekman A. Longitudinal associations between late-life depression dimensions and cognitive functioning: a cross-domain latent growth curve analysis. Psychol Med. 2017;47(4):690-702.

4. Maher JM, Markey JC, Ebert-May D. The other half of the story: effect size analysis in quantitative research. CBE Life Sci Educ. 2013; 12(3):345-51.

5. Fausto MA, Carneiro M, Antunes CM, Pinto JA, Colosimo EA. Mixed linear regression model for longitudinal data: application to an unbalanced anthropometric data set. Cad Saude Publica. 2008;24(3):513-24.