General Psychiatry

Analysis of correlated data with feedback for time-dependent covariates in psychiatry research

Elsa Vazquez Arreola ,[1] Jeffrey R Wilson,[2] Ding-Geng Chen[3]

To cite: Vazquez Arreola E, Wilson JR, Chen D-G. Analysis of correlated data with feedback for time-dependent covariates in psychiatry research. *General Psychiatry* 2020;**33**:e100263. doi:10.1136/gpsych-2020-100263

Received 07 June 2020
Revised 05 July 2020
Accepted 08 July 2020

[1]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, Arizona, USA
[2]Department of Economics, W.P. Carey School of Business, Arizona State University, Tempe, Arizona, USA
[3]School of Social Work and Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence to
Dr Elsa Vazquez Arreola; Elsa.Vazquez@asu.edu

ABSTRACT

In studies on psychiatry and neurodegenerative diseases, it is common to have data that are correlated due to the hierarchical structure in data collection or to repeated measures on the subject longitudinally. However, the feedback effect created due to time-dependent covariates in these studies is often overlooked and seldom modelled. This article reviews the methodological development of feedback effects with marginal models for longitudinal data and discusses their implementation.

INTRODUCTION

In the study of psychiatry and neurodegenerative diseases, it is common to have correlated observations. By correlation, one means that the mechanism that gives rise to the observation is not necessarily different from the one that gave rise to another observation. Observations may be correlated due to the hierarchical structure by which the data are obtained or because they are repeatedly measured.

Longitudinal models for psychiatry research

Longitudinal studies are a crucial component of psychiatric research.[1] Some of the most important research questions in psychiatry and mental health investigate symptoms and behaviour change over time with time-dependent factors that influence the development of pathological and normal behaviours.[2] Longitudinal studies are used to investigate mental health disorders, such as depression, schizophrenia, psychosis, bipolar disorder and post-traumatic stress disorder, among others.[1 3]

Longitudinal studies have been used in determining service use among patients with mental health diseases. Hospitalisation for psychiatric reasons and receiving psychiatric crisis services are two outcomes of interest for measuring service use. Such are of importance to help reduce service use and are particularly central for consumer-run organisations as they are usually government-funded.[4] Other longitudinal studies are used to understand the effectiveness of treatments and programmes in improving the quality of life of patients with mental health disorders.[5 6]

As an example, in a longitudinal study used to investigate the success of consumer-run organisations in promoting the mental health of their members, researchers gathered demographic data and obtained information about social support, community integration, personal empowerment, quality of life, symptom distress and service use.[6] Higher levels of social support, community integration and quality of life might decrease the probability of being hospitalised for psychiatric reasons or receiving psychiatric crisis services over time. Increased levels of symptom distress might increase the likelihood of using such services. Further, using such services might, in turn, increase levels of symptom distress in the future, resulting in feedback from the service use outcomes to levels of symptom distress.[7]

The marginal models presented in this paper are used to investigate these longitudinal studies, especially when feedback is involved in the research questions in psychiatry. These marginal models account for the different sources of correlation encountered in longitudinal data.

Types of correlation in longitudinal data

There are different types of correlation in longitudinal data. When analysing longitudinal binary data, it is essential to account for both the correlation inherent from the repeated measures of the responses and the correlation realised because of the feedback created between the responses at a particular time and the covariates at other times. Ignoring any of these correlations can lead to invalid conclusions. Such is the case, for example, when the covariates are time-dependent, and the standard logistic regression model is used.

There are three types of correlations discussed in the paper: responses with

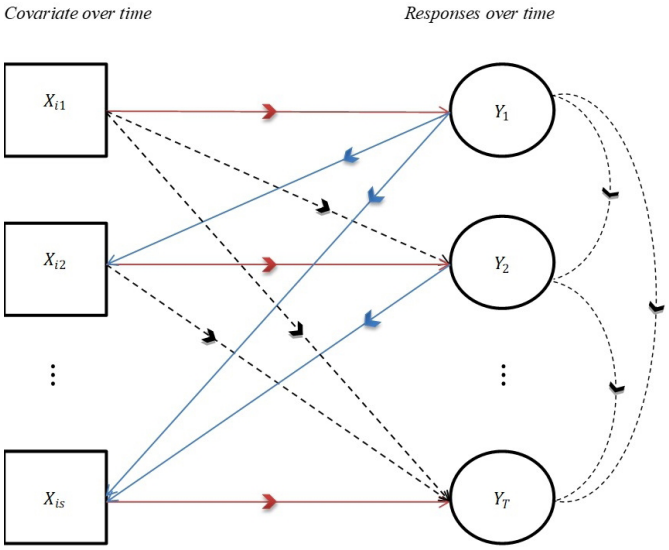

Figure 1 Types of correlation structures.

responses, covariates with current and future responses, and responses with future covariates. A model to address these types of relationships is the aim of this article and the book *Marginal Models in the Analysis of Correlated Data with Time-dependent Covariates*.[8] The different types of correlation presented are shown in figure 1,[8]

1. There are correlations among the responses which are denoted by $Y_1, \ldots, Y_T$ as time t goes from one to T.
2. There are correlations between the covariate X_s and outcome Y_t when covariates at time s impact the outcomes in time t where $s \leq t$. If $s = t$, one refers to these correlations as the direct or cross-sectional effects of the covariate on the outcome. If $s < t$, these correlations are called the lagged effects of the covariate on the outcome.
3. There are correlations between response Y_t and covariate X_s when the outcome in time t impacts the covariate in time s where $s > t$. These correlations are often referred to as feedback effects from Y_t to the futureX_s .

This paper provides an overview of modelling repeated responses with time-dependent, time-independent covariates and feedback effects. With this review, we offer some guidance in analysing correlated data due to repeated measurements.

Correlated models for longitudinal data

There are two common approaches used to analyse longitudinal data: population-averaged (also known as marginal models) and subject-specific models. Population-averaged models focus on understanding what affects the mean outcome of the population, while subject-specific models concentrate on determining what impacts the mean outcome of subpopulations.[9] The basic idea of population-averaged models is that instead of attempting to model the within-subject covariance structure, it is treated as a nuisance, and the focus turns to the marginal mean. In this framework, the covariance structure does not need to be specified correctly for one to get reasonable estimates of regression coefficients and SEs. In contrast, the subject-specific model distinguishes observations belonging to the same or different subpopulations. Random effects are commonly used to estimate the subject-specific models. For repeated responses, the hierarchical logistic regression models are used for multilevel analysis. There are two methods used for estimating subject-specific models, maximum likelihood approach (random-effects models) and the conditional likelihood procedure. This paper focuses on marginal models for longitudinal data.

Population-averaged or marginal model

For longitudinal data, Zeger and Liang[9] proposed the generalised estimating equation (GEE) marginal model. The GEE is an extension of generalised linear models to estimate the population-averaged estimates while accounting for the dependency between the repeated measurements.[10] Specifically, the dependency or correlation between repeated measures is accounted for by a robust estimation of the variances of the regression coefficients. In fact, the GEE approach treats the time dependency as a nuisance, and a 'working' correlation matrix for the vector of repeated observations from each subject is specified to account for the dependency among the repeated observations. The form of 'working correlation' is assumed to be the same for the subjects, reflecting average dependence among the repeated observations over subjects. Several different working correlation structures are possible, including independence, exchangeable, autoregressive and unstructured, to name a few.

A generalised method of moments (GMM) model for longitudinal data that provides reasonable estimates of the marginal regression coefficients and is more efficient than GEE is due to Qu *et al*.[11] However, this model does not distinguish between time-dependent and time-independent covariates. The GMM model for longitudinal data with continuous outcomes is extended to account for time-dependent covariates.[12] Their model estimated regression coefficients for time-dependent covariates by classifying them into three different types, which determined the group of valid moment conditions used in the estimation process.[12] This model is expanded to allow for the modelling of binary longitudinal outcomes in the presence of time-dependent covariates.[13] They provide researchers with the ability of testing for valid moment conditions for time-dependent covariates individually instead of assuming that a group of moment conditions are valid because of the type of time-dependent covariate. Although these marginal models provide reasonable estimates of the regression coefficients, they assume that the effects of time-dependent covariates are the same across time. However, a partitioned coefficient model allows for the estimation of current and future effects of time-dependent covariates on binary and continuous outcomes as discussed in Irimata *et al*.[14] For a thorough

discussion of these models, see *Marginal Models in the Analysis of Correlated Data with Time-dependent Covariates.*[8]

Feedback model with time-dependent covariates

When analysing longitudinal data with time-dependent covariates, there are usually three questions (Qs) of interest that researchers seek to answer[15]:

Q1. What is the cross-sectional relationship/association between the outcome Y_{it} and the covariate X_{ijt} (both *X* and *Y* are measured at the same time)?

Q2. Is the outcome at time t, Y_{it}, affected by the time-dependent covariate measured at a previous time period $t-s$, $X_{ij[t-s]}$; $(s = 1, 2, .., t-1)$ (lagged covariates related/associated with future values of the outcome)?

Q3. Does the outcome at time $t-s$, $Y_{[t-s]}$, associate with the j^{th} time-dependent covariate at time t, X_{ijt} (feedback effect of outcome on future values of the time-dependent covariate)?

A two-stage model allows researchers to answer all three questions simultaneously.[8] This two-stage model accounts for feedback effects while modelling the direct impact, as well as the delayed effect, of the covariates on future responses. However, modelling feedback might not always make sense or be significant. For example, higher levels of social support might decrease the probability of service use, but service use might not affect levels of social support in the future.

Model

Stage 1 of the model allows for the fit of the cross-sectional and the lagged effects of time-dependent covariates on the outcome of interest (Q1 and Q2). Let each covariate $\mathbf{X}_{ij*}$ be measured at times $t = 1, 2, \ldots, T$; resulting for subject *i* and covariate $\mathbf{X}_{ij*} = \left(X_{ij1}, \ldots..X_{ijT}\right)'$. Thus, the model

$$g\left(\mu_{it}\right) = \beta_0 + \beta_j^{tt}X_{ijt} + \beta_j^{[1]}X_{ij[t-1]} + \ldots. + \beta_j^{[T-1]}X_{ij1}$$

with $s \leq t$, so

$$g\left(\mu_i\right) = \mathbf{X}_{ij*}\beta$$

where the $\mathbf{X}_{ij*}^{[*]}$ matrix consists of a column of ones concatenated with a lower diagonal matrix as the systematic component, and $\mu_i = \left(\mu_{i1} \ldots..\mu_{iT}\right)'$ is dependent on the regression coefficients $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}_j\right)$ with $\beta_j = \left(\beta_j^{tt}, \beta_j^{[1]}, ..., \beta_j^{[T-1]}\right)$, where s and t go from 1 to T. The coefficient β_j^{tt} denotes the effect of the covariate $\mathbf{X}_{*jt}$ on the response $\mathbf{Y}_{*t}$ when both are measured at the *t*th time period. However, when $s \neq t$, it does not necessarily follow that one should interpret the past, using two different time periods in the same way as when $\mathbf{X}_{i*t}$ and Y_{it} are in the same time period, $s = t$. The impact of a covariate on the response from a previous time period is not intuitively the same as when they are measured in the same period. This is especially true in health research when time of dose will have impact on the reaction of the patient. Thus, current and future effects should not be combined but rather analysed separately. This is best explained by $\beta_j^{[1]}$ representing the effect of $X_{ij[t-1]}$ on Y_{*t}, and by $\beta_j^{[2]}$ representing the effect of $X_{ij[t-2]}$ on Y_t and so on. In general, one can consider the systematic component consisting of *P* covariates and let $\beta = \left(\beta_0, \beta_1, \ldots, \beta_P\right)'$ be the parameters associated with those covariates, with each β_j having maximum length *T*. Thus, **X** is of maximum dimension NT by $(PT + 1)$ and β is a vector of maximum dimension PT + 1.

Stage 1 of the model, based only on the valid moment conditions, is fitted as

$$\mu_{it}(\beta = \beta_0 + \beta_j^{tt}X_{jt} + \sum_{s=1}^{T-1}\beta_j^{[t-s]}X_{jt[t-s]} \mid_{\text{valid moments}}$$ when the valid moments conditions exist. In this model, β_j^{tt} denotes the regression parameter for the cross-sectional effects of time-dependent covariates on the outcome; these are cases where the moment conditions are always valid (the effect of the covariate in the same period as the response). The coefficient $\beta_j^{[t-s]}$ represents the lagged effects of the covariate on the response when the covariate is measured prior to the outcome $(s < t)$ and the moment conditions are valid.

As an example, in our earlier example, one can determine if the quality of life, social support, community integration and symptom distress measured 6, 12 and 18 months before had a positive or negative association with service use now. These are given by the coefficients in stage 1 of the model.

In stage 2 of the model, the feedback from the outcome to future values of the covariates (Q3) is addressed. The *j*th covariate measured at time *s*, X_{ijs}, is fitted as

$$g\left(\gamma_{ijs}\right) = \alpha_0 + \alpha_j^{ss}Y_{is} + \alpha_j^{[1]}Y_{i[s-1]} + \alpha_j^{[2]}Y_{i[s-2]} + \ldots + \alpha_j^{[s-1]}Y_{i1}$$

where γ_{ijs} is the mean of the covariate X_{ijs} . The, $\alpha_j^{[1]}$ represents the feedback of the outcome $Y_{i[s-1]}$ on the time-dependent covariate X_{ijs} measured immediately in the next time period. The regression coefficient $\alpha_j^{[2]}$ represents the feedback of the outcome on the covariate measured in the two time periods after and so on. As an example, in the study referred to earlier, one may want to investigate whether service use increases levels of symptom distress in the future.

In both stages of the two-stage model, estimates are obtained using GMM after determining valid moment conditions. For modelling the feedback from the outcome to two or more time-dependent covariates, the estimates are obtained through the use of simultaneous GMM. Computing code to fit this model can be found online (https://github.com/ElsaVazquez29/Feedback-Code).

CONCLUSIONS

In psychiatry, the correlation inherent in repeated measures is further affected by the presence of time-dependent covariates. It grossly impedes any interpretations the psychiatrist makes. In particular, the changes and feedback presented when the covariates are time-dependent cannot be ignored. Often, the feedback effects go unchecked.

However, any modelling of longitudinal data must address the impact from the feedback, as well as the immediate and the delayed effects of covariates on the responses. For the aspect of feedback, there is an advantage of using a two-part model to correlated data with time-dependent covariates as it allows one to use GMM methods to identify valid moments.

While there is merit in the models due to Lai and Small,[12] Zhou *et al*,[16] Lalonde *et al*[13] and Irimata *et al*,[14] they do not always account for the feedback. The two-stage GMM model allows one to account for the feedback effects across different time-periods. It partitions the regression coefficients and allows one to identify directional and delayed effects.

We have developed a new approach to marginal regression analysis for time-dependent covariates with feedback. We use the GMM to make optimal use of the estimating equations that are made available by the covariates in both the direct part of the model and the feedback. We have focused on marginal regression analysis. The approach is also useful for obtaining more efficient estimates. This model conditions on part of the past history of covariates and outcomes. The partly conditional model is intermediate between the marginal model that conditions only on the covariates at time t and the transition model that conditions on the full history of covariates and outcomes at time t.

In summary, there exists a correlation when modelling dependent data with time-dependent and time-independent covariates with feedback effects, which cannot be ignored. With this review on the methodological development, we recommend the marginal models as extensively discussed in *Marginal Models in the Analysis of Correlated Data with Time-dependent Covariates*.[8]

Contributors EV and JRW conceived of the presented idea and wrote the manuscript with support from D-GC.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement No additional data are available.

ORCID iD

Elsa Vazquez Arreola http://orcid.org/0000-0003-1960-1159

REFERENCES

1 McInnis MG, Greden JF. Longitudinal studies: an essential component for complex psychiatric disorders. *Neurosci Res* 2016;102:4–12.
2 Anstey KJ, Hoffer SM, designs L. Methods and analysis in psychiatric research. *Aust N Z J Psychiatry* 2004;38:93–104.
3 Youn S-J, Mackintosh M-A, Wiltsey Stirman S, *et al*. Client-level predictors of treatment engagement, outcome and dropout: moving beyond demographics. *Gen Psychiatr* 2019;32:e100153.
4 Ostrow L, Hayes SL. Leadership and characteristics of nonprofit mental health peer-run organizations nationwide. *Psychiatr Serv* 2015;66:421–5.
5 Fleury M-J, Grenier G, Bamvita J-M, *et al*. Predictors of quality of life in a longitudinal study of users with severe mental disorders. *Health Qual Life Outcomes* 2013;11:92.
6 Nelson G, Ochocka J, Janzen R, *et al*. A longitudinal study of mental health consumer/survivor initiatives: part V–Outcomes at 3-year follow-up. *J Community Psychol* 2007;35:655–65.
7 Loch AA, Andrade Loch A. Discharged from a mental health admission ward: is it safe to go home? A review on the negative outcomes of psychiatric hospitalization. *Psychol Res Behav Manag* 2014;7:137–45.
8 Wilson JR, Vazquez Arreola E, Chen D-G. *Marginal models in the analysis of correlated data with time-dependent covariates*. Springer, 2020.
9 Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986;42:121–30.
10 McCullagh P, Nelder JA. *Generalized linear models*. 2nd edn. London: Chapman and Hall, 1989.
11 Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000;87:823–36.
12 Lai TL, Small D. Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *J Royal Statistical Soc B* 2007;69:79–99.
13 Lalonde TL, Wilson JR, Yin J. GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications. *Stat Med* 2014;33:4756–69.
14 Irimata KM, Broatch J, Wilson JR. Partitioned GMM logistic regression models for longitudinal data. *Stat Med* 2019;38:2171–83.
15 Diggle P, Heagerty P, Liang K-Y, *et al*. *Analysis of longitudinal data*. Oxford, United Kingdom: Oxford University Press, 2002.
16 Zhou Y, Lefante J, Rice J, *et al*. Using modified approaches on marginal regression analysis of longitudinal data with time-dependent covariates. *Stat Med* 2014;33:3354–64.

Dr. Elsa Vazquez-Arreola is a biostatistician at the National Institute of Diabetes and Digestive and Kidney Diseases, USA. She obtained a Ph.D. degree in Statistics from Arizona State University, USA. She has consulted on numerous clinical and research studies and is currently a co-author of the book entitled "Marginal models in the analysis of correlated data with time-dependent covariates" with Dr. Jeffrey R. Wilson and Dr. Ding-Geng Chen. Her main research interests include models for correlated data, time-dependent covariates, generalized linear mixed models and propensity score models with applications in public health and behavioural health.