original article

The application of non-linear multilevel models to experience sampling data

Ben Richardson
Deakin University
Matt Fuller-Tyszkiewicz
Deakin University

A great deal of evidence demonstrates that state-based aspects of human functioning, such as moment-to-moment variation in affect, explain important psychological and behavioural outcomes (e.g., Colautti et al., 2011); often over and above more general measures that may be used in cross-sectional designs (Sturgeon & Zautra, 2013). For example, in clinical samples, findings demonstrate that a common feature of many disorders is higher levels of reactivity following stressful events (MyinGermeys et al., 2009).

State-based aspects of thoughts and behaviours can be assessed using the experience sampling method (ESM, aka ecological momentary assessment) (Bolger & Laurenceau, 2013; Csikszentmihalyi & Larson, 1987). ESM is a form of intensive longitudinal data collection where participants repeatedly respond, commonly multiple times per day, to questionnaires that assess their experience "right now". Participants' responses may be cued by prompts that occur at random or fixed intervals or by an event (e.g., when the person exercises). Although the method may be burdensome for participants and researchers (Palmier-Claus, MyinGermeys, Barkus, Bentley, & Udachina, 2011), the observations obtained have the advantages of offering a precise test of temporal relationships between variables of interest and increased ecological validity (Bolger & Laurenceau, 2013).

ESM data collection yields a hierarchical dataset where a series of observations (i.e., single responses at a particular time point) are nested within participants. A range of modeling options exists to analyse nested data including regression with robust standard errors and multilevel modeling (MLM). Although MLM is more complex than traditional regression, it allows explicit investigation of individual variability in relationships (i.e., investigation of 'random effects'). For example, a traditional regression approach to studying the relationship between affect in the morning and subsequent drinking in the evening assumes that this relationship is constant across individuals. However, it is possible that some individuals' drinking is more influenced by their mood than others; in other words, that the relationship between affect and subsequent drinking will be stronger for those particular individuals. MLM can test this possibility and also explain variation in this relationship using individual level variables (i.e., an individual's trait coping or impulsivity could explain variation in the strength of the relationship). For this reason, MLM is commonly employed to analyse ESM data.

Within the MLM framework, most commonly relationships between variables are represented in a linear fashion using either linear regression (continuous DV) or logistic regression (binary DV). Although in many cases a linear model may accurately represent the data, it is not guaranteed that the relationships are linear and other relationships are possible. Given this, when analysing ESM data, we recommend undertaking a comprehensive strategy that investigates a range of possible relationships.

More accurate modeling of relationships will contribute to greater understanding of the phenomena of interest.

In this paper we demonstrate such an approach in the context of an analysis undertaken to assess the relationship between positive affect and risky single occasion drinking (RSOD; consumption of 5+ standard drinks in one sitting). The study involved 37 participants (8 males; 29 females) responding to a smartphone-based questionnaire four times per day for ten days. At each time-point, the questionnaire measured participants' mood and whether they had engaged in RSOD. A baseline questionnaire included measures of demographic information and impulsivity (fun seeking and drive). In the context of this dataset, three models are illustrated and compared: a traditional linear model and two alternative models useful for studying non-linear effects: a piecewise regression model and a threshold dose-response model (Hunt & Rai, 2003).

Statistical Models

Traditional model

Commonly, ESM data are analysed using a log-linear model (a multilevel logistic regression) (Hox, 2002). In this model, a binary dependent variable (e.g., RSOD) is regressed onto one or more independent variables (e.g., previous positive mood). This is represented below in equation 1, where i represents the i^{th} individual and j represents the j^{th} assessment point; β_{01i} represents the intercept for the Level 1 equation (i.e., the average probability of engaging in risky drinking); β_{10i} is the unstandardised coefficient representing the relationship between the independent variable and the dependent variable (i.e., the relationship between positive mood and RSOD). β_{01i} is the random effect representing individual differences in the Level 1 IV-DV relationship (i.e., individual differences in the strength of the relationship between positive affect and RSOD). In the event that this random effect is significant, β_{01i} is regressed onto Level 2 (individual difference) variables (in this case: age, gender, fun seeking, drive). This is shown in equation 2, where γ_{001} is the intercept for the Level 2 model; γ_{001} - γ_{013} are the unstandardized coefficients representing the moderating influence of the Level 2 variables on the relationship between positive mood and drinking; u is the error term for Level 2.

Piecewise regression model

This model assumes that there is a cutting point (or knot) on the IV continuum at which the slope of the relationship between IV and DV changes. In a standard piecewise regression, the researcher must pre-specify the value of the knot (i.e., the value where the relationship between positive affect and RSOD changes) in

$$logit(Drink_{ij}) = \beta_{01i} + \beta_{10i} * (positive\, mood) + e_{ij} \qquad \text{(Equation1)}$$

$$\beta_{10} = \gamma_{001} + \gamma_{010} * (age) + \gamma_{011} * (gender) + \gamma_{012} * (fun) + \gamma_{013} * (drive) + u \qquad \text{(Equation 2)}$$

order to run the model. In the absence of prior evidence for what that cut value should be, researchers may trial different values. In brief, the equation incorporates two key predictors representing the slope below and above the knot. When an individual scores below the knot, the second predictor (above the knot) drops out of the equation:

$$logit(Drink_{ij}) = \beta_{01i} + \beta_{10i} * (positive\,mood) + D * \beta_{11i} * (positive\,mood - t)$$

$$D = \begin{cases} 0 & \text{positive mood} < \text{t} \\ 1 & \text{positive mood} \geq \text{t} \end{cases} \qquad \text{(Equation 3)}$$

Where β_{01i} represents the intercept; β_{10i} represents the slope below the knot; β_{11i} represents the slope at or above the knot; D is the dummy variable representing whether the knot value (t = cutting value on positive mood) has been met/exceeded (D=1) or not (D=0).

Threshold dose-response model

This model is differentiated from the traditional log-linear model in that it includes a threshold value around which the shape of slope for the IV-DV (i.e., positive affect-RSOD) relationship changes, thus in effect producing two lines of best fit (equation 4). The basis for this model is the notion that the relationship between the IV and DV is negligible (~ zero relationship) below a threshold because low-level exposure fails to influence the likelihood of the target event. Once exposure (in our example, positive mood) exceeds this threshold, a positive linear relationship between exposure level and likelihood of outcome (risky drinking) is evident. Importantly, the threshold level is empirically derived from the data rather than needing to be pre-specified by the researcher.

Where logit($drink_{ij}$) is the probability of drinking expressed in logit form; τ is the threshold dose of positive mood; β_0 is the intercept; β_1 is the slope parameter above the threshold; d is the actual dose (i.e., level of positive mood). As implied by Equation 4, the probability of a drinking episode is held constant when positive mood is below the threshold, and exhibits a dose response relationship beyond the threshold (see Figure 1).

$$Logit(Drink_{ij}) = \begin{cases} \beta_0 & \text{f or } d_i < \tau \\ \beta_0 + \beta_1(d_i - \tau) & \text{f or } d_i \geq \tau \end{cases} \qquad \text{(Equation 4)}$$

Data Analytic Strategy

Overview

The utility of three models was explored in the context of the relationship between positive mood and drinking. In each of the models, positive mood at one time point was used to predict likelihood of RSOD (Yes/No) at the next time point, in order to uphold the longitudinal nature of the data and to demonstrate temporal precedence of positive mood. The non-independence of observations arising from the repeated measures design was controlled using

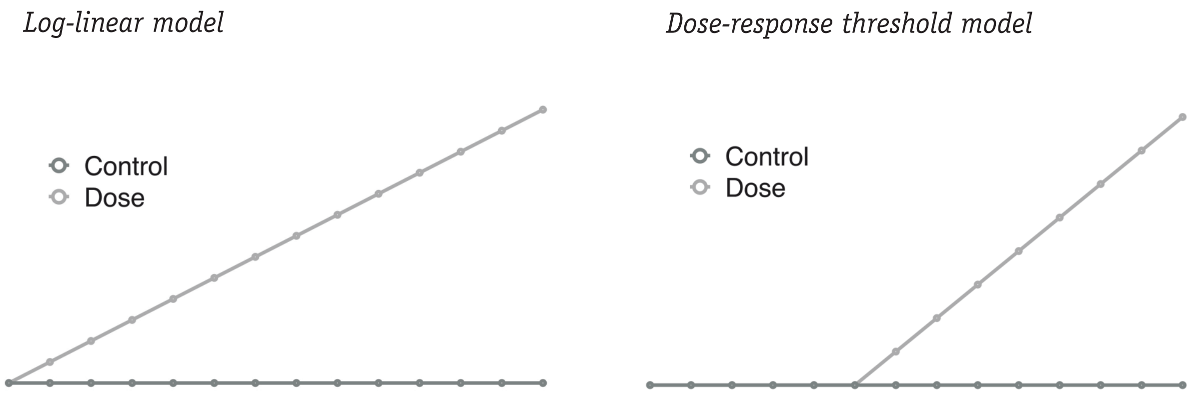


Figure 1: On the left-hand side is a standard log-linear representation of the relationship between positive mood and probability of drinking (traditional model), whereas the panel on the right shows the threshold model, which consists of two separate lines of best fit (a flat line for sub-threshold levels of positive mood, and accelerating probability proportional to exposure beyond the threshold level).

MLM. In each of the three models, random effects were tested for significance to determine whether the strength of the positive mood-drinking relationship varied from individual to individual.

Model comparison

The following indices are used to facilitate comparison of the different modeling approaches: (1) Odds ratios (ORs) were compared in order to compare strength of the IV-DV relationship, (2) standard errors of ORs permitted assessment of precision of these parameter estimates, and (3) log-likelihood, AIC, and BIC values were consulted to make comparisons of fit between these non-nested models, with the lower BIC value having best fit relative to the other models tested. We follow STATA convention of classifying a difference in BIC>10 between two competing models as strong support for the model with the lower BIC value.

Results

The standard multilevel logistic regression suggests that positive mood does not reliably predict the likelihood of a drinking episode (OR = 1.02, se = .012, p = .334). Moreover, this effect failed to vary significantly across individuals (Z = 0.02, p = .986).

The piecewise regression model was fit with different cutting points for the knot (10, 20, 30, and 40), and the best fitting model was achieved when positive mood was split above and below 30. Even so, in this model the positive mood-drinks relationship was positive but non-significant both below the knot (OR = 1.01, se = 0.019, p = .679) and above the knot (OR = 1.02, se = .03, p = .499). Furthermore, the two slopes failed to significantly vary ($Z<30$ = 0.285, p = .776; $Z\geq 30$ = 0.227, p = .821).

Finally, the threshold model suggests that the relationship between positive mood and likelihood of drinking is negative below the threshold (OR = 0.97, se = .11, p = .768) and positive above the threshold (OR = 1.01, se = .02, p = .566), but neither effect was significant. However, when these slopes were allowed to vary, the slope above the threshold significantly differed across participants (Z = 9.88, p < .001). Individual differences in this slope were regressed onto key trait-level variables, and it was found that the slope was strongest for individuals who were older (γ_{010} = .003, p <.001), male (γ_{011} = .009, p = .032), and who reported tendency to engage in behaviors because they are perceived as fun (γ_{012} = .002, p = .014). Reward drive was not a reliable moderator of the positive mood-drinking relationship (γ_{013} = -.001, p = .379). Finally, the slope below the threshold did not differ across individuals (Z = 0.28, p = .779).

Comparison of model fit statistics

As shown in Table 1, the threshold model produced the best fit of the data, followed by the traditional model and then the piecewise model. Using a difference of BIC > 10, the improvement in fit when using the threshold approach relative to the other two approaches provides strong support for this model.

Table 1. *Comparison of fit statistics for the three models*

Model	Log Likelihood	AIC	BIC
Traditional	-272.03	550.11	563.64
Spline	-281.51	570.01	589.17
Threshold	-247.43	504.86	527.39

Note: AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion

Discussion

Despite a significant increase in the volume of experience sampling studies (Mehl & Connor, 2011), there has been limited consideration of how to optimally model the state-based associations captured with this study design. The present study demonstrates several different modeling approaches for their ability to model the relationship between positive affect and likelihood of engaging in RSOD.

Although the positive affect-drinking relationship was weak across each of the tested models, the benefits of a threshold dose-response approach were still evident. First, this threshold model was the only model to detect that the relationship between positive affect and drinking has a negative slope at low levels of positive affect. The traditional multilevel logistic regression approach summarises a single line of best fit, and suggested that the relationship is positive. The piecewise approach also suggested that the relationship is positive across the range of positive affect levels, although the relationship may be slightly stronger at higher levels of positive affect. The stronger performance of the threshold model is further supported by commonly used model fit statistics (log likelihood, AIC, BIC), which suggested that the threshold approach provided a meaningful improvement in correspondence with the data relative to the other two models. Third, the threshold model was the only one to identify random effects for the positive affect-drinking relationship, and these random effects were in turn linked with age, gender and fun-seeking.

A further advantage of the threshold approach over the piecewise approach is that the former empirically derives the appropriate cutting point/threshold, whereas the latter requires researchers to pre-specify the cutting point(s) and then test their plausibility. This pre-specification threatens to be inaccurate: in instances where a predictor with a large range of scores is modeled, there are many different points to be possibly tested, increasing the likelihood that the researcher will miss the appropriate value. Indeed, although we presented results for the best of several knots tested, the poor performance of the piecewise model in this study may derive from choice of knot value.

It should be noted that the added complexity of the piecewise and threshold models appeared to come at a cost to efficiency in estimation as the standard error for the odds ratio of the slope in the traditional model was lower than the standard errors for any of the parameter estimates in the other two models. This is consistent with previous studies (e.g., Hunt & Rai, 2003). Insofar as this is a common effect in these models, the implication is that power may be lower when using this analysis, relative to a standard logistic regression model, and thus would necessitate a larger sample size and/or routine inclusion of covariates that may serve to reduce error variance.

References

Bolger, N., & Laurenceau, J-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford.

Colautti, L., Fuller-Tyszkiewicz, M., Skouteris, H., McCabe, M., Blackburn, S., & Wyett, E. (2011). Accounting for fluctuations in body dissatisfaction. *Body Image, 8*(4), 315-321. doi:10.1016/j.bodyim.2011.07.001

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease, 175*(9), 526-536. doi:10.1097/00005053-198709000-00004

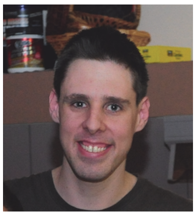
Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Hunt, D., & Rai. S. N. (2003). A Threshold Dose-Response model with random effects in teratological experiments. *Communications in Statistics - Theory and Methods, 32*(7), 1439-1457. doi:10.1081/STA-120021567

Mehl, M. R., & Conner, T. S. (2011). *Handbook of research methods for studying daily life*. New York, NY: Guilford.

MyinGermeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychological Medicine, 39*(9), 1533-1547. doi:10.1017/S0033291708004947

Palmier-Claus, J.E., MyinGermeys, I., Barkus, E., Bentley, L., & Udachina, A. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatrica Scandinavica, 123*(1), 12-20. doi:10.1111/j.1600-0447.2010.01596.x



Ben Richardson
is Associate Head of School at the School of Psychology at Deakin University, Australia

ben.richardson@deakin.edu.au

Matthew Fuller-Tyszkiewicz
is Senior Lecturer in Psychology (Research Methods) at the School of Psychology at Deakin University, Australia

matthew.fuller-tyszkiewicz@deakin.edu.au