Introduction to `mtm`:
An R Package for Marginalized Transition Models

Bryan A. Comstock and Patrick J. Heagerty
Department of Biostatistics
University of Washington

1 Introduction

Marginalized transition models are a general parametric class of serial dependence models that permit likelihood based marginal regression analysis of binary response data. The marginalized transition model may be used with data where subjects have variable lengths of follow-up, permitting likelihood analysis in settings where data may be missing at random (MAR). The methods developed in Heagerty (2002), are a natural extension of the first-order Markov models of Azzalini (1994).

Marginalized transition models are a convenient modeling choice in situations where the marginal mean regression structure is the primary target of inference and we would like to allow a general dependence structure for longitudinal binary outcome measures. While in Heagerty (2002) a general pth-order dependence structure was developed for a broad and flexible class of marginalized transition models, the `mtm` R library provides functions for modeling a dependence structure up to 2nd-order. This document is intended to serve as a supplementary, more detailed resource to the `mtm` library R help files.

2 General Framework and Notation

We restrict our focus to serial binary response data $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{in_i})$ observed on subjects $i = 1, \ldots, N$ at times $t = 1, \ldots, n_i$. We also assume that there are associated exogenous but possibly time-varying covariates $\boldsymbol{X_{it}} = (X_{it,1}, \ldots, X_{it,r})$ recorded for each subject at each occasion and our statistical objective is to obtain estimates for the regression of Y_{it} on $\boldsymbol{X_{it}}$. We assume that the regression model properly specifies the full covariate conditional mean defined as $\mu_{it}^M = \mathrm{E}(Y_{it} \mid \boldsymbol{X_{it}}) = \mathrm{E}(Y_{it} \mid \boldsymbol{X_{i1}}, \ldots, \boldsymbol{X_{in_i}})$. This condition assumes that stochastic time-varying covariates are properly modeled through $\boldsymbol{X_{it}}$ and that the current values of the response vector are not good predictors of future covariates. Finally, the marginal generalized linear model specifies $g(\mu_{it}^M) = \boldsymbol{X_{it}}\ \boldsymbol{\beta}$, where $g(\ )$ is a link function and $\boldsymbol{\beta}$ measures the influence of covariates on the average response. In the next sections, we describe the additional assumptions regarding the dependence among the response variables.

2.1 First Order Marginalized Transition Model

A binary Markov chain model was introduced in Azzalini (1994) to accommodate serial dependence commonly observed in longitudinal data. Given the immediate previous response in a first-order Markov model, the current response variable is assumed to be conditionally independent of any previous outcome variables, E($Y_{it} \mid Y_{ij}$, j < t) = E($Y_{it} \mid Y_{it-1}$). The probabilities that define the first order Markov process are given by $p_{it,0}$ = E($Y_{it} \mid Y_{it-1} =$ 0) and $p_{it,1}$ = E($Y_{it} \mid Y_{it-1} = 1$).

The first-order marginalized transition model is specified by assuming a regression structure for the marginal mean μ_{it}^M = E($Y_{it} \mid \boldsymbol{X_{it}}$) using a generalized linear model

$$g(\mu_{it}^M) = \mathbf{X_{it}}\boldsymbol{\beta} \tag{1}$$

The marginal mean regression model is constrained by the transition probabilities to satisfy:

$$\mu_{it}^M = p_{it,1}\mu_{it-1}^M + p_{it,0}\big(1 - \mu_{it-1}^M\big) \tag{2}$$

Serial dependence is then modeled using

$$\mu_{it}^C = E(Y_{it} \mid \mathbf{X_{it}}, Y_{ij} = y_{ij}, j < t) \tag{3}$$

$$logit\big(\mu_{it}^C\big) = \Delta_{it} + \gamma_{it} y_{it-1} \tag{4}$$

where Δ_{it} = logit($p_{it,0}$) and γ_{it} = log Ψ_{it} is the log odds ratio associated with the first-order transition probabilities

$$\Psi_{it} = \frac{p_{it,1}/(1 - p_{it,1})}{p_{it,0}/(1 - p_{it,0})} \tag{5}$$

Lastly, the serial dependence of Y_{it} on Y_{it-1}, given by the log odds ratio $\gamma_{it,1}$, is allowed to vary as a function of covariates $Z_{it,1}$ through

$$\gamma_{it,1} = \mathbf{Z_{it,1}}\boldsymbol{\alpha_1} \tag{6}$$

To summarize the above equations, the first-order marginalized transition model separates the specification of the regression model for outcome Y_{it} on covariates $\boldsymbol{X_{it}}$ from the dependence of Y_{it} on the previous outcome Y_{it-1} (autocorrelation). We are then interested in drawing inference on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha_1}$, the marginal association between outcome and covariates and how serial dependence is influenced by covariates.

2.2 Second Order Marginalized Transition Model

In Heagerty (2002), the natural extension of marginalized transition models to p^{th}-order dependence is explained in detail. The `mtm` library provides the ability to fit models up to second-order dependence and provides a score test for whether or not there is evidence for third-order dependence. Here, we briefly describe the extension of the first-order model to a second-order model.

Extensions of equations (3) and (4) allow Y_{it} to depend on the history through the previous two responses, Y_{it-1} and Y_{it-2}, by combining the marginal mean model of equation (1) with

$$\begin{aligned} \mu_{it}^{C} &= E(Y_{it} \mid \mathbf{X_{it}}, Y_{ij} = y_{ij}, j < t) && (7) \\ logit(\mu_{it}^{C}) &= \Delta_{it} + \gamma_{it,1} y_{it-1} + \gamma_{it,2} y_{it-2} && (8) \\ \gamma_{it,j} &= \mathbf{Z_{it,j}} \boldsymbol{\alpha_j} \qquad \text{for j = 1,2} && (9) \end{aligned}$$

Similar to first-order models, the second-order marginalized transition model separates the specification of the regression model of the outcome Y_{it} on covariates $\mathbf{X_{it}}$ from the dependence of Y_{it} on the previous two outcomes Y_{it-1}, Y_{it-2} (autocorrelation). Second order models allow dependence of Y_{it} on the two previous responses, Y_{it-1} and Y_{it-2}, where the parameters $\boldsymbol{\alpha_1}$ again reflect how the dependence on the previous response is affected by covariates $\boldsymbol{Z_{it,1}}$ and the parameters $\boldsymbol{\alpha_2}$ determine how the dependence on the second previous response varies by covariates $\boldsymbol{Z_{it,2}}$.

3 `mtm` Implementation in R

The `mtm` R library contains the functions `mtm.lag1()` and `mtm.lag2()` that fit marginal transition models with first and second order dependence respectively. The functions `print.mtm1()` and `print.mtm2()` are also contained in `mtm` and may be used to display a summary of the model output.

3.1 R Function Description: `mtm.lag1()`

A lag-1 marginalized transition model is called with the following R syntax:

```
mtm.lag1(marginal, trans1, id, beta=NULL, alpha1=NULL, offset=NULL,
                        data=NULL, tol = 1e-4)
```

`marginal` a symbolic description of the marginal model to be fit that generally takes the form `y ~ x` where `y` are serial binary outcome data and `x` are the covariates. The covariates `x` are a series of terms separated by `+` which specify the marginal linear predictor for `y`.

`trans1` covariates used to estimate the dependence of `y(t)` on `y(t-1)`. In general, `trans1` has the form $\sim$ `z1` where `z1` is a subset of covariates `x` and is a series of terms separated by `+`.

`id` a vector that identifies the clusters which correspond to the binary response vector given by `y`.

`beta` (optional) initial parameter estimate(s) of how the covariates `x(t)` influence the average response `y(t)`. The number of estimates provided in `beta` should correspond to the number of covariates in `x`, including an intercept.

`alpha1` (optional) initial estimate(s) of how the dependence of `y(t)` on `y(t-1)` varies as a function of covariate(s) `z1`. The number of estimates provided in `alpha1` should correspond to the number of covariates in `z1` used to assess the serial dependence in the outcome measure.

`offset` an optional argument used to specify an *a priori* known component to be included in the linear predictor during fitting

`data` (optional) a data frame containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`.

`tol` tolerance is a measure used in the numerical calculations to determine whether or not convergence of the point estimates has occurred. The default is 1e-4.

3.2 R Function Description: `mtm.lag2()`

The R call for a lag-2 marginal transition model is similar to a lag-1 model with the exceptions detailed below:

```
mtm.lag2(marginal, trans1, trans2, id, beta=NULL, alpha1=NULL,
         alpha2=NULL, tol = 1e-4, iter = 50, data)
```

`marginal` is a symbolic description of the model to be fit that generally takes the form `y` $\sim$ `x`, where `y` are serial binary outcome data and `x` are the covariates. The covariates `x` are a series of terms separated by `+` which specify the marginal linear predictor for `y`.

`trans1` covariates used to estimate the dependence of `y(t)` on `y(t-1)`. In general, `trans1` has the form $\sim$ `z1` where `z1` is a subset of covariates `x`.

`trans2` covariates used to estimate the dependence of `y(t)` on `y(t-2)`. In general, `trans2` has the form $\sim$ `z2` where `z2` is a subset of covariates `x`.

`beta` (optional) initial parameter estimate(s) of how the covariates `x(t)` influence the average response `y(t)`. The number of estimates provided in `beta` should correspond to the number of covariates in `x`, including an intercept.

`alpha1` (optional) initial estimate(s) of how the dependence of `y(t)` on `y(t-1)` varies as a function of covariate(s) `z1`. The number of initial estimates provided in `alpha1` should correspond to the number of covariates in `z1` used to assess the serial dependence in the outcome measure.

`alpha2` (optional) initial estimate(s) of how the dependence of `y(t)` on `y(t-2)` varies as a function of covariate(s) `z2`. The number of initial provided in `alpha2` should correspond to the number of covariates in `z` used to assess the serial dependence in the outcome measure.

3.3 `mtm.lag1` and `mtm.lag2` Function Output

`Maximized log-likelihood` Maximized log-likelihood of model.

`Beta estimates` Marginal model log-odds ratio estimates (`est.`), model-based standard errors (`s.e.M`), empirical standard errors (`s.e.E`), and z-score test statistics derived using the model-based standard errors.

`Alpha1 estimates` First-order dependence model log-odds ratio estimates, model-based standard errors, empirical standard errors, and z-score test statistics derived using the model-based standard errors.

`Alpha2 estimates` Second-order dependence model log-odds ratio estimates, model-based standard errors, empirical standard errors, and z-score test statistics derived using the model-based standard errors.

`Lag-3 score test` A score test for evidence of outcome dependence beyond first and second order.

3.4 An Example

In this section, we present analyses of the Madras Longitudinal Schizophrenia Study data (Thara et al., 1994) that were explored with these methods in detail in Heagerty (2002). The Madras data are included in the `mtm` library and may be loaded in R with `data(madras)`.

The data contain serial binary outcome measures y_{it} that denote the presence of positive psychiatric symptoms over the course of $t = 0, \ldots, 11$ months during the first year following hospitalization for schizophrenia for patients $i = 1, \ldots, 86$ (denoted by `id`). The dataset also contains the binary indicator of whether or not the patient's `age` at hospitalization $<$ 20 (0 = `age` $\geq$ 20, 1 = `age` $<$ 20), and `gender` (0 = male, 1 = female), and the interactions between both of these covariates with time (`month`). Finally, the data contain a binary indicator, labeled `initial` of whether or not `month = 1`. This indicator variable allows α_1 to be used for both the second-order model $Y_{it} \mid Y_{it-1}, Y_{it-2}$, and the initial state, $Y_{i1} \mid Y_{i0}$.

The goal of this data analysis is to determine factors that may correlate with the course of illness. Specifically, we would like to examine whether the rate of decline in symptoms prevalence differs across gender and age-at-onset subgroups. This question will be explored by evaluating marginal effects of the interaction terms of `age` and `gender` with `month`.

The following R code is used to explore this data with a second-order marginalized transition model with `mtm.lag2()`:

```
## Load mtm library and madras data
>  library(mtm)
>  data(madras)
>  attach(madras)
>  madras[1:10, ]  # Print the first 10 lines...
    id  y   month age gender   monthXage monthXgender initial
1    1  1     0   0      0           0            0        0
2    1  1     1   0      0           0            0        1
3    1  1     2   0      0           0            0        0
4    1  1     3   0      0           0            0        0
5    1  1     4   0      0           0            0        0
6    1  0     5   0      0           0            0        0
7    1  0     6   0      0           0            0        0
8    1  0     7   0      0           0            0        0
9    1  0     8   0      0           0            0        0
10   1  0     9   0      0           0            0        0
```

As an example of a second-order transition model, the first and second order time trend terms of Model 5 in Heagerty (2002) are presented below:

```
## lag-2 transition model -- model 5 of Heagerty (2002):
>  model2 <- mtm.lag2(marginal = y ~ month + age + gender + monthXage + monthXgender,
                trans1 = ~ initial + month, trans2 = ~ 1, id=id, data=madras)
>  print.mtm2(model2)

Marginalized Transition Models - lag 2

  maximized logLikelihood = -332.931

Beta estimates:

               est. s.e. M s.e. E      Z
(Intercept)   0.568  0.295  0.291  1.924
```

```
month        -0.234  0.054  0.056 -4.324
age           0.619  0.434  0.461  1.425
gender       -0.160  0.407  0.424 -0.395
monthXage    -0.100  0.091  0.090 -1.098
monthXgender -0.149  0.089  0.096 -1.672

Alpha1 estimates:

             est. s.e. M s.e. E     Z
(Intercept) 2.099  0.559  0.568 3.755
initial     0.403  0.740  0.732 0.544
month       0.156  0.096  0.093 1.626

Alpha2 estimates:

             est. s.e. M s.e. E     Z
(Intercept) 0.597  0.293  0.262 2.035

Score test for lag-3 coefficient = 0.072 p-val = 0.789
```

The model output provides both model-based (`s.e.M`) and empirical (`s.e.E`) standard errors for the log-odds ratio estimates. The estimates provided under the *Beta estimates* model output indicate to what extent symptoms in the current month ($month = t$) are correlated with specific covariates. For example, among males with later age-at-onset the estimated rate of decline in the log odds of schizophrenic symptoms is -0.234 per month, denoted by the beta estimate `month` (z = -4.324). This translates to more than a 20% average reduction in the presence of symptoms per month after the initial hospitalization (OR = 0.79, 95% CI: 0.71 - 0.88) for this particular subgroup. The other marginal estimates may be interpreted in a similar manner.

The dependence of current outcomes on previous outcomes is summarized under the headings `Alpha 1 estimates` and `Alpha 2 estimates`. For instance, one may obtain an estimate of the odds ratio comparing the odds of symptoms at $month = 1$ ($Y_1 = 1$) given symptoms at baseline ($Y_0 = 1$) to the odds of symptoms at one month ($Y_1 = 1$) given no symptoms at baseline ($Y_0 = 0$). In this case, the odds ratio is estimated by:

$$
\begin{aligned}
OR &= e^{Intercept_{\alpha_1} + initial_{\alpha_1} + month_{\alpha_1}} \\
&= e^{2.099+0.403+0.156} \\
&= 14.27
\end{aligned}
$$

Similarly, one may obtain an estimate of the odds ratio comparing the odds of $Y_5 = 1$ given $Y_4 = 1$ and $Y_3 = 1$ to the odds of $Y_5 = 1$ given $Y_4 = 0$ and $Y_3 = 0$. Here, the odds ratio is estimated by:

$$\begin{aligned} OR &= e^{Intercept_{\alpha_1}+5*month_{\alpha_1}+Intercept_{\alpha_2}} \\ &= e^{2.099+5*0.156+0.597} \\ &= 32.33 \end{aligned}$$

The individual estimates provided under the `Alpha 1 estimates` model output indicate to what extent symptoms in the current month (*month = t*) are dependent on the previous month's (*month = t-1*) symptoms for specific subgroups. For instance, while we observed above that there was a 20% decline in the marginal odds of symptoms per month, we see that patients whose symptoms persist over time (from one month to the next) are more likely to continue to have symptoms. For every one-month increase post-hospitalization, those with persistent symptoms in the preceding month were exp(0.156) times, or 17%, more likely to have symptoms in the current month. However, this result was not significant at the nominal 0.05 level (p = 0.104).

Similarly, the estimates provided under the *Alpha 2 estimates* model output indicate to what extent symptoms in the current month (*month = t*) are dependent on the presence of symptoms two months ago (*month = t-2*). In this particular model, an intercept was included to describe the dependence of the currents months symptoms on those that were reported two months previously. The model output suggests that the presence of symptoms two months earlier are associated with an 82% higher odds of schizophrenic symptoms in the present month independent of the presence symptoms in the previous month (*month = t-1*), (OR = 1.82, 95% CI: 1.02 - 3.23).

4 Conclusions

Marginalized transition models allow for simultaneous likelihood-based estimation of the average response and for the correlation among longitudinal observations. Such models are permissible both when subjects have varying lengths of follow-up and when data may be missing at random (MAR). The latter characteristic is an advantage over semiparametric approaches such as generalized estimating equations (GEE) which may produce biased results unless data are missing completely at random (MCAR). In this document, we introduced a new R-package `mtm` which contains functions to fit first and second order marginalized transition models to equally spaced binary response data.

References

[1] A Azzalini. Logistic regression for autocorrelated daata with application to repeated measures. *Biometrika*, (81):767–775, 1994 (correction: 1997, 84, 989).

[2] Patrick J. Heagerty. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, (58):342–351, June 2002.

[3] R. Thara, M. Henrietta, A. Joseph, S. Rajkumar, and W. Eaton. Ten year course of schizophrenia - the madras longitudinal study. *Acta Psychiatrica Scandinavica*, (90):329–336, 1994.