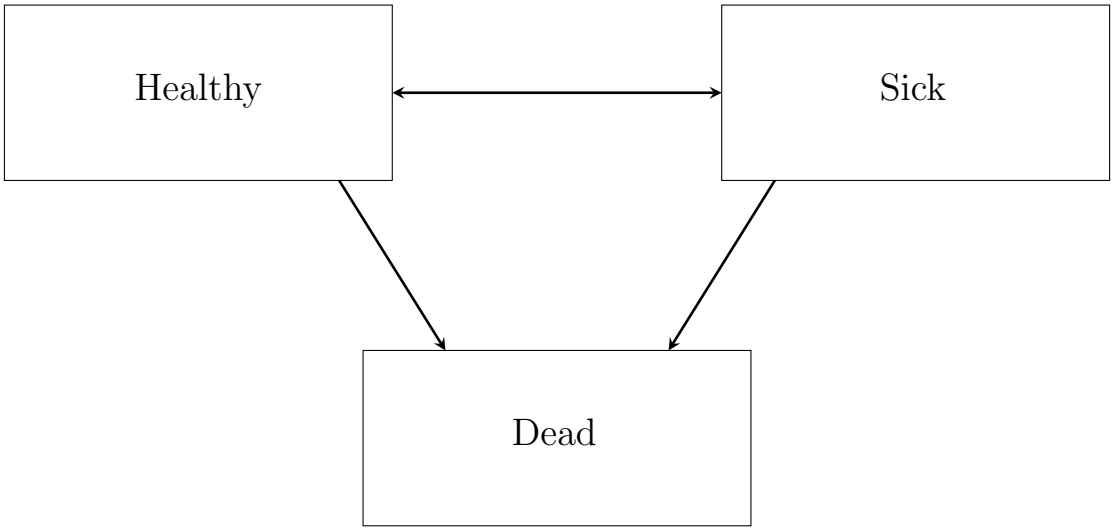
Transition Models for Categorical Longitudinal Data

Unlike marginal models, which model within-subject correlations based on a correlation pattern matrix, and mixed effects models, which model within-subject correlations based on random effects, *transition models* explicitly model the development of longitudinal outcomes based on previously observed outcomes. They are most natural to think about when we have categorical, longitudinal data, where there are several states that an individual may find themselves in. In the following diagram, we can imagine tracking the longitudinal development of some illness, where individuals are healthy, sick, or have died from the illness. They are free to move between healthy and sick, but once they have died, their process stops entirely.



If we consider consecutive timepoints, j and $j+1$, and we want to know $P(Y_{i,j+1} = \text{Dead})$ for this process, it evidently helps to know the value at Y_{ij}. In the extreme case we have $P(Y_{i,j+1} = \text{Dead}|Y_{i,j} = \text{Dead}) = 1$, but presumably we also know that $P(Y_{i,j+1} = \text{Dead}|Y_{i,j} = \text{Sick}) > P(Y_{i,j+1} = \text{Dead}|Y_{i,j} = \text{Healthy})$. Generalizing this process naturally lends itself to the idea of modelling the state of Y_i at time j, based on the previously observed trajectory of Y_i. It becomes natural to think of the longitudinal process as an emerging stochastic process.

To formalize notation, consider a state space for Y, denoted $\mathcal{S}$. In our previous example, this would make $\mathcal{S} = \{\text{Healthy}, \text{Sick}, \text{Dead}\}$. Then, assume that we have discretized our time measurement so that there are fixed time points, $t_1 < t_2 < \cdots < t_K$, with corresponding measurements of our outcomes $Y_{i1}, Y_{i2}, \ldots, Y_{iK}$. At time j, we define a history vector for individual i to be $\mathcal{H}_{ij} = (Y_{i1}, Y_{i2}, \ldots, Y_{i,j-1})$. That is, $\mathcal{H}_{ij}$ contains all of the information available prior to time point j. We are interested in estimating the parameters given by

$$P(Y_{ij} = \ell|\mathcal{H}_{ij}).$$

The Markov Assumption

One simplifying assumption that we often will make when dealing with stochastic processes in general (and transition models in particular) is that the *Markov property* is satisfied. In

general, a stochastic process is said to be an *rth order Markov chain* if

$$P(Y_{ij} = \ell|\mathcal{H}_{ij}) = P(Y_{ij} = \ell|Y_{i,j-1}, Y_{i,j-2}, \ldots, Y_{i,j-r}).$$

That is, under this Markovian assumption, the probabilities at time point j only rely on the most recent previous r states. This is often a reasonable assumption, and frequently we will take $r = 1$ or $r = 2$. In the event of a *first order Markov chain* we can take

$$P(Y_{ij} = m|\mathcal{H}_{ij}) = P(Y_{i,j} = m|Y_{i,j-1}).$$

At each time point, t, there are then a set of *transition probabilities* that define the model. We denote this as

$$p_{\ell,m}(t) = P(Y_t = m|Y_{t-1} = \ell).$$

That is, $p_{\ell,m}(t)$ represents the probability of the process being observed in state m at time t when we know at the previous time, $t-1$, the process was in state ℓ. In addition to the transition probabilities, we also need to know initial distributions of states. That is, $P(Y_{i1} = \ell)$, since there is no previous information to condition on. We will denote these starting probabilities as $\pi_\ell = P(Y_{i1} = \ell)$.

Maximum Likelihood in the First-Order Markov Model

In order to estimate the probability parameters specified by the model, we can apply maximum likelihood estimation. The central idea is that we can divide one individuals trajectory up into single step transitions. For instance, if $K = 3$, at $t = 1$ the individual is observed in state ℓ, at $t = 2$ the individual is observed in state m, and at $t = 3$ the individual is back in state ℓ, then the probability of this pathway can be written out as

$$\begin{aligned}
L_i &= P(Y_{i1} = \ell, Y_{i2} = m, Y_{i3} = \ell) \\
&= P(Y_{i3} = \ell|Y_{i1} = \ell, Y_{i2} = m)P(Y_{i1} = \ell, Y_{i2} = m) \\
&= P(Y_{i3} = \ell|Y_{i1} = \ell, Y_{i2} = m)P(Y_{i2} = m|Y_{i1} = \ell)P(Y_{i1} = \ell) \\
&= P(Y_{i3} = \ell|Y_{i2} = m)P(Y_{i2} = m|Y_{i1} = \ell)P(Y_{i1} = \ell) \\
&= p_{m,\ell}(3) \cdot p_{\ell,m}(2) \cdot \pi_\ell \\
&= \pi_{Y_{i1}} \cdot p_{Y_{i1},Y_{i2}}(2) \cdot p_{Y_{i2},Y_{i3}}(3).
\end{aligned}$$

More generally, we can write this as

$$L_i = \pi_{Y_{i1}} \prod_{j=2}^{K} p_{Yi,j-1,Y_{i,j}}(j).$$

Combining this with other individuals gives us the full likelihood as

$$L(p) = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \left\{ \pi_{Y_{i1}} \prod_{j=2}^{K} p_{Yi,j-1,Y_{i,j}}(j) \right\}.$$

Now, when maximizing this likelihood, note that we have restrictions on $p_{\ell,m}(j)$. In particular, if we denote our sample space $\mathcal{S} = \{1, 2, \ldots, S\}$, then we can constrain

$$p_{m,S}(j) = 1 - \sum_{\ell=1}^{S-1} p_{m,\ell}(j).$$

That is, if we are currently in state ℓ, with probability 1 we will be in one of the S states next time step, and so the sum $\sum_{m=1}^{S} p_{\ell,m}(j) = 1$. As a result, we actually wish to consider the constrained optimization problem here, which we can get by adding on (to our log-likelihood) a term which forces the summation to be 1.

In the following derivation, we take $n_{\ell,m}(j)$ to be the number of individuals in our sample who were at ℓ at $t = j-1$ and at m at $t = j$. That is,

$$n_{\ell,m}(j) = \sum_{i=1}^{n} I(Y_{i,j-1} = \ell) I(Y_{i,j} = m).$$

For notational convenience, we take $n_{\ell,\cdot}(j)$ to be the total number of people who are in state ℓ at time $j-1$, without concern for where they end up at time j. That is,

$$n_{\ell,\cdot}(j) = \sum_{m=1}^{S} n_{\ell,m}(j) = \sum_{i=1}^{n} I(Y_{i,j-1} = \ell).$$

Taking this notation, and adding on the constraint to the above likelihood (after having taken the log) we can optimize this with respect to our transition probabilities, by differentiating with respect to the parameters, and solving for $\widehat{p}$ such that the score is 0.

$$\begin{aligned}
\ell(p) &= \sum_{i=1}^{n} \left\{ \log \pi_{Y_{i1}} + \sum_{j=1}^{K} \log\left(p_{Y_{i,j-1},Y_{i,j}}(j)\right) \right\} + \sum_{j=1}^{K} \sum_{s=1}^{S} \lambda_{sj} \left(1 - \sum_{\ell=1}^{S} p_{s,\ell}(j) \right) \\
\frac{\partial}{\partial p_{\ell,m}(j)} \ell(p) &= \sum_{i=1}^{n} I(Y_{i,j-1} = \ell) I(Y_{i,j} = m) \frac{1}{p_{\ell,m}(j)} - \lambda_{\ell,j} = \frac{n_{\ell,m}(j)}{p_{\ell,m}(j)} - \lambda_{\ell j}. \\
\implies \widehat{p}_{\ell,m}(j) &= \frac{n_{\ell,m}(j)}{\lambda_{\ell,j}} \\
1 = \sum_{m=1}^{S} \widehat{p}_{\ell,m}(j) &= \sum_{m=1}^{S} \frac{n_{\ell,m}(j)}{\lambda_{\ell,j}} \\
\implies \lambda_{\ell,j} &= \sum_{m=1}^{S} n_{\ell,m}(j) = n_{\ell,\cdot}(j) \\
\implies \widehat{p}_{\ell,m}(j) &= \frac{n_{\ell,m}(j)}{n_{\ell,\cdot}(j)}.
\end{aligned}$$

This gives us a (fairly intuitive) estimator for the transition probabilities, based on the maximum likelihood. The transition probability from ℓ to m at $t = j$ is given by

$$\widehat{p}_{\ell,m}(j) = \frac{n_{\ell,m}(j)}{n_{\ell,\cdot}(j)} = \frac{\{\#\text{ from } \ell \text{ to } m \text{ at } j\}}{\{\#\text{ in } \ell \text{ at } j-1\}}.$$

Time Homogeneity Assumption

One way that we can further simplify this setting is by making the assumption of *time homogeneity*. This corresponds to the assumption that the transition probabilities are constant across all time points. Using our notation, under the first-order Markov model, if we have that

$$p_{\ell,m}(j) = p_{\ell,m}(j') \quad \forall\{\ell, m\} \in \mathcal{S} \text{ and } \forall j \neq j' \in \{1, \ldots, K\},$$

then the Markov chain is time homogeneous. Under this assumption all we care about is the probability of transitioning from $\ell \to m$, regardless of what time that transition is happening at. We can work out the same likelihood derivation under this simplified assumption. Using the same notation as before, we get

$$\begin{aligned}
\ell(p) &= \sum_{i=1}^{n} \left\{ \log \pi_{Y_{i1}} + \sum_{j=1}^{K} \log \left(p_{Y_{i,j-1}, Y_{i,j}} \right) \right\} + \sum_{s=1}^{S} \lambda_s \left(1 - \sum_{\ell=1}^{S} p_{s,\ell} \right) \\
\frac{\partial}{\partial p_{\ell,m}} \ell(p) &= \sum_{i=1}^{n} \sum_{j=2}^{K} I(Y_{i,j-1} = \ell) I(Y_{i,j} = m) \frac{1}{p_{\ell,m}} - \lambda_\ell = \frac{\sum_{j=2}^{K} n_{\ell,m}(j)}{p_{\ell,m}} - \lambda_\ell. \\
\implies \widehat{p}_{\ell,m} &= \frac{\sum_{j=1}^{K} n_{\ell,m}(j)}{\lambda_\ell} \\
1 = \sum_{m=1}^{S} \widehat{p}_{\ell,m} &= \sum_{m=1}^{S} \frac{\sum_{j=2}^{K} n_{\ell,m}(j)}{\lambda_\ell} \\
\implies \lambda_\ell &= \sum_{m=1}^{S} \sum_{j=2}^{K} n_{\ell,m}(j) = \sum_{j=2}^{K} n_{\ell,\cdot}(j) \\
\implies \widehat{p}_{\ell,m} &= \frac{\sum_{j=2}^{K} n_{\ell,m}(j)}{\sum_{j=2}^{K} n_{\ell,\cdot}(j)}.
\end{aligned}$$

This gives rise to a similar interpretation as in the previous setting where here we can say that the transition probability from ℓ to m at any time is given by

$$\widehat{p}_{\ell,m} = \frac{\sum_{j=2}^{K} n_{\ell,m}(j)}{\sum_{j=2}^{K} n_{\ell,\cdot}(j)} = \frac{\{\#\text{ from } \ell \text{ to } m \text{ by } K\}}{\{\#\text{ in } \ell \text{ before } K\}}.$$

Inclusion of Covariates with Logistic Regression

One major shortcoming of this method is that we are unable to accommodate the effects of covariates on the transition probabilities. This is a problem primarily since the questions of interest tend to rely on testing the significance of covariates on the outcomes. However, if we consider the simplified setting of $\mathcal{S} = \{0, 1\}$, so that the categorical data are binary, there are only two (unique) parameters to estimate at each stage j. That is, $p_{00}(j) + p_{01}(j) = 1$ and $p_{10}(j) + p_{11}(j) = 1$, so by estimating $p_{01}(j)$ and $p_{11}(j)$, we have all the required parameters for this model. Consider expressing

$$\text{logit} \left\{ P(Y_{ij} = 1 | Y_{i,j-1}) \right\} = \alpha_{0j} + \alpha_{1j} Y_{i,j-1}.$$

From this relationship, we would find that

$$\begin{aligned}
&\text{logit}(p_{01}(j)) = \text{logit}(P(Y_{ij} = 1|Y_{i,j-1} = 0)) = \alpha_{0j} \\
&\text{logit}(p_{11}(j)) = \text{logit}(P(Y_{ij} = 1|Y_{i,j-1} = 1)) = \alpha_{0j} + \alpha_{1j} \\
&\implies p_{01}(j) = \text{expit}(\alpha_{0j}) \\
&\implies p_{11}(j) = \text{expit}(\alpha_{0j} + \alpha_{1j}).
\end{aligned}$$

As a result, this model represents a simple re-parameterization of the model expressed above. In the event that we are using the time homogeneous assumption, we would simply drop the j subscript from these models.

The form $\text{logit}(P(Y_{ij} = 1|Y_{i,j-1})) = \alpha_{0j} + \alpha_{1j}Y_{i,j-1}$ expresses a standard logistic regression model, where we are regressing the outcome Y_{ij} on the explanatory factor, $Y_{i,j-1}$. As a result, we could fit the parameter estimates here using all of our standard GLM theory! All we would need to do is transform our data frame to include a column with the lagged outcome ($Y_{i,j-1}$ at time j) and then use that as the predictor in the model. From here, we can use all of the standard logistic regression diagnostic, hypothesis testing, and so forth as we are used to.

ID	Y_1	Y_2	$\cdots$	Y_{K-1}	Y_K
1	$Y_{1,1}$	$Y_{1,2}$	$\cdots$	$Y_{1,K-1}$	$Y_{1,K}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
n	$Y_{n,1}$	$Y_{n,2}$	$\cdots$	$Y_{n,K-1}$	$Y_{n,K}$

Table 1

ID	j	Y_j	Y_{j-1}
1	2	$Y_{1,2}$	$Y_{1,1}$
1	3	$Y_{1,3}$	$Y_{1,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	K	$Y_{1,K}$	$Y_{1,K-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	2	$Y_{n,2}$	$Y_{n,1}$
n	3	$Y_{n,3}$	$Y_{n,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	K	$Y_{n,K}$	$Y_{n,K-1}$

Table 2

ID	Y_j	Y_{j-1}
1	$Y_{1,2}$	$Y_{1,1}$
1	$Y_{1,3}$	$Y_{1,2}$
$\vdots$	$\vdots$	$\vdots$
1	$Y_{1,K}$	$Y_{1,K-1}$
$\vdots$	$\vdots$	$\vdots$
n	$Y_{n,2}$	$Y_{n,1}$
n	$Y_{n,3}$	$Y_{n,2}$
$\vdots$	$\vdots$	$\vdots$
n	$Y_{n,K}$	$Y_{n,K-1}$

Table 3

Using the above examples, we have that Table 1 representing the standard wide-format data, with K observations for each of n subjects. Table 2 transforms this data to be suitable for an analysis where we do not make the time-homogeneous assumption. Here, we would regress Y_j on both Y_{j-1} and j, where j is treated as a factor. This would estimate separate parameters for (α_0, α_1) across all $j = 1, \ldots, K$. Table 3 displays the long-format for a time-homogeneous process, where we would estimate α_0 and α_1 by regressing Y_j on Y_{j-1}.

There are two reasons that we want to frame this analysis as a logistic regression. First, we can very easily move from a first-order Markov assumption, to an r-th order Markov assumption, by simply adding additional lag terms. For instance,

$$P(Y_{i,j} = 1|Y_{i,j-1}, Y_{i,j-2}) = \text{expit}\left(\alpha_0 + \alpha_1 Y_{i,j-1} + \alpha_2 Y_{i,j-2} + \alpha_3 Y_{i,j-1}Y_{i,j-2}\right),$$

will cover all possible combinations. In particular we will get that

$$\begin{aligned}
p_{(0,0),1} &= \text{expit}(\alpha_0) & p_{(0,1),1} &= \text{expit}(\alpha_0 + \alpha_1) \\
p_{(1,0),1} &= \text{expit}(\alpha_0 + \alpha_2) & p_{(1,1),1} &= \text{expit}(\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3).
\end{aligned}$$

The second benefit to this framing is that it provides an obvious way of allowing the transition probabilities to depend on additional variates, beyond simply the lagged terms. To see this, consider the time-homogeneous model, with the first-order Markov assumption. If we want the transition probabilities to depend on X_i as well, we can take

$$\text{logit}\left\{P(Y_{ij}=1|X_{ij},Y_{i,j-1})\right\} = X_{ij}'\beta + Y_{i,j-1}\left(X_{ij}'\alpha\right).$$

Then, by the same arguments presented above we are estimating separate transition probabilities based on whether the previous state was 0 (given by β) or 1 (given by $\alpha+\beta$). We can of course extend this, making higher order Markov assumptions, by simply replacing the $\alpha_\cdot$ terms above with $X_{ij}'\beta_\cdot$. These models can also be fit through standard, logistic regression (alongside all the benefits of asymptotic inference which come along with that)!

Some Notes and Further Considerations

- If we do not care about the impact of covariates and just wish to compute transition probabilities directly, these can be computed based on a summary table of the counts of one-step transitions.
- In the logistic regression model, β is interpreted as the change in the (logit) of p_{01} that is associated with a 1 unit increase in the relevant covariate, where α is interpreted as the difference in the effects of the covariates on transition probabilities ($p_{11}-p_{01}$).
- By using formal hypothesis tests, we can see whether simpler models fit the data equally well (e.g., testing whether the the rth order is necessary).
- In the logistic model, we have only specified the likelihood for the conditional component, and have ignored the baseline, $f(y_{i1})$, or, (in the case of an r-th order assumption),

  $$f(y_{i1},y_{i2},\ldots,y_{ir}).$$

  We then proceeded by ignoring this marginal, baseline term and working with the conditional likelihood,

  $$L_i^C = \prod_{j=r+1}^{K} f(y_{ij}|y_{i,j-1},\ldots,y_{i,j-r}).$$

  This conditional likelihood works out to be the standard GLM, and we used standard software. Alternatively, we could have also explicitly modelled the baseline distribution (e.g., assign the equilibrium distribution to it) which will increase efficiency **assuming that the model is correct**.
- In theory, similar extensions would work when the data are not binary (e.g. with multinomial GLMs)