Paper 7861-2016

Probability Density for Repeated Events

Bruce Lund

Magnify Analytic Solutions, a Division of Marketing Associates, LLC
Detroit MI, Wilmington DE, Charlotte NC

ABSTRACT

In customer relationship management (CRM) or consumer finance it is important to predict the time of repeated events. These repeated events might be a purchase, service visit, or late payment. Specifically, the goal is to find the probability density for the time to first event, the probability density for the time to second event, etc. Two approaches are presented and contrasted. One approach uses discrete time hazard modeling (DTHM). The second, a distinctly different approach, uses multinomial logistic model (MLM). Both DTHM and MLM satisfy an important consistency criterion which is to provide probability densities whose average across all customers equals the empirical population average. DTHM requires fewer models if the number of repeated events is small. MLM requires fewer models if the time horizon is small. SAS® code is provided through a simulation in order to illustrate the two approaches.

INTRODUCTION

A Company wants to predict whether its customers will perform actions such as re-purchasing the Company's products, coming in for service, or be late in making a payment. In the simplest case these predictions give a probability that the customer will perform such an action in the next 90 days (6 months, etc.). Such predictions are often provided by logistic regression models. These models provide a useful tactical tool.

But customer relationship management (CRM) requires a model that makes a more comprehensive prediction of the customer's future engagement with the Company. Specifically, a model is needed that gives the probability densities for the times of future events (purchases, service, payments) by the customer. The goal of the paper is to discuss this type of model. Two modeling approaches are presented and contrasted. One approach uses discrete time hazard modeling. The second uses multinomial logistic regression.

Notation and Definitions

Customers are observed for up to T periods of time (months, quarters, years, etc.). The start time of these observations could vary by customer as can the number of periods observed. In each time period it is recorded whether the customer made a purchase (one or more) of the Company's products. Formal mathematical notation is presented next.

If a customer makes one or more purchases during the t^{th} period after the customer's start time, then $\delta_t = 1$. Otherwise, $\delta_t = 0$. By definition $\delta_0 = 0$.

Event $\mathbf{E_{j,t}}$ occurs when $\sum_{t'=0}^{t-1} \delta_{t'} = j-1$ and $\delta_t = 1$. In words, event $\mathbf{E_{j,t}}$ occurs when a purchase in the t^{th} time-period is the j^{th} period in which the customer has made a purchase.

For example: Event $\mathbf{E_{2,3}}$ occurs when a purchase in the customer's 3^{rd} time-period is the 2^{nd} period where the customer has made a purchase.

Table 1 gives 4 examples to illustrate event $\mathbf{E_{2,3}}$

Table 1 Examples of Events

Time periods from start ➔	**1**	**2**	**3**	$\mathbf{E_{2,3}}$
Cust 001	purchase(s)	no purchase	purchase(s)	**YES**
Cust 002	purchase(s)	purchase(s)	purchase(s)	**NO**
Cust 003	no purchase	purchase(s)	purchase(s)	**YES**
Cust 004	purchase(s)	purchase(s)	no purchase	**NO**

It seems natural to use "time" instead of "time-period from start" and this simplification is used from this point on. The maximum number of events, logically, must be less than or equal to T. This maximum will be denoted by J. If events are rare, then J could be pre-set to be less than T. In the discussions that follow, whether J = T or J < T will not matter.

REVIEW OF THE DISCRETE TIME HAZARD MODEL FOR EVENT $E_{1,t}$

For an individual customer the hazard probability $\mathbf{h_{1,t}}$ gives the probability that $\mathbf{E_{1,t}}$ occurs given that $\mathbf{E_{1,t'}}$ has not occurred where t' < t. The discrete time hazard model is a logistic regression that estimates $\mathbf{h_{1,t}}$ as a function of predictor variables X_1, .., X_k and time t or some function of t. This model is discussed in books by Allison (2010) and Malthouse (2013). For an overview of supporting theory see Allison (1982).

The data set for the logistic regression includes the observations for each customer for the time periods where the event has not yet occurred, and for these observations the Target variable is set to 0. If the event occurs an observation is added for this time period, and the Target variable is set to 1.

An illustration of a discrete time hazard model data set is given below.

Example Data Set for the Discrete Time Hazard Model for Event $E_{1,t}$

Purchase histories for the five customers are shown in Table 2. In this example the time-periods in which an event can occur are "years". In Table 2 the column Y indicates, by a value of 1, that a purchase was made by the customer in the period. In this small hypothetical sample the maximum time period is T = 2.

In 2014 customer 002 purchased for the second time (observation #4). Customer 002 had already purchased for the first time in 2013 (observation #3). Observation #4 is not relevant to predicting $\mathbf{h_{1,t}}$ and this observation will be removed before proceeding to modeling. Observation #8 shows that customer 004 did not make an additional purchase in 2014 after purchasing at 2013. This observation is not relevant to predicting $\mathbf{h_{1,t}}$ and will be removed before modeling.

Table 2 Example Data Set for the Discrete Time Hazard Model for Event $E_{1,t}$

Obs	Cust_ID	Start_Date	MEMO: Year	t = Periods_Since_Start	X1	Y
1	001	12/31/2012	2013	1	0	0
2	001	12/31/2012	2014	2	0	0
3	002	12/31/2012	2013	1	1	1
4	002	12/31/2012	2014	2	1	1
5	003	12/31/2012	2013	1	1	0
6	003	12/31/2012	2014	2	1	1
7	004	12/31/2012	2013	1	1	1
8	004	12/31/2012	2014	2	1	0
9	005	12/31/2013	2014	1	0	1

The two irrelevant observations are removed by the DATA STEP and the remaining observations are saved in data set "example1b". In "example1b" the customers are observed during the periods after the start-date until either event $\mathbf{E_{1,t}}$ occurs or until the end of the observation period is reached. The target variable called New_Y is created for use by PROC LOGISTIC in order to compute $\mathbf{h_{1,t}}$. The values of NEW_Y are either 1 (for observations where $\mathbf{E_{1,t}}$ occurs) or are 0.

A predictor variable is shown in column X1.

EXAMPLE 1

```
Data example1;
length start_date $10;
input Cust_ID $ start_date $ t X1 Y;
datalines;
001 12/31/2012 1 0 0
001 12/31/2012 2 0 0
002 12/31/2012 1 1 1
002 12/31/2012 2 1 1
```

```
003 12/31/2012 1 1 0
003 12/31/2012 2 1 1
004 12/31/2012 1 1 1
004 12/31/2012 2 1 0
005 12/31/2013 1 0 1
;
Proc Sort data = example1; by Cust_ID t;
Data example1b; set example1; by Cust_ID t;
retain Cum_Y Flag;
if first.Cust_ID then do;
  Cum_Y = 0;
  Flag = 0;
  end;
Cum_Y = Cum_Y + Y;
if Cum_Y < 1 then do;
  New_Y = 0;
  output;
  end;
if Cum_Y = 1 & Flag = 0 then do;
  New_Y = 1;
  Flag = 1;
  output;
  end;
run;
```

Two observations for customer 999 are added to a new data set “example1_more” with Start-Data equal to 12/31/2015 and Years 2016 and 2017 respectively. The hazards $\mathbf{h}_{1,1}$ and $\mathbf{h}_{1,2}$ are computed for customer 999 and are printed by PROC PRINT. These hazards are predictions for two future periods.

Also, a third observation is added which reproduces the observation #4 for customer 002 in “example1”. Even though observation #4 was unused in the fitting the model, it can be scored by the fitted model.

Table 3A Data Set “example1_more”

Obs	Cust_ID	Start_Date	MEMO: Year	t = Periods_Since_Start	X1
1	999	12/31/2015	2016	1	1
2	999	12/31/2015	2017	2	1
3	002	12/31/2012	2014	2	1

```
Data example1_more;
length start_date $10;
input Cust_ID $ start_date $ t X1;
datalines;
999 12/31/2015 1 1
999 12/31/2015 2 1
002 12/31/2012 2 1
;
Proc Logistic data = example1b desc;
model New_Y = X1 t;
score data = example1_more out = predict(rename=(P_1 = hazard));
run;
Proc Print data = predict;
var Cust_ID start_date t X1 hazard;
run;
```

Table 3B Logistic Model Scores for "example1_more"

Obs	Cust_ID	start_date	t	X1	hazard
1	999	12/31/2015	1	1	0.76388
2	999	12/31/2015	2	1	0.70835
3	002	12/31/2012	2	1	0.70835

RESTRICTIONS AND ASSUMPTIONS

Table 2 is reproduced to aid in the following discussion:

Table 2 Reproduced

Obs	Cust_ID	Start_Date	MEMO: Year	t = Periods_Since_Start	X1	Y
1	001	12/31/2012	2013	1	0	0
2	001	12/31/2012	2014	2	0	0
3	002	12/31/2012	2013	1	1	1
4	002	12/31/2012	2014	2	1	1
5	003	12/31/2012	2013	1	1	0
6	003	12/31/2012	2014	2	1	1
7	004	12/31/2012	2013	1	1	1
8	004	12/31/2012	2014	2	1	0
9	005	12/31/2013	2014	1	0	1

Mixing observations with different start-dates: In Table 2 the observation #9 for customer 005 with start-date 12/31/2013 contributes to the calculation of $\mathbf{h}_{1,1}$. The other observations where t = 1 (which all have a start-date of 12/31/2012) also contribute to the calculation of $\mathbf{h}_{1,1}$. This grouping of observations with different start-dates is appropriate in many applications. But this practice also raises an issue: Customer 002 buys in 2013 and Customer 005 buys in 2014. Both records have t = 1. But markets in 2013 and 2014 might have very dissimilar characteristics so that these observations would not be comparable for the purpose of fitting the model.

A work-around is to add a dummy variable with value 1 for observations with a start-date of 12/31/2012 and set this dummy to equal 0 when making predictions. But this work-around adds to model complexity and still may not resolve the possible distortion. (There may be interactions that are not captured by this dummy variable.)

But more importantly the combining of records with different start-dates for CRM applications is not necessary to achieve sample size goals. Large-size Company customer databases provide large samples of customer records for any given start-date.

No censoring needed: In a survival study in medical science there are subjects that drop out before the end of study. For these subjects it will be unknown whether the target event occurred before the end of the observation period. These subjects are censored in the sense that no observations are added to the discrete time hazard data set after the subject's drop-out date. In many CRM applications it is unlikely that the Company would be aware of the customer "dropping out" from being a potential buyer.[1] From the point-of-view of Marketing the customer simply remains a non-buyer. In this situation each customer will have an observation for each period from start to end.

NO MIXING AND NO CENSORING: From this point forward it is assumed that all customers have the same start-date and have an observation for each time period from t = 1 through t = T. **In brief, no mixing and no censoring.**

Let $\mathbf{f}_{i,j,t}$ give the probability of event $\mathbf{E}_{i,j,t}$ for the i^{th} customer. The $\mathbf{f}_{i,j,t}$ (fixed j) give the probability density functions for t = 1 to T for the time to first event (which is j=1), the probability density for the time to second event (which is j=2), etc. Customer-level subscripts "i" in notation for events $\mathbf{E}_{i,j,t}$ and probabilities $\mathbf{f}_{i,j,t}$ are omitted after this point.

[1] The Company would become aware of a drop-out if the customer has a billing relationship with the Company. Such companies include telephone service providers, credit card companies, and so on.

Goal of this Paper: To discuss models which produce $\mathbf{f_{j,t}}$ $(X_1, .., X_k)$ for individual customers as a function time and of predictor variables $X_1, ..., X_k$. For a given customer the values of the predictors $X_1, ..., X_k$ can vary with time and also interact with time.[2]

A PROPOSAL FOR A DISCRETE TIME HAZARD MODEL FOR EVENTS $E_{j,t}$

An extension of the discrete time hazard model for the time to the first event is proposed for computing the hazards $\mathbf{h_{j,t}}$ for the events $\mathbf{E_{j,t}}$ where j = 1, 2, …, J

For a fixed j a data set is prepared as follows: All customers are observed during the periods after the start-date only until either $\mathbf{E_{j,t}}$ occurs or until the end of the observation period is reached. The values of the target New_Yj are either equal to 1 for observations where $\mathbf{E_{j,t}}$ occurs or are equal to 0. PROC LOGISTIC is used with the target variable New_Yj and with predictors involving both "time" and characteristics of the customer.[3]

A total of J models is required for the discrete time hazard model for events $\mathbf{E_{j,t}}$ where j = 1 to J.

Dummy Variable Parameterization of Time

As described above the hazard models for events $\mathbf{E_{j,t}}$ can incorporate time t as a predictor in any way that is effective, such as linear, cubic splines, or dummy variable parameterization. But to obtain an optimal property, to be described by Equations A and B later in the paper, it is necessary to use the dummy variable parameterization of time t. This parameterization is used in Example 2 below.

An Example to Illustrate the Hazard Model for Events $E_{j,t}$

Example 2 illustrates the fitting of discrete time hazard models for the events $\mathbf{E_{j,t}}$ where j = 1 to 4 and T = 4. In Example 2 the data set "example2" has no mixing and no censoring. The columns in data set "example2" include a variable t that counts the time-periods since the start-date, a predictor X1, and a variable Y with Y=1 if an event (purchase) occurs in the time period, else Y = 0.[4]

The value of X1 depends on t, being inside the t do-loop. We might assume for this example that these values of X1 are lagged by T = 4 time-periods so that hazard models fitted to this data set could be used in making predictions.

EXAMPLE 2

```
Data example2;
do Cust_ID = 1 to 1000;
  do t = 1 to 4; /* T = 4 */
    cumLogit = ranuni(3);
    e = 1*log(cumLogit/(1-cumLogit));
    X1 = 0.5*rannor(3);
    Y_star = ((4-t)/t) * X1**2 + e;
    Y = (Y_star > 1.5);
    output;
    end;
  end;
/*Proc Sort data = example2; by Cust_ID t;*/ /*example2 is already sorted*/
```

---

[2] For a model that is used for prediction the time-varying predictors are not available for model-scoring into future periods. However, time-dependent lagged predictor variables can be extended into the future and used for prediction.

[3] A different formulation of the data set is to observe only those customers from the time of experiencing $\mathbf{E_{j-1,t}}$ until either $\mathbf{E_{j,t}}$ occurs or until the end of the observation period is reached. This gives the "risk set" for $\mathbf{E_{j,t}}$. But this formulation does not lead to the desired relationship between the hazards and the densities $\mathbf{f_{j,t}}$ (defined at the beginning of the paper) nor to the "population" densities described later in the paper, and does not allow comparison to the multinomial logistic model which is also described later in the paper. See Allison (2010 chapter 8) for an example of repeated events modeling where separate models are fitted for each risk set.

[4] In data set example2 Y is derived from Y_star which is a latent (hidden) variable. Because Y_star = X + ε where X is a predictor variable and ε is an error term with a logistic distribution, the probability that Y = 1 is given by a logistic model. The construction of X includes a multiplier (4-t)/t which causes the probability of Y = 1 to decrease, on average, with t. The goal is to generate a Y with enough complexity to illustrate the concepts in this paper.

The preparation of the data sets "Event_j" where j = 1 to 4 for use in PROC LOGISTIC is shown below. The macro "Target" allows for re-use of some code in the DATA STEP. The DATA STEP includes code to create a dummy variable parameterization of t.

```
%Macro Target(j);
  If Flag&j = 0 & Cum_Y <= &j then do;
    if Cum_Y = &j then New_Y&j = 1;
    else New_Y&j = 0;
    output Event_&j;
    if Cum_Y = &j then Flag&j = 1;
    end;
%Mend;

Data Event_1 Event_2 Event_3 Event_4; set example2; by Cust_id t;
array New_Yj {*} New_Y1 - New_Y4;
array Flagj {*} Flag1 - Flag4;
retain Cum_Y Flag1 - Flag4;
drop i;
if first.Cust_id then do;
  Cum_Y = 0;
  do i = 1 to 4;
    New_Yj{i} = 0;
    Flagj{i} = 0;
    end;
  end;
tdum1 = (t=1); tdum2 = (t=2); tdum3 = (t=3);
Cum_Y = Cum_Y + Y;
%Target(1);
%Target(2);
%Target(3);
%Target(4);
run;
```

Below, the macro "LOGIT(j)" starts with PROC LOGISTIC and ends with a DATA STEP.

Each of the four PROC LOGISTIC's includes three dummies (tdum1 - tdum3). This gives a dummy variable parameterization of t. A CLASS statement for t is an alternative to the use of the dummies.

The hazards $\mathbf{h}_{j,t}$ where t < j must be zero by definition. But PROC LOGISTIC can be run, without adjustment, in spite of this constraint. There will be a warning from PROC LOGISTIC about quasi-separation; however, the correct model will be produced. The estimated $\mathbf{h}_{j,t}$ where t < j will be essentially zero. The modeler can re-code these $\mathbf{h}_{j,t}$ to be 0 in a later data step.

Alternatively, to remove the quasi-separation warning a "WHERE t >= j" can be added to PROC LOGISTIC. It is added as a comment in the code below.

In PROC LOGISTIC output the naming of output option "pred" as "hazard" is appropriate. For example, considering the logistic model for New_Y2, when t = 3 the observations that are fitted by PROC LOGISTIC are those where New_Y2 = 0 for t < 3 and New_Y2 = 0 or 1 for t = 3. The logistic regression is modeling the probability that event $\mathbf{E}_{2,3}$ occurs given that $\mathbf{E}_{2,1}$ and $\mathbf{E}_{2,2}$, have not occurred. This is the hazard $\mathbf{h}_{2,3}$.

The DATA STEP in the macro reads the scored data set "score_event_&j" by Cust_ID and creates a single denormalized record for each customer with the hazards $\mathbf{h}_{j,t}$ for t = 1 to 4. Also the values of $\mathbf{h}_{j,t}$ are set to zero when t < j.

```
%Macro Logit(j);
Proc Logistic data = Event_&j desc;
model New_Y&j = tdum1 tdum2 tdum3 x1;
output out = score_event_&j pred = hazard;
* WHERE t >= &j.;
run;
Data score_event_&j._denorm; set score_event_&j; by Cust_ID;
retain h&j.1 h&j.2 h&j.3 h&j.4;
array h&j._{*} h&j.1 - h&j.4;
keep Cust_ID h&j.1 h&j.2 h&j.3 h&j.4;
if first.Cust_id then do;
  do i = 1 to 4; /* T = 4 */
    h&j._{i} = .;
    end;
  end;
h&j._{t} = hazard;
if t < &j. then h&j._{t} = 0;
if last.Cust_ID then output;
run;
%Mend;

%Logit(1);
%Logit(2);
%Logit(3);
%Logit(4);
```

As an illustration of results of the program above, the first 5 observations from "score_event_2_denorm" are printed. Three hazards are not computed (see observations #2 and #5). These correspond to "irrelevant" observations. (E.g. customer 2 purchased for the second time during period 3 and, therefore, customer 2 does not contribute an observation to the fitting of $\mathbf{h}_{2,4}$.)

```
Proc Print data = score_event_2_denorm(obs=5); run;
```

Table 4 First Five Observations from "score_event_2_denorm"

Obs	Cust_ID	h21	h22	h23	h24
1	1	0	0.059404	0.080831	0.11915
2	2	0	0.059568	0.079677	.
3	3	0	0.059807	0.082499	0.10631
4	4	0	0.066134	0.085474	0.12322
5	5	0	0.067742	.	.

POPULATION HAZARDS AND POPULATION DENSITIES

Discussion of the discrete time hazard logistic model is set aside for the moment in order to define and illustrate the concepts of "population hazards" and "population densities" related to the events $\mathbf{E}_{j,t}$.

Notation:

N is the number of customers (unique Cust_ID's) in data set

$\mathbf{N}_{j,t}$ is the number customers experiencing event $\mathbf{E}_{j,t}$

$\mathbf{S}_{j,t}$: For each j: If t ≤ j then $\mathbf{S}_{j,t}$ = N and if t > j then $\mathbf{S}_{j,t} = \mathbf{N} - \sum_{t'=j}^{t-1} \mathbf{N}_{j,t'}$

$\mathbf{H}_{j,t}$: For each j: the hazard at time t is: $\mathbf{H}_{j,t} = \mathbf{N}_{j,t} / \mathbf{S}_{j,t}$

$\mathbf{F}_{j,t}$: For each j: the density at time t is: $\mathbf{F}_{j,t} = \mathbf{N}_{j,t} / \mathbf{N}$

For fixed j, the $\mathbf{H}_{j,t}$ give the probability that a randomly selected customer at time t will experience event $\mathbf{E}_{j,t}$ given that $\mathbf{E}_{j,t'}$ has not occurred where t' < t. The $\mathbf{F}_{j,t}$ give the probability that a randomly selected customer at time t will experience event $\mathbf{E}_{j,t}$.

The example below illustrates these concepts for a small data set of 6 customers who are tracked across 4 time periods.

EXAMPLE 3

Table 5A Example Of 6 Customers With Events Across t = 1,.., 4

	t			
Cust_ID	1	2	3	4
001	E	E	E	E
002		E		E
003		E	E	
004			E	E
005				E
006				

E = Event Occurred

Table 5B The Population Hazards and Densities from Table 5A

	$\mathbf{N_{J,t}}$					$\mathbf{S_{J,t}}$					$\mathbf{H_{J,t}}$					$\mathbf{F_{J,t}}$			
	t (time)					t (time)					t (time)					t (time)			
J	1	2	3	4		1	2	3	4		1	2	3	4		1	2	3	4
1	1	2	1	1		6	5	3	2		1/6	2/5	1/3	1/2		1/6	1/3	1/6	1/6
2	0	1	1	2		6	6	5	4		0	1/6	1/5	1/2		0	1/6	1/6	1/3
3	0	0	1	0		6	6	6	5		0	0	1/6	0		0	0	1/6	0
4	0	0	0	1		6	6	6	6		0	0	0	1/6		0	0	0	1/6

The following property is easy to verify:

For each j: $\mathbf{H_{j,t}}$ and $\mathbf{F_{j,t}}$ are related by: $\mathbf{H_{j,t} * (1 - H_{j,t-1}) * \ldots * (1 - H_{j,1})} = \mathbf{F_{j,t}}$.

For example, if j = 2 and t = 4, then: 1/2 * (1 – 1/5) * (1 – 1/6) * (1 – 0) = 1/3

See Appendix 1.

Discrete Time Hazard Model for Event $E_{j,t}$ and a Connection between $H_{j,t}$ and $h_{j,t}$

If hazards $\mathbf{h_{J,t}}$ are fitted by a discrete time hazard model with no mixing or censoring and with dummy variable parameterization of time (and other predictors), then:

For each j and t: Average of $\mathbf{h_{j,t}}$ across customers = $\mathbf{H_{j,t}}$ … **Eq. A**

Eq. A is derived from the likelihood functions of the discrete time hazard models. See Appendix 2.

However, Eq. A does not hold without the dummy variable parameterization of t. For example, Eq. A fails if predictor t enters only linearly (if T > 2).

Relationship Between $p_{J,t}$ for the Discrete Time Hazard Model and Population $F_{J,t}$ Density

For individual customers for fixed j the quantities $\mathbf{p_{j,t}}$ will be defined by these equations:

$$\text{if } t = 1, \text{ then } \mathbf{p_{j,t}} = \mathbf{h_{j,t}}$$
$$\text{if } t > 1, \text{ then } \mathbf{p_{j,t}} = \mathbf{h_{j,t}} \prod_{t'=1}^{t-1} (1 - \mathbf{h_{j,t'}})$$

For fixed j the $\mathbf{p_{j,t}}$ gives the probability density function, based on the discrete time hazard model, that $\mathbf{E_{j,t}}$ occurs for the customer. See Appendix 3 for a proof.

On page 5 the goal of the paper was stated to be: For each j we want models which can be used to estimate $\mathbf{f_{j,t}}$ (the probability of event $\mathbf{E_{j,t}}$ for the i^{th} customer.) The $\mathbf{p_{j,t}}$ provide an estimate of the $\mathbf{f_{j,t}}$.

Under the assumptions of Eq. A, it would be satisfying if Eq. B (below) is also true. This relationship holds for the simulation in this paper and for other simulations I have run.

For each j and t: Average of $\mathbf{p_{j,t}}$ across customers = $\mathbf{F_{j,t}}$ … **Eq. B**

I do not have an "arithmetic" proof of Eq. B (as is given for Eq. A in Appendix 2). But if we assume that $\mathbf{h_{j,t}}$ are independent random variables (for each fixed j), then the expected value of $\mathbf{p_{j,t}}$ equals $\mathbf{F_{j,t}}$ [5]

The simulations show the $\mathbf{p_{j,t}}$ are correct on average. Naturally, good predictions of $\mathbf{h_{j,t}}$ and $\mathbf{p_{j,t}}$ at the customer level depend on successful selection of predictor variables for the J models.

Using the Fitted Model to Score Unused Observations

Continuing Example 2: Earlier Table 4 showed the customer's hazards $\mathbf{h_{2,t}}$ for t = 1 to 4 for the first 5 customers. Three hazards were not computed (see observations #2 and #5). These correspond to the "irrelevant" observations. Also in Table 4 the values of $\mathbf{h_{2,1}}$ were replaced by zeros.

To create the complete probability density $\mathbf{p_{j,t}}$ for all individual customers, the fitted model (i.e. the fitted coefficients) is used to fill in values for the missing hazards $\mathbf{h_{j,t}}$ and to compute $\mathbf{p_{j,t}}$. A convenient way to do this is by using the CODE statement in PROC LOGISTIC.

This is illustrated by re-running PROC LOGISTIC for Target New_Y2 which was created in Example 2.

```
Proc Logistic data = Event_2 desc;
model New_Y2 =  tdum1 tdum2 tdum3 x1;
CODE file = "<your path>\Event_2.txt";
Data score_event_2; set example2; by Cust_ID;
retain h21 h22 h23 h24  p21 p22 p23 p24  _1_minus_h;
array h2_{*} h21 - h24;
array p2_{*} p21 - p24;
if first.Cust_ID then _1_minus_h = 1;
tdum1 = (t = 1); tdum2 = (t = 2); tdum3 = (t = 3);
%include "<your path>\Event_2.txt";
h2_{t} = P_New_Y21*(t > 1);
p2_{t} = h2_{t}*_1_minus_h;
_1_minus_h = (1-h2_{t})*_1_minus_h;
if last.Cust_ID then output;
Proc Print data = score_event_2(obs=5);
var Cust_ID h21 h22 h23 h24 p21 p22 p23 p24;
```

Table 6 Hazards and Densities for the first 5 Cust_ID's from Event_2

Obs	Cust_ID	h21	h22	h23	h24	p21	p22	p23	p24
1	1	0	0.059404	0.080831	0.11915	0	0.059404	0.076029	0.10302
2	2	0	0.059568	0.079677	0.10625	0	0.059568	0.074931	0.09196
3	3	0	0.059807	0.082499	0.10631	0	0.059807	0.077565	0.09171
4	4	0	0.066134	0.085474	0.12322	0	0.066134	0.079822	0.10523
5	5	0	0.067742	0.085254	0.10388	0	0.067742	0.079479	0.08859

Now, it is no longer true that the relationship of Eq. A between customer-level average hazards and the population hazards holds exactly. But based on simulations the approximation is very good.

For each j and t: Average ($\mathbf{h_{j,t}}$) ~ $\mathbf{H_{j,t}}$

Similarly, a very good approximation is given by:

For each j and t: Average ($\mathbf{p_{j,t}}$) ~ $\mathbf{F_{j,t}}$

For j = 2 the average $\mathbf{p_{2,t}}$ from the scoring of "Event_2" using the CODE file is shown below:

```
Proc Means data = score_event_2 mean; var p21 p22 p23 p24; run;
```

[5] Independence of $h_{j,t}$ follow from treating each customer's history as independent observations, one for each time unit. Then, $E[h_{j,t} \prod_{t'=1}^{t-1}(1-h_{j,t'})] = E[h_{j,t}] \prod_{t'=1}^{t-1}(1-E[h_{j,t'}])$. From Eq. A: $E[h_{j,t}] = H_{j,t}$ and $H_{j,t}\,(1 - H_{j,t-1})\ \ldots\ (1 - H_{j,1}) = F_{j,t}$. See Allison (1984 p. 81) for discussion of "Problems with the Discrete-Time Approach"

Table 7 Average of $\mathbf{p_{2,t}}$ by time periods t

Variable	Mean
p21	0
p22	0.0620
p23	0.0790
p24	0.0951

To compare the average $\mathbf{p_{2,t}}$ to the $\mathbf{F_{2,t} = N_{2,t} / N}$ the values of $\mathbf{N_{2,t}}$ are needed (where N = 1000). The data set Event_2 has this information.

```
Proc Means data = Event_2 sum; var New_Y2; class t; run;
```

Table 8 Sum of New_Y2 by time periods t

t	Sum
1	0
2	62
3	79
4	95

Using the values from this table the $\mathbf{F_{2,t}}$ can be computed, and they are equal to the average $\mathbf{p_{2,t}}$.

Conclusion: A fully scored hazards data set provides, for fixed j, complete customer probabilities $\mathbf{p_{j,t}}$ whose average closely approximates the population densities. That is, the $\mathbf{p_{j,t}}$ are approximately correct on average.

This is a reassuring property. It justifies the scoring of new data sets by a fitted discrete time hazard model for predicting $\mathbf{E_{j,t}}$. Naturally, good predictions of $\mathbf{h_{j,t}}$ and $\mathbf{p_{j,t}}$ at the customer level depend on the successful selection of predictor variables for the J models.

THE MULTINOMIAL LOGISTIC MODEL ALTERNATIVE

For fixed j, another approach to estimating $\mathbf{f_{j,t}}$ (the probability of $\mathbf{E_{j,t}}$ for the i^{th} customer) is presented in this section. It is based on fitting an unordered multinomial logistic model for each time-period t = 1 to T. The first step is to define the target variable ML_t where t = 1 to T for these models.

Creating the Target Variables for the Multinomial Logistic Models

For each customer here are the step-by-step instructions:

Step 0: Fix a value for t
Step 1: Count the number of events between 1 and t. Call it "C"
Step 2: Multiply C by 10. New value is 10*C
Step 3: If 10*C = 0 then stop. ML_t = 0
Step 4: If there is no event at period t, then ML_t = 10*C + 5
Step 5: Else ML_t = 10*C

The usage of the "+5" in Step 4 provides a convenient but arbitrary labeling for those patterns where there is no event at time t. Examples of computing ML_t appear in Table 9. For example, the 5 steps to compute ML_4 for Cust_ID =1 are shown below:

Step1: C = 2. Step 2: 10*C = 20. Step 3: Skip. Step 4: Skip. Step 5: ML_4 = 20.

Table 9 Target Variable Values for Multinomial Logistic Models

Cust_ID	t=1	t=2	t=3	t=4	ML_1	ML_2	ML_3	ML_4
1	0	1	0	1	0	10	15	**20**
2	1	0	0	1	10	15	15	20
3	0	1	1	0	0	10	20	25
4	1	1	1	1	10	20	30	40
5	0	0	0	0	0	0	0	0

Consolidation or Expansion of Target Values

When t = 4 there are 8 target values: 0, 10, 15, 20, 25, 30, 35, 40. In general, there are 2*t values for time-period t and, specifically, 20 values for t = 10. This is a fairly large number of target values for PROC LOGISTIC but is feasible when the model is built on a large sample.

Consolidation of target values “15” and “25” into, say, “101” would reduce the complexity of the logistic model but would adversely affect the fitting of the model since value “15” has a very different meaning than does value “25” (one event versus two).

But a value such as “15” could be expanded into, say, “12” and “18” to distinguish between the patterns of events such as 1, 0, 0, 0 and 0, 1, 0, 0. This distinction is unlikely to create important changes in the model. However, the wide-spread expanding of such responses would significantly increase the number of target values for large t and could make the multinomial logistic model approach unworkable.[6] [7]

EXAMPLE 4

The data set “example2” of Example 2 is again used. The DATA STEP computes the targets ML_t for t = 1 to 4. Then four multinomial logistic regressions are run with the single predictor X1. The logistic probabilities are written to four data sets “scored_ML_1” to “scored_ML_4”. These probabilities are named “P_<value of target>”. For instance, in the second logistic model the logistic probabilities are named P_0, P_10, P_15, P_20.

```
Data ML; set example2; by Cust_ID t;
array Yt{*} Y1 - Y4;
array X1t{*} X1_1 - X1_4;
retain Y1 - Y4 X1_1 - X1_4;
Yt{t} = Y;
X1t{t} = X1;
if last.Cust_ID then do;
  ML_1 = 0; ML_2 = 0; ML_3 = 0; ML_4 = 0;
  do i = 1 to 4;
    ML_4 = ML_4 + 10*Yt{i} + (i = 4)*(Yt{4} = 0)*(ML_4 > 0)*5;
    ML_3 = ML_3 + 10*Yt{i}*(i < 4) + (i = 3)*(Yt{3} = 0)*(ML_3 > 0)*5;
    ML_2 = ML_2 + 10*Yt{i}*(i < 3) + (i = 2)*(Yt{2} = 0)*(ML_2 > 0)*5;
    ML_1 = ML_1 + 10*Yt{i}*(i < 2);
    end;
  output;
  end;
%let rnm2 = (P_20 = q22); %let rnm3 = (P_20 = q23); %let rnm4 = (P_20 = q24);
Proc Logistic data = ML;
model ML_1(ref = "0")= X1_1 / link = glogit;
score data = ML out = scored_ML_1;
Proc Logistic data = ML;
model ML_2(ref = "0")= X1_2 / link = glogit;
score data = ML out = scored_ML_2(rename = &rnm2);
Proc Logistic data = ML;
model ML_3(ref = "0")= X1_3 / link = glogit;
```

[6] If a multinomial logistic model with a predictor X1 has target values 12 and 18, then the TEST statement in PROC LOGISTIC will test for the equality of the coefficients X1_12 and X1_18.

```
 PROC LOGISTIC DATA = WORK;
   MODEL Target(ref="0") = X1 / LINK = glogit;
   TEST X1_12 = X1_18;
```

[7] A case can be made for assuming the values of the target variable are, in fact, ordered. (“15” is greater than “10” because the event of “15” was earlier. Earlier is better than later.) If so, then a proportional odds or partial proportional odds logistic model might be used. This reduces the number of coefficients to be fitted. But, as discussed later in the paper, an important property of the general (unordered) logistic model is not guaranteed to be satisfied by an unordered model. The property: The average probability for a target value is equal to the percentage of that target value within the distribution of the target values in the sample.

```
score data = ML out = scored_ML_3(rename = &rnm3);
Proc Logistic data = ML;
model ML_4(ref = "0")= X1_4 / link = glogit;
score data = ML out = scored_ML_4(rename = &rnm4);
Proc Means data = scored_ML_2 mean; var q22;
Proc Means data = scored_ML_3 mean; var q23;
Proc Means data = scored_ML_4 mean; var q24;
Proc Freq data = ML; table ML_2 ML_3 ML_4;
run;
```

In the following discussion, as an illustration, j is fixed at 2. In the code above in the 2nd, 3rd, and 4th logistic models there was a variable renaming to rename P_20 as $\mathbf{q_{2,2}}$ $\mathbf{q_{2,3}}$ $\mathbf{q_{2,4}}$ respectively.

Table 10 will show that the averages of $\mathbf{q_{2,2}}$ $\mathbf{q_{2,3}}$ $\mathbf{q_{2,4}}$ across the customers are very close approximations to the averages of $\mathbf{p_{2,2}}$ $\mathbf{p_{2,3}}$ $\mathbf{p_{2,4}}$. Additionally, as will be discussed after this example, the averages of $\mathbf{q_{2,2}}$ $\mathbf{q_{2,3}}$ $\mathbf{q_{2,4}}$ are exactly equal to $\mathbf{F_{2,2}}$ $\mathbf{F_{2,3}}$ $\mathbf{F_{2,4}}$.

There are three sources for Table 10: (1) The average of the $\mathbf{p_{2,t}}$ are retrieved from Table 7. (2) The averages of the $\mathbf{q_{2,t}}$ are computed by the PROC MEANS (above). (3) The percentage distributions for **ML_t = 20** are found from the distributions of **ML_t** from the data set "ML" by PROC FREQ (above).

For each t the values for $\mathbf{p_{2,t}}$, $\mathbf{q_{2,t}}$, and **ML_t=20** are essentially the same.

Table 10 $\mathbf{p_{2,t}}$ From Table 7, $\mathbf{q_{2,t}}$ from PROC MEANS, **ML_t** from PROC FREQ

Variable	Mean	Variable	Mean	Value	% of total
p21	0	q21	0	ML_1 = 20	0%
p22	0.0620	q22	0.0620	ML_2 = 20	0.0620
p23	0.0790	q23	0.0790	ML_3 = 20	0.0790
p24	0.0951	q24	0.0950	ML_4 = 20	0.0950

The equality of $\mathbf{q_{2,t}}$ and "% of total for **ML_t=20"** was pre-destined by this general relationship:

For fixed t the average across customers of $\mathbf{q_{j,t}}$ is equal to $\mathbf{F_{j,t}}$

This follows from the well-known important property of logistic regression: If the target of an unordered multinomial logistic model has values x = 0, 1, … L, then the logistic probabilities P_x satisfy:

Average(P_x) = (Count of x) / Total sample

A proof follows directly from likelihood functions of the logistic models. See Hosmer, et al (2013, p.9)

Conclusion: The customer-level multinomial logistic probabilities $\mathbf{q_{j,t}}$, for fixed t are correct <u>on average</u>. This is a reassuring property. It justifies the scoring of new data sets by fitted multinomial logistic regression models for predicting $\mathbf{E_{j,t}}$. Naturally, good predictions of $\mathbf{q_{j,t}}$ at the customer level depend on the successful selection of predictor variables for the T models.

COMPARISON OF METHODS

In this section the two methods, multinomial logistic model (MLM) and discrete time hazard model (DTHM), are compared and contrasted.

- A disturbing property of the MLM is that examples can be constructed where for a fixed j and for a given customer the cumulative density exceeds 1, that is: $\sum_{t=1}^{T} q_{j,t} > 1$. An example is constructed in Appendix 4. In this example a "very poor" model is carefully constructed to produce a customer whose cumulative density exceeds 1. My suspicion is that good modeling will prevent this occurrence in practical applications. One of the steps in model validation for MLM will be to check the values of the cumulative densities for each j and for all customers. But if cumulative densities for a good model modestly exceed 1 for some customers, these densities could be trimmed so that the maximum cumulative value for these customers is, perhaps, 0.99.
- For DTHM it is mathematically not possible for $\sum_{t=1}^{T} p_{j,t} > 1$, See Thomas (1957) and Appendix 5.

- DTHM requires the time dependency to be included as dummy variable parameters in order to satisfy Equation B. For MLM the dependency on t is fully incorporated by having separate models for each time period and Equation B will be satisfied.
- MLM allows the predictor variables to be changed for each of the T time-period models. DTHM allows the predictor variables to be changed for each of the J event models.
- MLM introduces a large number of target values into a component model when t is large. This issue was previously discussed.
- Both DTHM and MLM allow predictions to T future periods. For either method prediction beyond T periods is problematic.
- For fixed j, by using $\mathbf{p}_{i,t}$ (or $\mathbf{q}_{i,t}$) various statistics of the density $f_{i,t}$ for individual customers can be estimated. For example, the median time to the j^{th} event occurs at the first t_{50} (if any) where $\sum_{t'=1}^{t50} p_{j,t'} > 0.5$.

FURTHER WORK

1. Study is needed of how DTHM or MTM performs without the no mixing, no censoring assumption.
2. Tests of goodness-of-fit and of predictive accuracy for DTHM ($\mathbf{p}_{j,t}$) and MLM ($\mathbf{q}_{j,t}$) need to be developed or be found in the literature.[8] These measures of fit and accuracy would normally be performed on a randomly selected validation sample. See Appendix 6 for comments.

CONCLUSION

Both DTHM and MTM offer workable approaches to the task of finding the $\mathbf{f}_{j,t}$ for individual customers.

SAS Global Forum 2016, Las Vegas, NV

REFERENCES

Allison, P. (1982). Discrete-Time Methods for the Analysis of Event Histories, *Sociological Methodology 1982,* Samuel Leinhardt (ed.), 61-99.

Allison, P. (2010). *Survival Analysis Using SAS: A Practical Guide, 2nd Ed*., SAS Institute Inc.

Allison, P. (2012). *Logistic Regression Using SAS*, *2nd Ed*., SAS Institute Inc.

Allison, P. (2014). Measures of Fit for Logistic Regression, *Proceedings of the SAS Global Forum 2014 Conference*, Paper 1485-2014.

Hosmer D., Lemeshow S., and Sturdivant R. (2013). *Applied Logistic Regression, 3rd Ed.,* John Wiley & Sons, New York.

Malthouse, E. (2013). *Segmentation And Lifetime Value Models Using SAS*, SAS Institute Inc.

Thomas, G. (1957). Probability of Sums of Series, *American Mathematical Monthly,* 64, 586-589.

ACKNOWLEDGMENTS

Colleagues at Magnify Analytic Solutions provided helpful suggestions in the writing of the paper: Dave Brotherton, Xinran Liu

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bruce Lund
Magnify Analytic Solutions, a Division of Marketing Associates, LLC
777 Woodward Ave, Suite 500
Detroit, MI, 48226
blund@marketingassociates.com or blund.data@gmail.com

[8] See Allison (2014) for a survey of measures of predictive accuracy and goodness-of-fit for binary logistic regression. Some of these measures may generalize to the case of multiple time periods and events.

All code in this paper is provided by Marketing Associates, LLC. "as is" without warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Recipients acknowledge and agree that Marketing Associates shall not be liable for any damages whatsoever arising out of their use of this material. In addition, Marketing Associates will provide no support for the materials contained herein.

APPENDICES

Appendix 1 – Population Density and Population Hazard

The formula is demonstrated for J = 2 and t = 3.

$H_{2,3} * (1 - H_{2,2}) * (1 - H_{2,1}) = N_{2,3} / S_{2,3} * (1 - N_{2,2} / S_{2,2}) * (1 - N_{2,1} / S_{2,1})$
$= (N_{2,3} / (N - N_{2,2} - N_{2,1})) * (1 - N_{2,2} / (N - N_{2,1})) * (1 - N_{2,1} / N)$
$= (N_{2,3} / (N - N_{2,2} - N_{2,1})) * ((N - N_{2,2} - N_{2,1}) / (N - N_{2,1})) * ((N - N_{2,1}) / N)$
$= N_{2,3} / N = F_{2,3}$

Appendix 2 – Proving Equation A

Theorem: If hazards $\mathbf{h_{j,t}}$ are fitted by a discrete time hazard model with no mixing or censoring and with dummy variable parameterization of time (and other predictors), then:

For each j and t: Average of $\mathbf{h_{j,t}}$ across customers = $\mathbf{H_{j,t}}$ … **Eq. A**

Let j be fixed and assume that the intercept in the logistic model for j is replaced by a parameterization of t that includes a dummy variable for each value of t. ($tdum_1$ … $tdum_T$).

For each k the likelihood functions have the general form (Hosmer, et al 2013, p.37):

$$\sum_{i=1}^{All\ Obs} x_{i,k} y_i = \sum_{i=1}^{All\ Obs} x_{i,k} \pi_i$$

where $x_{i,k}$ is the value of the k^{th} predictor for observation i, y_i is the target value for observation i, and π_i is the probability from the model for observation i.

Considering the left-hand-side: Let x_k refer to the predictor $tdum_{t0}$. Then $tdum_{t0}$ equals 1 for observations where t = t0 and otherwise 0. Therefore, y can contribute a "1" to the sum only when t = t0, producing the sum equal to $N_{j,t0.}$

Considering the right-hand-side: The probabilities π are only computed when t = t0, giving $\sum_{i|t=t0} \pi_i$. The likelihood equation becomes:

$$N_{j,t0} = \sum_{i|t=t0} \pi_i$$

Now each side is divided by the count of Event j survivors at t = t0 which is denoted by $S_{j,t0}$. This gives:

$$N_{j,t0} / S_{j,t0} = H_{j,t0} = \sum_{i|t=t0} \pi_i / S_{j,t0} = \text{average of } h_{j,t} \text{ across customers}$$

Appendix 3 – Connection Between Customer Hazards and Densities

Definitions using ***new*** notation for Appendix 3 only:

f(t) is the probability density across values of time t where t = 1, …
The cumulative distribution for f(t) is $F(t) = \sum_{i=1}^{t} f(i)$ and the survival function is S(t) = 1 - F(t)
The hazard function h(t) is given by h(t) = f(t) / S(t-1) … with the convention that S(0) = 1

Derivation:

1 - h(t) = [S(t-1) - f(t)] / S(t-1)
S(t-1) (1 - h(t)) = S(t)
By iterating the formula above: $\prod_{i=1}^{t}(1 - h(i)) = S(t)$
$1 - S(t) = 1 - \prod_{i=1}^{t}(1 - h(i)) = F(t)$

Therefore, f(t) = F(t) - F(t-1) = h(t) $\prod_{i=1}^{t-1}(1-h(i))$

Appendix 4 – It is possible that $\sum_{t=1}^{T} q_{j,t}$ > 1 in the MLM

EXAMPLE 5

```
Data example5;
input Cust_ID t X1 Y;
datalines;
1  1 0  0
2  1 0  0
3  1 0  0
4  1 0  0
5  1 0  0
6  1 0  0
7  1 0  1
8  1 0  1
9  1 0  1
10 1 0  1
1  2 1  1
2  2 2  1
3  2 3  1
4  2 10 1
5  2 1  0
6  2 2  0
7  2 1  0
8  2 2  0
9  2 1  1
10 2 2  1
;
Proc Sort data = example5; by Cust_ID t;
Data ML; set example5; by Cust_ID t;
array Yt{*} Y1 - Y2;
array X1t{*} X1_1 - X1_2;
retain Y1 - Y2 X1_1 - X1_2;
Yt{t} = Y;
X1t{t} = X1;
if last.Cust_ID then do;
  ML_1 = 0; ML_2 = 0;
  do i = 1 to 2;
    ML_2 = ML_2 + 10*Yt{i}*(i < 3) + (i = 2)*(Yt{2} = 0)*(ML_2 > 0)*5;
    ML_1 = ML_1 + 10*Yt{i}*(i < 2);
    end;
  output;
  end;
Proc Logistic data = ML;
model ML_1(ref = "0")= / link = glogit;
score data = ML out = scored_ML_1(rename=(P_10 = q11));
Proc Logistic data = ML;
model ML_2(ref = "0")= X1_2 / link = glogit;
score data = ML out = scored_ML_2(rename=(P_10 = q12));
Data q11_q12; merge scored_ML_1 scored_ML_2; by Cust_ID;
  sum_q11_q12 = q11 + q12;
Proc Print data = q11_q12; Var Cust_ID q11 q12 sum_q11_q12;
run;
```

Table 11 Cust_ID's 3 and 4 have cum. density > 1

Obs	Cust_ID	q11	q12	sum_q11_q12
1	1	0.4	0.17235	0.57235
2	2	0.4	0.40515	0.80515
3	3	0.4	0.69017	**1.09017**
4	4	0.4	0.99989	**1.39989**
5	5	0.4	0.17235	0.57235
6	6	0.4	0.40515	0.80515
7	7	0.4	0.17235	0.57235
8	8	0.4	0.40515	0.80515
9	9	0.4	0.17235	0.57235
10	10	0.4	0.40515	0.80515

Appendix 5 – Customer Cumulative Densities for DTHM do not exceed 1

G. Thomas (1957) shows for any j and any customer that the sum of the densities satisfies: $\sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{p_{j,t}} \leq 1$. From this point on, j is fixed and does not enter into the logic.

Here is the proof: The proof recasts the problem in terms of probabilities. Assume that $\mathbf{h_{j,t}}$ gives the probability of "heads" for the (independent) toss of the t^{th} unfair coin from a sequence of coins.

The product $\prod_{\mathbf{t'=1}}^{\mathbf{t}}(1-\mathbf{h_{j,t'}})$ is the probability that no heads occur in the first t tosses.

The product $\mathbf{p_{j,t}} = \mathbf{h_{j,t}} \prod_{\mathbf{t'=1}}^{\mathbf{t-1}}(1-\mathbf{h_{j,t'}})$ is the probability that the first head occurs at the t^{th} toss.

And so, $\sum_{\mathbf{t'=1}}^{\mathbf{t}} \mathbf{p_{j,t'}}$ is the probability of at least one head in the first t tosses.

Therefore, $\sum_{\mathbf{t'=1}}^{\mathbf{t}} \mathbf{p_{j,t'}} + \prod_{\mathbf{t'=1}}^{\mathbf{t}}(1-\mathbf{h_{j,t'}}) = 1$ for all t.

Since $\prod_{\mathbf{t'=1}}^{\mathbf{t}}(1-\mathbf{h_{j,t'}}) \leq 1$, it follows that $\sum_{\mathbf{t'=1}}^{\mathbf{t}} \mathbf{p_{j,t'}} \leq 1$ for any t.

Appendix 6

The indicator $\mathbf{e_{j,t}}$ is defined for a customer to be 1 if the j^{th} event occurred at t or 0 otherwise.

1. One approach to measuring the fit of the models lies in measuring the agreement of the average of the $\mathbf{p_{j,t}}$ ($\mathbf{q_{j,t}}$) with the average of the $\mathbf{e_{j,t}}$ on subsets of the customers. For fixed j and t the deciles (or other percentiles) of the $\mathbf{p_{j,t}}$ ($\mathbf{q_{j,t}}$) are formed and their average within each decile is computed. This average is compared to the average of the $\mathbf{e_{j,t}}$ within the decile. The closer the average within decile, the better. The total of J*T tables is required.

2. A second approach reflects the motivating reason for fitting the models (DTHM and MLM): To estimate the probability densities $\mathbf{f_{j,t}}$ for each customer across t = 1 to T for fixed j. There are two ways to compare the estimated densities to empirical densities.

   - Suppose a predictor X1 has only a few distinct values. For a fixed value of X1 and for fixed j, the average probabilities $\mathbf{p_{j,t}}$ ($\mathbf{q_{j,t}}$) are computed and are compared with the averages of $\mathbf{e_{j,t}}$ across t = 1 to T. The hoped-for outcome is that the average of $\mathbf{p_{j,t}}$ ($\mathbf{q_{j,t}}$) closely agree with average of $\mathbf{e_{j,t}}$ across t = 1 to T for each value of X1 and for each j.

   - The comparison $\mathbf{p_{j,t}}$ ($\mathbf{q_{j,t}}$) with the average of $\mathbf{e_{j,t}}$ can be applied to "covariate patterns" (subsets of observations having identical predictor values) for the patterns where there is sufficient sample size to meaningfully compute the average of $\mathbf{e_{j,t}}$.[9]

     Finally, the idea might be further extended by creating approximate covariate patterns where a pattern is defined by fixing values of discrete predictors and fixing the mid-points of ranks of the continuous predictors.

[9] The following discussion applies to DTHM. It also applies to MTM provided MTM uses the same predictors across t.