PENNSTATE

The Methodology Center
advancing methods, improving health

%TVEM_zip (Time-Varying Effect Model) SAS Macro Users' Guide

For zero-inflated Poisson models

Version 2.1.1

Runze Li

Xianming Tan

Liying Huang

Aaron Wagner

Jingyun Yang

NOTE: For normal, logistic, and Poisson distributions, a newer version of the macro is available.

Please send questions and comments to *MChelpdesk@psu.edu*.

The development of the SAS %TVEM_zip macro was supported by National Institute on Drug Abuse Grant P50-DA10075 to the Center for Prevention and Treatment Methodology. The authors of this document would like to thank Jessica Trail, Stephanie Lanza, Sara Vasilenko, and Amanda Applegate for their helpful comments.

The suggested citation for this users' guide is

Li, R., Tan, X., Huang, L. Wagner, A. T., & Yang, J. (2015). *%TVEM_zip (time-varying effect model) SAS macro users' guide* (Version 2.1.1). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Table of Contents

1. Overview of the %TVEM_zip Macro

1.1 Users' Guide Overview

This users' guide describes how to use the %TVEM_zip macro. This SAS macro can be used to estimate time-varying coefficient functions in time-varying effect models for intensive longitudinal data (ILD) when the outcome can be described by a zero-inflated Poisson model. ILD refers to longitudinal data with more frequent measurements than traditional longitudinal data, in which there are typically only a few widely spaced waves of data for each individual. Traditional analytic methods assume that covariates have constant (i.e., time-invariant) effects on a time-varying outcome. This macro estimates the *time-varying* effects of the covariates.

This guide assumes you have a working knowledge of time-varying effect models (TVEMs). Tan, Shiyko, Li, Li, & Dierker (2012) provide an introduction to TVEMs for audiences in psychological science. An empirical demonstration of the %TVEM macro appears in Shiyko, Lanza, Tan, Li, & Shiffman (2012).

Note: Version 2.1.1 includes the capability to fit TVEMs with zero-inflated Poisson distributed response variables. However, it does not include random effects; rather, it treats each observation as independent. This limitation has been removed in version 3.1.0. Version 3.1.0 handles normally distributed, logistic distributed, or Poisson distributed response variables. However, it does not work with zero-inflated Poisson distributed response variables. For this reason, continue using Version 2.1.1 to fit zero-inflated Poisson models. For other outcome distributions, please download the newest version. The latest version of the %TVEM SAS macro is available at http://methodology.psu.edu/downloads/tvem.

1.2 %TVEM_zip Macro Features

Features of the %TVEM_zip macro include

- Option to include multiple time-invariant covariates
- Option to include multiple time-varying covariates

1.3 Version Features

Important changes from version 2.1

- Minor updates to improve usability

Important changes from version 2.0

- Substantial revision of the plots produced by the macro, including the plotting of the confidence interval and the adding of a reference line

2. System Requirements

The macro requires

- SAS v. 9.x (Windows version)
- SAS/IML (to generate the B-spline basis functions)
- SAS/STAT (to estimate linear mixed effects models using PROC MIXED)

Note: SAS/IML and SAS/STAT are sold separately from the base SAS package, but most university licenses include them.

3. Time-Varying Effect Models

Time-varying effect models (TVEMs) are a natural extension of linear regression models. The fundamental difference is this: in linear regression models, a single estimate of each covariate's effect is provided, but in TVEMs the coefficients can vary over time (Hastie & Tibshirani, 1993). Intensive longitudinal data are generally collected to capture temporal changes in a process, so it is natural to expect that both the outcome and the relationships between the covariates and the outcome might change over time. TVEMs are designed to evaluate whether and how the effects of covariates change over time.

Suppose we observe intensive longitudinal data $\{(\boldsymbol{x}_{ij}, y_{ij}, t_{ij}),\ i = 1,2,\dots,n,\ j = 1,2,\dots,n_i\}$, where y_{ij} is individual $i's$ response variable measured at time t_{ij}, and $\boldsymbol{x}_{ij}$=$(x_{ij1}, x_{ij2}, \dots, x_{ijp})'$ is a corresponding p-dimensional **covariate vector**. Traditional linear models are

$$y_{ij} = \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \varepsilon_{ij}, \tag{1}$$

where we can set $x_{ij1} \equiv 1$ in order to include an intercept, and ε_{ij} is a random error term.

A TVEM is defined as

$$y_{ij} = \beta_1(t_{ij}) x_{ij1} + \cdots + \beta_p(t_{ij}) x_{ijp} + \varepsilon_{ij}, \tag{2}$$

where $\beta_1(t), \cdots, \beta_p(t)$ are unknown coefficient functions that are assumed to be smooth over time t. TVEM thus allows the effects to change over time, which may be particularly useful in the analysis of longitudinal data.

The Generalized Time-Varying Effect Model

Traditional generalized linear model theory (McCullagh and Nelder, 1989) assumes that

$$E(y_{ij}) = \mu_{ij},\ \ g(\mu_{ij}) = \eta_{ij} = \boldsymbol{\beta}' \times \boldsymbol{x}_{ij}, \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$, η_{ij} is called a linear predictor, and the function $g(\cdot)$ is called the link function, which may vary from case to case. For example, we can let $g(a) = a$ for continuous responses, let $g(a) = \log(a/(1-a))$ if the outcome y_{ij} is binary, and let $g(a) = \log(a)$ if we assume y_{ij} follows a Poisson distribution. In model (1), all coefficients in $\boldsymbol{\beta}$ are assumed to be constant over time. This generalized linear model can be extended similarly to allow its coefficients in the linear predictor to change over time t:

$$E(y_{ij}) = \mu_{ij},\ \ g(\mu_{ij}) = \eta_{ij} = \boldsymbol{\beta}'(t_{ij}) \times \boldsymbol{x}_{ij}, \tag{4}$$

where $\boldsymbol{\beta}(\cdot)=(\beta_1(\cdot),\ \beta_2(\cdot),\ldots,\beta_p(\cdot))'$ consists of p unknown coefficient functions of time that are assumed to be smooth over time t. If we let $g(\mu_{ij})=\ln(\mu_{ij}/(1-\mu_{ij}))$, the proposed model can handle binary responses, and setting $g(\mu_{ij})=\ln(\mu_{ij})$, the model could handle count responses that follow a Poisson distribution. Choosing $g(\mu_{ij})=\mu_{ij}$, we obtain the TVEM as in model (2), which could be considered a special case of the generalized TVEM.

The %TVEM_zip macro enables estimation of coefficient functions $\boldsymbol{\beta}(\cdot)$ in model (2) for a count income having a zero-inflated Poisson distribution.

Note that the %TVEM_zip macro uses all available data for every individual over time. Time-specific observations with missing values either for response *y* or for any covariates are automatically excluded. The %TVEM_zip macro can also include *time-invariant* covariates, such as gender, whose effects remain static over time. Details about the TVEM mathematical model can be found in Tan et al. (2012).

4. Using the %TVEM_zip Macro

4.1 Preparation

A SAS macro is a special block of SAS commands. First the block is defined, and then it is called when needed. Three steps need to be completed before running the %TVEM_zip macro.

1. If you haven't already done so, download and save the macro to the designated path (e.g., *S:\myfolder*).

2. Direct SAS to read the macro code from the path, using a SAS %INCLUDE statement, such as

   ```
   %INCLUDE 'S:\myfolder\Macro_TVEMv211_zip.sas';
   ```

   Note: we suppose that the SAS macro file exists in the folder S:\myfolder\. This path represents any user-specified folder. This convention will be followed in the examples and the appendices.

3. Use a `libname` statement to direct SAS to the data file. The statement should give the libname command, name the library, and then identify the path to the data. For example,

   ```
   libname sasf 's:\myfolder\';
   ```

4.2 Running the %TVEM_zip Macro and Syntax Definitions

Call the macro using a percent sign, its name, and user-defined arguments in parentheses. The macro parameters, which we will refer to as "arguments," are shown below.[1]

```
%TVEM_zip (mydata = filename,
           id = variable,
           time = variable,
           dep = variable,
           class_var = variables,
           tcov = variables,
           cov_knots = numbers,
           cov = variables,
           evenly = number,
           scale = number,
           oddsratio = number,
           outfilename = file path and name,
           cov_zip = variable
      );
```

[1] The argument `method` exists in the %TVEM_zip macro for compatibility reasons, but it is always set to the default value. It is not discussed in this users' guide.

Argument (= default)	Description	Mandatory
mydata	Input data set; should have longitudinal data structure (one row for each assessment)	Y
id	Subject's identification variable	Y
time	Measurement time variable	Y
dep	Dependent variable	Y
class_var	Classification variables (See the 2nd appendix for each macro for an example with gender and day of the week as classification variables.)	N
tcov	Covariates assumed to have time-varying coefficients. To include a time-varying intercept, a variable should be created where all values = 1, and the variable name should be included in this argument. Classification variables (class_var) should **NOT** be included in this argument. (See the 3rd appendix for each macro to include class variables that are assumed to have time-varying coefficients.)	Y
cov_knots	Number of inner knots for each covariate in the tcov argument. (See section 5.5 for instructions on specifying knots.)	Y
deg (=3)	Degree of B-spline functions, default is cubic spline (e.g., deg=3; see chapter 7 for technical details.)	N
cov	Covariates assumed to have constant coefficients	N
evenly (=0)	Positions of inner knots. Two methods are available. One (evenly=1) evenly spaces the inner knots over the range of measurement times of all observations. The other (evenly=0) positions the knots on evenly distributed quantiles of these observations. Default value = 0. (See chapter 7 for technical details.)	N
scale (=100)	Number of points to be plotted in the graphs of the estimated coefficient functions and their confidence bands. Default value =100.	N
outfilename	Output file and path name. The macro generates a .csv file with the path *and* name specified by this parameter. This file contains the data for plotting the coefficient curves and their confidence bands for more specific use. By default, the file is saved in the same folder where your SAS program file is located as "plot_data.csv".	N
cov_zip	Variable(s) generating the excessive zeros in the count response	Y

4.3 Output

When generated, the output will be located in the same folder on your computer where your SAS program file is located. In the output, only the fit statistics and the plots of the curves with CI should be interpreted. **The rest of the output including p-values, should NOT be interpreted.** The output for time-invariant covariates will still be accurate.

The output for this macro contains the standard output for PROC GENMOD. Model fit statistics like AIC and BIC are included. The macro will output a data set (users can specify where this data set is saved) containing the estimates of all the time-varying covariates and their 95% confidence intervals at various measurement times. For each time-varying covariate, this information is also presented in plots to show how the estimates change over time.

5. Empirical Demonstration of the %TVEM_zip Macro

In this section, we first describe the structure of the data sets and the variables that the %TVEM_zip macro uses. Then, we illustrate how to fit a model using the %TVEM_zip macro. We also include example output. For more examples with the %TVEM_zip macro, see the appendices in chapter 6.

5.1 Example Data

First, we will examine the structure of the database and the variables to be analyzed. The data below are sample observations from the SAS data set **simulated_data_zip.sas7bdat** (from Appendix 6.1):

Subj	Y	T	X_0	X_1	X_2
1	0	0.006122	1	1.155575	0.082545
1	3	0.010204	1	-1.70564	-0.00701
1	0	0.014286	1	0.644903	-1.2941
1	0	0.018367	1	-1.23753	-1.07619
1	3	0.022449	1	1.057264	1.43559
2	4	0.002041	1	0.712153	-0.23305
2	0	0.006122	1	1.369241	1.249073
2	0	0.010204	1	-0.02551	0.506621
2	0	0.014286	1	1.082895	-1.03818
2	0	0.018367	1	1.849346	0.197081

where

subj= subject's identification variable

y= the dependent variable

t= measurement time for each observation (NOTE: different subjects may have different measurement times)

$\mathbf{x_0}$= the intercept variable (equal to 1 for all observations)

$\mathbf{x_1}$= a covariate)

$\mathbf{x_2}$= another covariate

5.2 Example Model and Syntax

Include a "libname" statement prior to running the macro to direct SAS to the data file.

```
libname sasf 'S:\myfolder\';
```

Tell SAS to read the macro file.

```
%INCLUDE 'S:\myfolder\Macro_Tvem_v211_zip.sas';
```

Next, we will fit the model and enter the proper syntax in SAS. The following TVEM is to fit the data:

$$y_{ij} = \beta_0(t_{ij}) \times x_{ij0} + \beta_1(t_{ij}) \times x_{ij1} + \beta_2(t_{ij}) \times x_{ij2} + \varepsilon_{ij} \quad (3)$$

using the following syntax

```
%TVEM_zip( mydata = sasf.simulated_data_zip,
        id = subj,
        time = t,
        dep = y,
        tcov = x0 x1 x2,
        cov_knots = 6 6 6
        cov_zip = x1
);
```

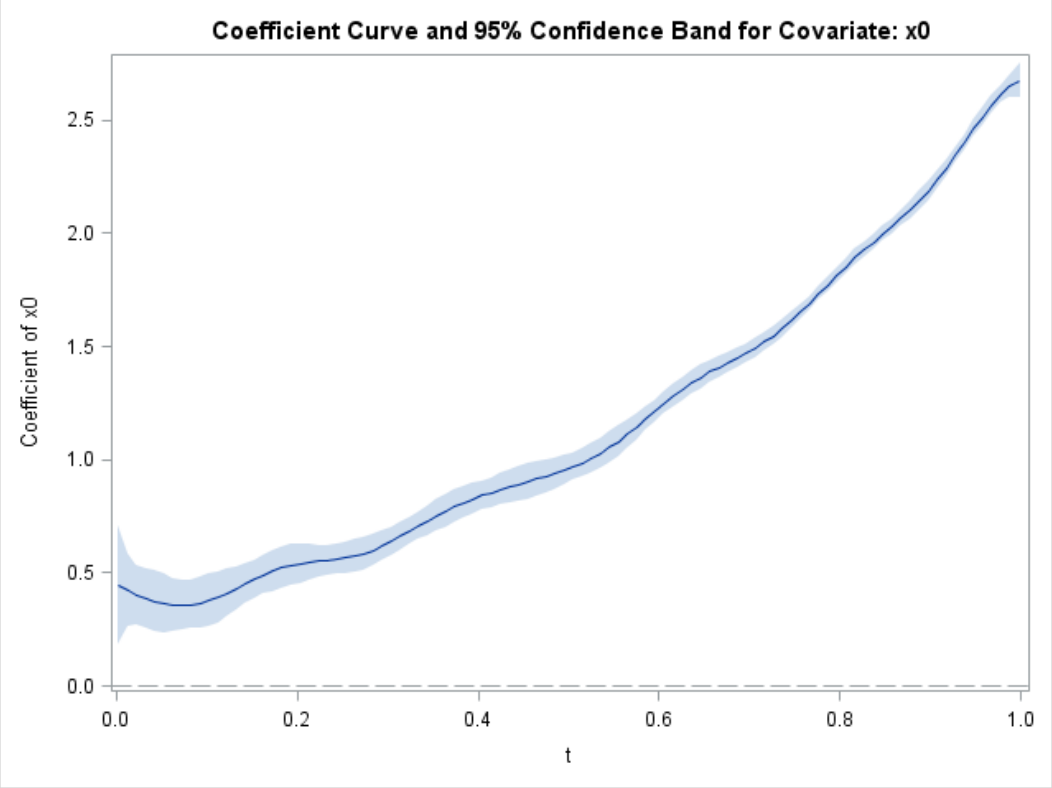
The meaning of the first four arguments is self-evident. The argument **tcov** lists those covariates that are assumed to have time-varying coefficients. The argument **cov_knots** indicates the number of inner knots to be used in the estimation of the corresponding coefficient function. Inner knots determine the number of basis functions; the technical detail in chapter 7 explains the meaning of inner knots and how they affect the estimation. See section 5.5 below for details on how to determine the proper value for the **cov_knots** argument. The **cov_zip** argument indicates the covariate(s) relevant to the probability of excessive zeros in the response (see appendices 6.2-6.4).

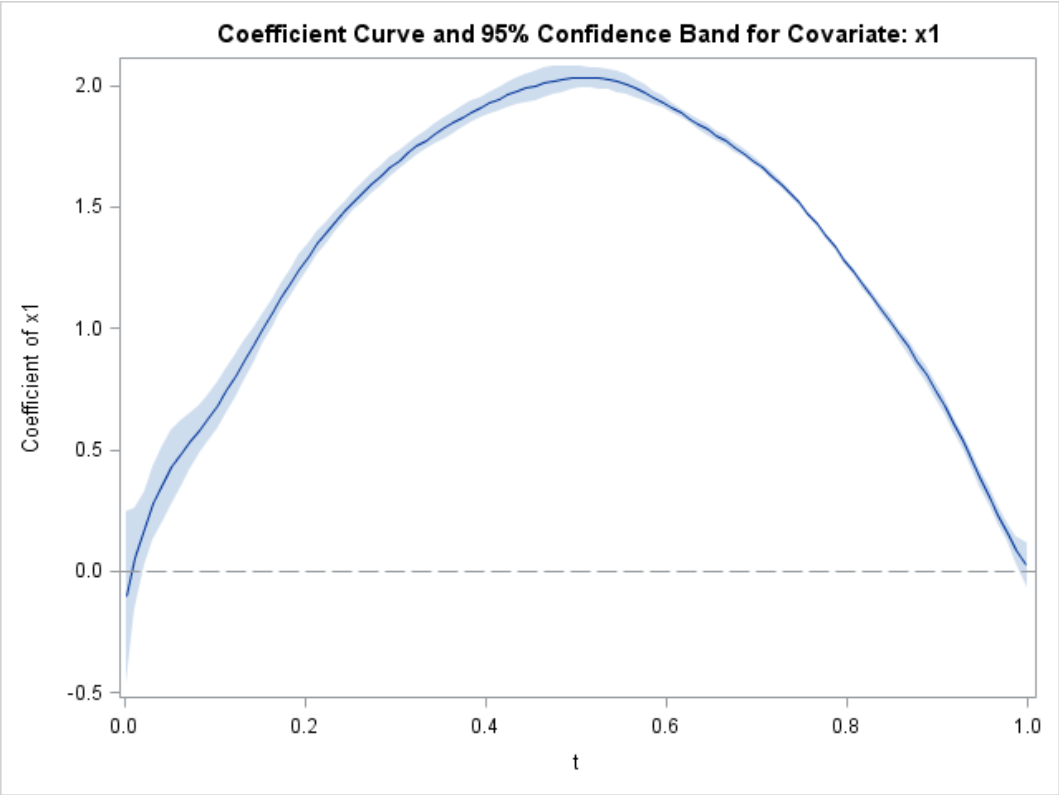
5.3 Example Output

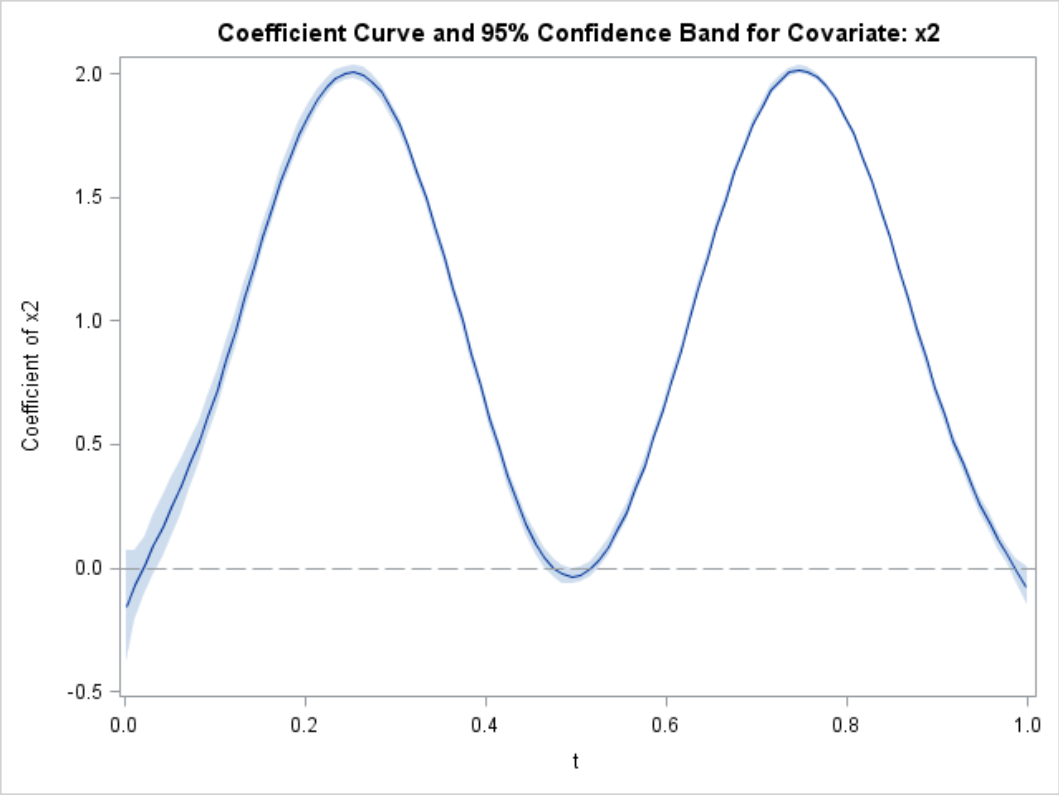
For the %TVEM_zip macro the interpretable part of the output is the same: the fit statistics and the plots of the curves with 95% CI**. The rest of the output, including p-values, should NOT be interpreted.** The output includes a plot for the estimated curve for each coefficient. The fit statistics from PROC GENMOD can be interpreted, but **the standard output for PROC GENMOD cannot be interpreted.**

Model fit statistics like AIC and BIC are included (see below), as are three plots of the coefficient functions, $\beta_0(\cdot), \beta_1(\cdot), \beta_2(\cdot)$ with 95% confidence bands (see next page).

Criterion	DF	Value	Value/DF
Deviance	_	25737.6014	_
Scaled Deviance	_	25737.6014	_
Pearson Chi-Square	9744	9777.6390	1.0035
Scaled Pearson X2	9744	9777.6390	1.0035
Log Likelihood	_	540800.2430	_
Full Log Likelihood	_	-12868.8007	_
AIC (smaller is better)	_	25825.6014	_
AICC (smaller is better)	_	25826.0079	_
BIC (smaller is better)	_	26141.9136	_




Coefficient Curve and 95% Confidence Band for Covariate: x1
Coefficient of x1
2.0
1.5
1.0
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
0.8
1.0
t


Coefficient Curve and 95% Confidence Band for Covariate: x2
Coefficient of x2
2.0
1.5
1.0
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
0.8
1.0
t

5.4 About Plotting Interactions

The %TVEM_zip macro can be used to plot interactions. This could be done by modifying the syntax above as listed below.

```
%TVEM_zip( mydata = sasf.simulated_data_zip,
        id = subj,
        time = t,
        dep = y,
        tcov = x0 x1 x2 x1_x2,
        cov_knots = 10 10 10 10,
        cov_zip = x1
);
```

Please note that the interaction plot generated would show *how the effect of the interaction varies over time*. If you are interested in the interaction of one variable (e.g., x1) with another categorical variable (e.g., x2) on the response variable (e.g., y), then you could draw a plot of x1 vs. y separately for the different values of x2. Note that the interaction variable must be created in the dataset prior to running the macro; the macro will NOT accept X1 * X2 as input if it is not created in the dataset.The example output above s for the model without an interaction.

5.5 About Spline Bases

The %TVEM_zip macro uses B-spline basis functions to approximate unknown coefficient functions. **P-spline basis functions are NOT available in %tvem_zip.** The difference between using B-spline basis functions and using truncated power spline basis functions is described in the %TVEM macro v. 3.1.0 and higher users' guides. Information on B-spline basis functions can be found in Chapter 7.

Specification of number of inner knots

Users need to specify the argument **cov_knots** (i.e., number of inner knots for each coefficient function). When using B-spline basis functions to approximate the unknown coefficient functions, countering over-fitting caused by large number of knots is not as straightforward as it is when using a truncated power spline basis. We propose the following procedure, which works quite well in our experience, for the selection of optimal numbers of inner knots:

1) Set each number of inner knots equal to a given number (e.g., 5);

2) Estimate the time-varying coefficient model, and record the AIC and BIC;

3) Repeat steps 1 and 2, but increase the number of inner knots by one, until the number of inner knots reach a large enough number (e.g., 10); and

4) Compare AICs and BICs for different numbers of inner knots, and select the model with the optimal AIC and/or BIC.

6. Appendices for %TVEM_zip

Appendix 6.1. A Brief Introduction to Zero-Inflated Poisson Model for Count Outcomes

In statistical analysis of count data, often there are more zeros than expected from a Poisson distribution. The zero-inflated Poisson (ZIP) model is a useful approach to model this scenario. It assumes that the count is generated by two processes, with process one generating only zeros and process two being a Poisson process. For each observation, a Bernoulli trial determines which process is used such that process one is chosen with probability φ_i and process two with probability 1- φ_i. More specifically, let y_i be the count for observation i. We have

$$y_i = \begin{cases} 0 & \text{with probability } \varphi_i \\ g(y_i \mid x_i) & \text{with probability 1-}\varphi_i \end{cases},$$

where $g(y_i \mid x_i)$ is given by the Poisson distribution with mean being μ_i,

$g(y_i \mid x_i) = \dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, $y_i = 0,1,2\cdots$ and x_i is the vector of covariates.

It is often reasonable to assume that the probability φ_i depends on some characteristics (z_i) of observation i and is written as a function of $z_i'\gamma$, $\varphi(z_i'\gamma)$, where γ is the vector of zero-inflated coefficients to be estimated. The function linking $z_i'\gamma$ with the probability φ_i is called the link function. Logistic function (the logit function) is used here as as the link function.

Then the probability of $\{Y_i = y_i \mid x_i\}$ becomes

$$P\{Y_i = y_i \mid x_i\} = \begin{cases} \varphi(\gamma' z_i) + \left[1-\varphi(\gamma' z_i)\right] g(0 \mid x_i) & \text{if } y_i = 0 \\ \left[1-\varphi(\gamma' z_i)\right] g(y_i \mid x_i) & \text{if } y_i > 0 \end{cases}.$$

In using ZIP to model count data, users have to specify the covariate(s) z_i. In the %TVEM_zip macro, this is done by specifying variable(s) for the cov_zip argument.

Appendix 6.2. A Web-Based Survey Example for %TVEM_zip

Note: the following example is based on sample data that was included in the download.

This example was inspired by the UpTERN study (Tiffany et al. 2006). We assumed that

52 participants were asked to complete reports on their daily cigarette smoking, alcohol consumption, and depression scores for the previous week. In total, 245 days (35 weeks) of data were generated with an average response rate for the sample slightly over 90% for each week of the study. In addition, the gender of each subject and the day of week for each of the 245 days were included in the data set **web_survey_zip.sas7bdat**.

Some observations in this data set:

Ind	day	d_cig	smoking	Int	d_alc	depr	gender	weekday
1	247	0	0	1	0	7	2	5
1	248	1	1	1	1	7	2	6
1	249	0	0	1	0	7	2	0
1	250	1	1	1	6	7	2	1
8	5	3	1	1	3	9	1	1
8	7	0	0	1	0	9	1	3
8	8	0	0	1	0	9	1	4
8	9	0	0	1	0	9	1	5

As in the empirical example in Appendix 8.2, we speculate that there is a similar relationship between the number of cigarettes smoked per day (y_{ij}) and other covariates (weekday, gender, alcohol drinks and depression) except that we believe that there are excessive zeros in y_{ij}. For illustrative purpose, we assume that the probability φ_i depends on the number of alcoholic drinks (d_alc) a participant had. The %TVEM_zip macro can be called as follows:

```
%TVEM_zip(
        mydata = sasf.web_survey_zip,
        id = idn,
        time = day,
        dep = d_cig,
        class_var = weekday gender,
        tcov = int d_alc depr,
        cov_knots = 7 7 7,
        cov = weekday gender,
        cov_zip=d_alc
 );
```

The output (not shown) will include the estimated time-constant effects of weekday and gender on number of cigarettes smoked, which represent how smoking behavior varies across days of the week and between genders. The plot of the coefficient function β_1 (·) shows how the association between smoking and drinking changed over time, and the plot of the coefficient β_2 (·) shows how the association between depression and smoking changed over time.

Appendix 6.3. A Web-Based Survey Example for %TVEM_zip (continued)

It is possible that the effect of a class variable may change over time. For example, in the web-based survey example from Appendix 6.3, we may want to study whether the effect of gender changes over time. Or, equivalently, we may think that the baseline curve of female students will be different than that of male students. In the zero-inflated Poisson models, we may have the following equations for the Poisson distribution:

$\eta_{ij} = \alpha_m(t_{ij}) \times g_i + \beta_1(t_{ij}) \times x_{ij1} + \beta_2(t_{ij}) \times x_{ij2} + \gamma \times w_{ij} + e_{ij}$, (for male students),

$\eta_{ij} = \alpha_f(t_{ij}) \times g_i + \beta_1(t_{ij}) \times x_{ij1} + \beta_2(t_{ij}) \times x_{ij2} + \gamma \times w_{ij} + e_{ij}$, (for female students).

Note that the equations above indicate that male students and female students share the common coefficient functions for alcohol use $\beta_1(\cdot)$ and depression score $\beta_2(\cdot)$, although they have different baseline curves $(\alpha_m(\cdot), \alpha_f(\cdot))$.

To fit models specified by the two equations above, we proceed as follows:
Step1: Define two new variables (because the class variable "gender" has 2 levels), say gender_M and gender_F are derived from the variable "gender," as shown in the following SAS code:

```
DATA new_web_survey;
SET sasf.web_survey_zip;
IF gender = 1 THEN DO; gender_M=1; gender_F=0; END;
IF gender = 2 THEN DO; gender_F=1; gender_M=0; END;
RUN;
```

Note: when a class variable has $k(> 2)$ levels and we want to estimate the baseline curves corresponding to different levels, we could define $k(> 2)$ new variables corresponding to the $k(> 2)$ levels of the class variable.

Step 2: Run the %TVEM macro as follows:

```
%TVEM_zip(
        mydata = new_web_survey,
        id = idn,
        time = day,
        dep = d_cig,
        class_var = weekday gender,
        tcov = gender_M gender_F d_alc depr,
        cov_knots = 7 7 7 7,
        cov = weekday gender,
        cov_zip=d_alc
);
```

Compared to the corresponding SAS code in Appendix 6.2, this SAS code removes "gender" from the list of class variables (**class_var**) and "int" (intercept variable) from the list of variables with time-varying coefficients (**tcov**). This code adds **"**gender_M" and "gender_F" to **tcov**. (Output not shown.)

7. Technical Details on B-Spline Basis Functions

Without loss of generality, we consider the following generalized time-varying coefficient model:

$$E(y_{ij}) = \mu_{ij}, \quad g(\mu_{ij}) = \eta_{ij} = \beta(t_{ij}) \times x_{ij}. \tag{11}$$

The process of estimating the unknown coefficient function $\beta(\cdot)$ involves approximating this function with certain combinations of B-spline basis functions which are determined by knots and degree. Given m+1 knots, say, $t_0 \le t_1 \le t_1 \le \cdots t_m$, the m-d basis B-splines of degree d can be defined using the **Cox-de Boor recursion formula** (de Boor, 1972) with

$$b_{j,0}\{t\} = \begin{cases} 1, & t_j \le t \le t_{j+1} \\ 0, & t < t_j \text{ or } t \ge t_{t+1} \end{cases}$$

and

$$b_{j,d}(t) = \frac{t - t_j}{t_{j+d} - t_j} \times b_{j,d-1}(t) + \frac{t_{j+d+1} - t}{t_{j+d+1} - t_{j+1}} \times b_{j+1,d-1}(t), \text{ for } j = 0,1,\cdots,K-d-1.$$

In addition, knots $t_{d+1} \le t_{d+2} \le \cdots \le t_{m-d-1}$ are called internal knots (or inner knots). In %TVEM_zip macro, the knots t_0, $\cdots$, t_d and t_{m-d}, $\cdots$, t_m are determined by the minimal ($t_{(min)}$) and the maximal ($t_{(max)}$) observation time in the input data set as follows:

$$t_j = \begin{cases} t_{(min)} - (d + 1 - j) \times \varepsilon, & j = 0, 1, 2, \cdots, d \\ t_{(min)} + (j - m + d) \times \varepsilon, & j > m - d, \end{cases}$$

where ε is a small positive number which is set at 10^{-12} in this macro. In addition, we employ cubic splines (d = 3) in this macro, as many applications do.

Given the number of inner knots, the inner knots are equally distributed over the range of study period or uniformly distributed on quantiles of measurement times (depending on the designation in the **evenly** parameter). For example, suppose that the study period is from year 0 to year 1. If we use four knots and distribute them equally over the study period, the inner knots will be at year 0.2, 0.4, 0.6, and 0.8. Or, if we distribute the 4 inner knots uniformly on the quantiles of measurement times, they will be at the 20%, 40%, 60% and 80% quantiles of the pooled measurement times.

Hence, when we input the number of inner knots = k (using the **cov_knots** parameter), and we define the method to posit these inner knots (using the evenly parameter), %TVEM_zip macro will calculate the k inner knots, and then all the $k + 2 \times (d + 1)$ knots, and then $k + d + 1$ B-spline basis functions by using the Cox-de Boor recursion formula. For example, if k = 4, then there are 8(= 4+3+1) cubic B-spline basis functions.

To estimate model (11), we approximate $\beta(\cdot)$ by k +3+1 cubic B-spline basis functions:

$$\beta_0(t) \approx \sum_{k=0}^{K+3} a_j b_{j,3}(t) \tag{12}$$

where $b_{j,3}(\cdot)$, j= 0, 1, ⋯, k+3, are cubic B-spline basis as defined by the Cox-de Boor recursion formula, and a_j, j= 0, 1, ⋯, k+3, are unknown parameters. In this way, we transfer the problem of estimating the function $\beta(\cdot)$ into a problem of estimating a_j, j= 0, 1, ⋯, k+3. Combining (11) and (12), we get the following linear regression model:

$$E(y_{ij}) = \mu_{ij}, \qquad g(\mu_{ij}) = \eta_{ij} = \sum_{j=0}^{k+3} a_j \times b_{j,3}(t_{ij}) \times x_{ij} + \varepsilon_{ij}.$$

The estimate of a_j, j= 0, 1, ⋯, k+3, can be obtained using available software, such as the PROC GENMOD or the GLM package in R. We use PROC GENMOD in our macro. The number of inner knots determines the number of basis functions used to approximate $\beta(\cdot)$. Intuitively, the larger the number of inner knots, the better the approximation. However, using too many basis functions could cause over-fitting, which can cause near interpolation of the data and undesirable "wiggly" curves. So, we need to select the ideal number of inner knots to ensure good approximation and avoid over-fitting. This can be done by using different numbers of inner knots and running the model repeatedly. Then, we select the optimal number based on the model fit statistics (AIC and/or BIC) provided in the output.

References

Buu A., Johnson, J. J., Li, R., & Tan, X. (2010). New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*, *30*(18), 2326-2340.

de Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory, 6*, 50–62.

Erdman, D., Jackson, L., & Sinko A. (2008). Zero-inflated Poisson and zero-inflated negative Binomial models using the COUNTREG procedure. *SAS Global Forum 2008*, paper 322.

Hastie, T. J., & Tibshirani, R. J. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society B*, *55*, 757-796.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Shiyko, M. P., Lanza, S. T., Tan, X., Li, R., & Shiffman, S. (2012). Using the time-varying effect model (TVEM) to examine dynamic associations between negative affect and self-confidence on smoking urges: differences between successful quitters and relapsers. *Prevention Science.* Advance online publication. doi: 10.1007/s11121-011-0264-z

Tan, X., Shiyko, M. P., Li, R., Li, Y., & Dierker, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychological Methods, 17,* 61 77.

Tiffany, S. T., Agnew, C. R., Maylath, N. K., Dierker, L., Flaherty, B., Richardson, E., ..., Tobacco Etiology Research Network (TERN). (2007). Smoking and college freshmen: University project of the Tobacco Etiology Research Network (UpTERN). *Nicotine & Tobacco Research, 9*(S4), S611-S625.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, *18*, 223-249.