PennState The Methodology Center
advancing methods, improving health

Weighted TVEM SAS Macro Users' Guide
Version 2.6

John Dziak
Runze Li
Aaron Wagner
The Methodology Center, Penn State

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

Suggested citation for this users' guide:
Dziak, J. J., Li, R., & Wagner, A. T. (2017). *WeightedTVEM SAS macro users' guide* (Version 2.6). University Park, PA: The Methodology Center, Penn State. Retrieved from methodology.psu.edu

The developers thank Deb Kloska, Sara Vasilenko, and Rebecca Evans-Polce for their assistance with testing. This work was supported by Award Number P50 DA039838 from the National Institute on Drug Abuse (NIDA). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIDA or the National Institutes of Health.

Contents

1 Introduction

1.1 Overview

The %WeightedTVEM macro was developed by the Methodology Center for fitting time-varying effects models (TVEMs; see Hastie & Tibshirani, 1993; Shiyko, Lanza, Tan, Li, & Shiffman, 2012; Tan, Shiyko, Li, Li, & Dierker, 2012) on complex datasets, such as Add Health, a nationally representative survey dataset (see Chen & Chantala, 2014). Such datasets may involve clustering (e.g., students are nested within schools) and survey weights (e.g., participants represent different numbers of population members due to systematically unequal probabilities of selection). Datasets may be cross-sectional (e.g., one observation per student) or longitudinal (e.g., multiple related observations per student over time). **This brief users' guide assumes familiarity with TVEMs and with the original %TVEM macro without survey weights** (TVEM SAS Macro, version 3.1, 2015; see Li, Dziak, Tan, Huang, Wagner, & Yang 2015). Below we explain how the %WeightedTVEM macro differs from the usual macro.

1.2 Differences between the %TVEM and %WeightedTVEM macros

The standard %TVEM macro does not accommodate survey weights or clustering other than the clustering of observations within subjects. The %WeightedTVEM macro is designed for situations in which these complex survey features are important. Failing to include survey weights of clusters when they are appropriate can result in biased standard errors and incorrect conclusions.

Currently, the %WeightedTVEM macro supports continuous or binary outcomes; count outcomes are not supported as they are in %TVEM. Also, the options available for fitting the data are currently not as rich as in the %TVEM macro, version 3.1.0. The %TVEM macro allows either unpenalized B-splines (with or without random effects), or penalized truncated power splines ("P-splines") without random effects but with robust standard errors. The %WeightedTVEM macro only supports unpenalized B-splines without random effects, although it provides robust standard errors. The robust standard errors used here are constructed using the sandwich formula (Taylor linearization) as in generalized estimating equations (GEEs; Liang & Zeger, 1986).

The main computations of the standard %TVEM macro, version 3.1, are performed in SAS PROC GLIMMIX, which is not designed for survey data. Although PROC GLIMMIX does have a WEIGHT statement, its weighting system is intended not for sampling weights, but instead for cases believed to have systematically differing measurement error variances (see Chen and Chantala 2014). Because of this limitation, the %WeightedTVEM macro uses the SURVEYREG and SURVEYLOGISTIC procedures instead of PROC GLIMMIX. In the context of longitudinal data, the SURVEYREG and SURVEYLOGISTIC procedures take a marginal (population-average) rather than multilevel (subject-specific) approach. More specifically, they use working-independence GEE with robust standard errors, instead of random effects, to take within-cluster and within-subject correlation into account. The marginal and multilevel approaches in the context of survey data are compared by Heeringa, West, and Berglund (2010) and Chen and Chantala (2014).

2 System requirements

The %WeightedTVEM macro requires

- SAS version 9.2 or above
- SAS/IML (to generate B-spline basis functions)
- SAS/STAT (to estimate linear mixed effects models using PROC SURVEYREG and PROC SURVEYLOGISTIC)

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

The macro has not been extensively tested on versions of SAS for operating systems other than Microsoft Windows, but may function there. One of the plotting options offered by the %TVEM macro requires Java and the SAS SGRENDER procedure, but a simpler plotting option that requires only the usual SAS GPLOT procedure is available.

3 Model fit in the %WeightedTVEM macro

As is typical for GEE procedures, the SURVEYREG and SURVEYLOGISTIC procedures allow only a single level of nesting to be explicitly specified. At first, this might appear to make it impossible to do regression analysis (with or without time-varying effects) clustered longitudinal datasets such as Add Health because they include at least two levels of nesting (multiple measurements per student and multiple students per school). The workaround we recommend is to specify simply the highest level of clustering (school) and ignore person-level clustering. This is because GEE requires an assumption that clusters are independent, but does not necessarily require correct specification of within-cluster correlation in order to obtain unbiased estimates. A pair of observations from two different schools can be treated as independent. Two observations from the same school are indeed more correlated if they come from the same student than if they come from different students, but we leave the nature of this unspecified. Further research, perhaps using simulations, might be beneficial in verifying that this strategy is valid and reasonably efficient; however, it is the most correct longitudinal analysis that can be done using the survey regression procedures currently available in SAS. Also, questions involving TVEM with survey data will often be basically marginal (i.e., population-averaged rather than subject-specific; see Hu, Goldberg, Hedeker, Flay, and Pentz, 1998, for a discussion of these terms) in nature, so (in the binary case particularly) the results of a GEE approach may be more straightforward to interpret than those of a random effects approach would be.

Specifically, for observation y_{ij} taken within cluster i at time t_{ij}, let $E(y_{ij})$ denote the expected value for the response variable of interest (averaging across clusters such as schools and/or students). We assume that each such observation has a corresponding p-dimensional covariate vector $\boldsymbol{x}_{ij}$=$(x_{ij1}, x_{ij2}, \dots, x_{ijp})'$. Usually, $x_{ij1} = 1$ for all observations in order to provide an intercept term. For a linear model, we assume

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

and for a logistic model, we assume

$$\mathrm{logit}(\mathrm{E}(y_{ij})) = \beta_1(t_{ij})x_{ij1} + \cdots + \beta_p(t_{ij})x_{ijp} \ .$$

As with the usual TVEM macro, some of the covariates can be assumed to have time-invariant effects (that is, some coefficient functions $\beta_k(t)$ can be specified to be constants in *t*).

As is usual for the marginal approach with survey data, estimation is performed by optimizing a weighted pseudo-likelihood function. For purposes of estimation, this function treats the survey weights as though they were frequency counts and treats observations as independent instead of clustered. Robust standard errors, using the sandwich (Taylor linearization) method, then take the weights and clustering into account for purposes of tests and confidence intervals.

Because the %WeightedTVEM macro uses only working-independence GEEs, it is not affected by the severe bias that can be caused by missing data on time-varying covariates in GEEs, as described in Pepe & Anderson (2004). However, even with working independence, GEE approaches such as those used here still theoretically require the missingness completely at random (MCAR) assumption when handling missing data, unlike multilevel approaches, which require only missingness at random (MAR). Although this could be problematic, it may be the case that the effect of the violations of this assumption can be reduced by using weights, which are recalculated at each wave to adjust for attrition. These are available in Add Health (see Chen & Chantala, 2014).

The pseudolikelihood function also allows us to calculate pseudolikelihood AIC and BIC fit statistics (see Xu, Chen, & Mantel, 2013). These can be used heuristically for model selection (e.g., choosing the number of knots). Instead of using a penalized sample log-likelihood, a pseudolikelihood criterion consists of a penalized weighted sum of individual log-likelihood functions; it is the same thing as a likelihood criterion if all of the weights are 1. Internally to the macro, the weights provided are standardized to have a mean of 1 across the sample (including all domains but not including individuals excluded for having missing data). This standardization is done because the fit criteria would otherwise fail to work properly when weights are very large, even though proper adjustments were made to the estimates and standard errors. The standardization can be prevented if the user sets the optional argument “normalize_weights=no,” but it is recommended that the user maintains the default setting “normalize_weights=yes.”

4 Macro syntax

The syntax for the macro is as follows.

```
%MACRO WeightedTVEM(data ,
    time ,
    dv ,
    tvary_effect ,
    knots ,
    dist ,
    weight ,
    cluster ,
    domain ,
    which ,
    degree = 3 ,
    evenly = 0 ,
    normalize_weights = yes,
    invar_effect = ,
    show_all = no,
    output_prefix = tvem_ ,
    outfilename = ,
    plot = full ,
    plot_scale = 100
);
```

The first six arguments, shown in bold, are required, and all of the others are optional. Most of the arguments have the same function and meaning as in the %TVEM macro version 3.1. The arguments are described in the next four tables.

Table 1
Required Arguments for the %WeightedTVEM Macro

Name	Description
`data`	The name of the input dataset. This dataset should have longitudinal data structure (one row for each assessment).
`time`	The measurement time variable. One and only one must be provided.
`dv`	The dependent variable. One and only one must be provided.
`tvary_effect`	The covariates which are assumed to have time-varying coefficients. At least one must be provided. To include a time-varying intercept, a variable should be created where all values equal 1, and this variable name should be included in this argument.
`knots`	A positive integer (such as 10) for each variable in `tvary_effect.` These numbers indicate the number of knots (a measure of flexibility) to be used in estimating the corresponding coefficient functions for the `tvary_effect` variables.
`dist`	A single word describing the kind of conditional distribution assumed for the dependent variable, and therefore the type of model to be fit. There are currently two choices: • `normal` assumes that the dependent variable is a continuous and normally distributed (i.e., Gaussian) numerical variable and that a linear regression link function should be used. The macro will accept the word `Gaussian` instead of `normal.` • `logistic` assumes that the dependent variable is on a binary (0 or 1) scale and that a logistic regression link function should be used. The macro will accept the word `binary` or `binomial` instead of `logistic.` **Technical note:** The logistic regression model will be for predicting the probability of a 1 (like the descending option in PROC LOGISTIC).

Table 2
Arguments Found in the %TVEM Macro but Not in the %WeightedTVEM Macro

Name	Reason
`Id`	`cluster` is the equivalent argument in %WeightedTVEM.
`random`	**Not available in the %WeightedTVEM macro.** Instead of random effects, a marginal model with adjusted standard errors is fit.
`stderr`	**Not available in the %WeightedTVEM macro**. Only one method is currently available in this macro, namely Taylor linearization (sandwich standard errors).
`method`	**Not available in the %WeightedTVEM macro**. Only unpenalized B-splines are currently available in this macro.

Table 3
Arguments Found in the %WeightedTVEM Macro But Not in the %TVEM Macro

Name	Description
`cluster`	The name of the variable identifying the cluster (at the highest level of clustering, such as school). If subjects are not clustered, then use the subject identification variable here instead. If nothing is specified for `cluster`, then all observations are assumed to be independent. This argument is directly based on the `cluster` statement in PROC SURVEYREG and PROC SURVEYLOGISTIC.
`weight`	The name of the variable identifying the survey weight. If nothing is specified here, then all observations are assumed to have weight one. This argument is directly based on the `weight` statement in PROC SURVEYREG and PROC SURVEYLOGISTIC.
`domain`	The name of the domain variable for restricting analysis to a subset of the data. The values of this variable should be integers representing different categories.
`which`	The level of the domain variable for which the analysis is currently being done. See note on domain specification, below, for more information.
`normalize_weights`	This tells whether the macro should standardize the weights to average to 1 across the usable dataset. It can be reset to "no," but we recommend leaving it at the default value of "yes."
`show_all`	If this option is set to "yes," then the results of the intermediate calculations in PROC SURVEYREG or PROC SURVEYLOGISTIC are displayed; otherwise they are hidden. Most users will not need these intermediate results.

Note on `domain` *specification*: When a domain variable (e.g., one representing biological sex) is specified in SURVEYREG or SURVEYLOGISTIC, the model is automatically fit within each level of the domain variable (e.g., male and female). However, in the %WeightedTVEM macro an additional argument, called `which`, needs to be specified whenever `domain` is used. In the %WeightedTVEM macro, the fit will be reported for only a single level of the `domain` variable, namely the level specified by the `which` argument. In the %WeightedTVEM macro, the domain variable is assumed to be coded to have integer values (e.g. 1 for male and 2 for female), and likewise the `which` argument is assumed to have a single integer value. The reason for introducing the `which` argument is that, in TVEM, unlike ordinary regression or logistic regression, it is necessary to do some tuning (e.g., select a number of knots) for each model. There is no reason why the best tuning (e.g., best number of knots) should be the same for males and for females. Thus, it is convenient to be able to explore various candidate models separately for males and for females. This process would be less confusing if the sexes were considered one at a time, rather than the output for each sex being given concurrently for each possible model option. Also, when standardizing weights, it was not entirely clear whether weights should be standardized for the usable sample as a whole or for each domain separately. For simplicity, the former approach (standardize once for the whole sample) was implemented.

Table 4
Additional Optional Arguments in the %WeightedTVEM Macro

Name	Description	Default
`invar_effect`	The covariates, if any, which are assumed to have a non-time-varying effect. Variables listed here must not be listed in `tvary_effect.`	Note: If omitted, there will be no non-time-varying-effects variables in the model.
`plot`	The technical mechanism used to plot the coefficient functions. Specify `full` to generate polished-looking plots; however, some users' SAS or Java installations may not work with this option. Specify `simple` to generate a less polished-looking plot that is likely to work in more systems. Specify `none` to suppress plotting; this would mainly be used in loops or simulations in which the macro is being called many times.	`full` -Try `simple` if plots are not generated.
`degree`	The degree of the spline used between each knot in the coefficient function. This is another measure of flexibility. The value specified must be 1 for linear, 2 for quadratic, or 3 for cubic.	3 (cubic) -Default recommended unless there is a reason to choose otherwise.
`evenly`	A number (0 or 1) telling how to choose the positions of knots within the time interval. 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 = 0.	`0` (quantiles) -Default recommended unless there is a reason to choose otherwise.
`plot_scale`	The number of points to be plotted in the graphs of the estimated coefficient functions and their confidence bands.	`100`
`output_prefix`	Letters to be prefixed to the names of the output datasets generated by the macro, to indicate which analysis they came from. The user can specify any name of 15 characters or less—may be useful in distinguishing one analysis from another if the macro is called multiple times. Advanced note: For output_prefix (but not outfilename) you can include a SAS library name in the prefix. For example, you can use *mylib.tvem_* instead of *tvem_* if you want to save in a different SAS library. If you don't know what a SAS library is, ignore this note.	If `domain` is not used, the default is "`tvem_.`" If `domain` is used, the default is "`tvem_`*n*`_`" where *n* = the number of domains specified.
`outfilename`	The path and file name for an output file to be generated by the macro. 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 in another application such as R or Microsoft Excel.	If `outfilename` is not specified, no such file is generated. However, SAS datasets will still be generated; they can be exported from SAS in the usual way.

5 Data analysis examples: Longitudinal TVEM in Add Health

Before using the macro, it is necessary to read it into SAS using syntax such as the following:

```
%INCLUDE "C:\Users\Me\Documents\WeightedTvem_v26.sas";
```

Prior to running this example code, the data for this example had been merged from the first three public-use waves of Add Health and variables were recoded so that variables have consistent names across waves; code for doing this merge is available separately. For example, SampleWeight refers to the Core2 variable for Wave1, the Core2_2 variable for Wave 2, and the variable MEXQ1003 for Wave 3. This is the core weight with poststratification. The variable named Cluster holds the value of the original value named Cluster2 at Wave 1, copied across waves. Indices such as depression (Depression) and self-reported closeness to mother (MomCloseness) were also calculated for each wave.

5.1 Predicting depression from closeness to mother

In this example, the outcome variable is a scale measuring the degree of depression and is treated as normally distributed. The code below calls the %WeightedTVEM macro to estimate a TVEM predicting depression level from closeness to mother, assuming that gender has a non-time-varying effect. The macro, in turn, calls the SURVEYREG procedure and post-processes the results.

```
%WeightedTVEM(dist=normal,
    data = Waves123,
    time = AgeCalculated,
    dv = Depression,
    weight = SampleWeight,
    cluster = cluster,
    tvary_effect = Intercept MomCloseness,
    knots = 5 5,
    invar_effect = Male);
```

The following output is produced. It can be interpreted in essentially the same way as output from the %TVEM macro, version 3.1. Note that the AIC and BIC are not calculated from the true likelihood function, because this is considered unknown in GEE. Instead, they are calculated from a pseudolikelihood function, which assumes that the working correlation structure (here, independence) is true. AIC and BIC can still be used in a heuristic way for model fitting.

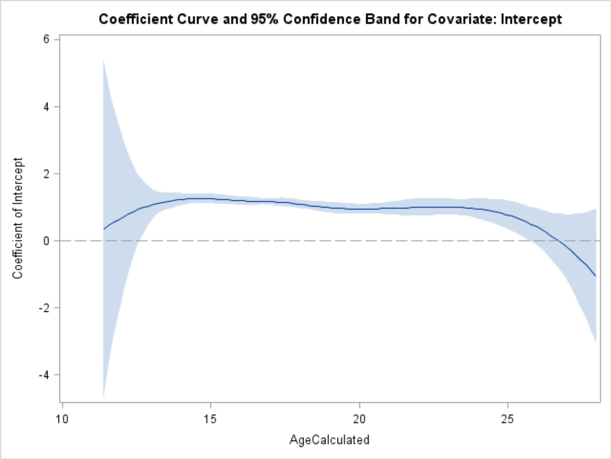
TVEM Macro Output Summary

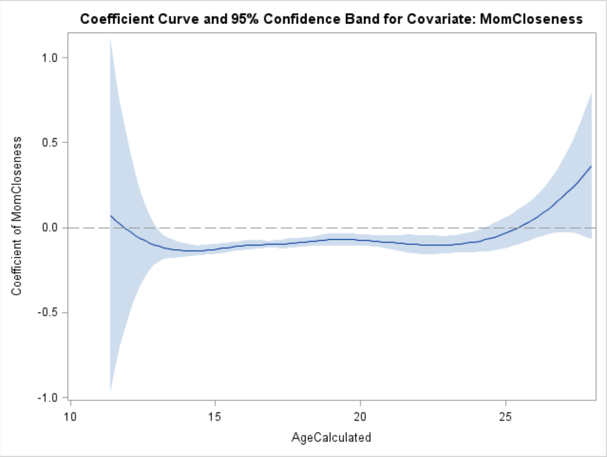
```
================================================================
WEIGHTED Time-Varying Effects Modeling (TVEM) Macro Output
================================================================
Dataset:                    Waves123
Time variable:              AgeCalculated
Response variable:          Depression
Weighting variable:         SampleWeight
Clustering variable:        cluster
Response distribution:      Normal (Gaussian)
Non-time-varying effects:   Male
Time-varying effects:       Intercept MomCloseness
Knots for splines:          5 5
Degree for splines:         3
================================================================
Fit Statistics:

Number of observations used:                  12466
Negative Two Pseudolikelihood:  14705.86
Pseudolikelihood AIC:       14743.86
Pseudolikelihood BIC:       14885.044
================================================================
The estimated coefficient functions are stored in the dataset
tvem_plot_data.
================================================================
                          Fixed Effects Covariates

Obs Parameter Estimate   StdErr     DenDF tValue Probt

1   Male      -0.1196239 0.00974191 131   -12.28 <.0001
```





5.2 Predicting past-month tobacco use from depression

Here, past month tobacco use is a binary variable coded as 0 for no and 1 for yes. The following code can be used to fit a logistic regression.

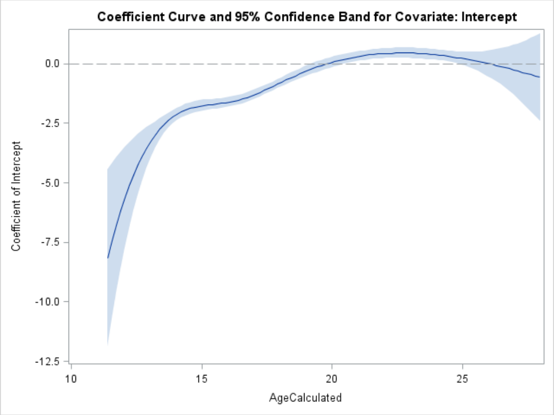
```
     %WeightedTVEM(dist=binary,
           data = Waves123,
           time = AgeCalculated,
           dv = PastMonthTobacco,
           weight = SampleWeight,
           cluster = cluster,
           tvary_effect = Intercept Depression,
           knots = 2 2,
           invar_effect = Male);
```

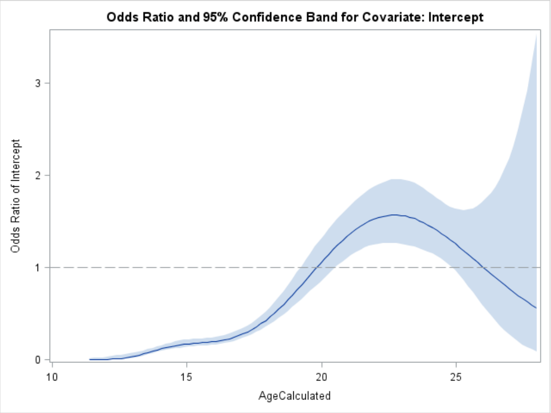
TVEM Macro Output Summary

```
================================================================
WEIGHTED Time-Varying Effects Modeling (TVEM) Macro Output
================================================================
Dataset:                   Waves123
Time variable:             AgeCalculated
Response variable:         PastMonthTobacco
Weighting variable:        SampleWeight
Clustering variable:       cluster
Response distribution:     Binary (Logistic)
Non-time-varying effects: Male
Time-varying effects:      Intercept Depression
Knots for splines:         2 2
Degree for splines:        3
================================================================
Fit Statistics:
Number of observations used:   13190
Negative Two Pseudolikelihood:  14972.02
Pseudolikelihood AIC:       14998.02
Pseudolikelihood BIC:       15095.354
================================================================
The estimated coefficient functions are stored in the dataset
tvem_plot_data.
================================================================
```

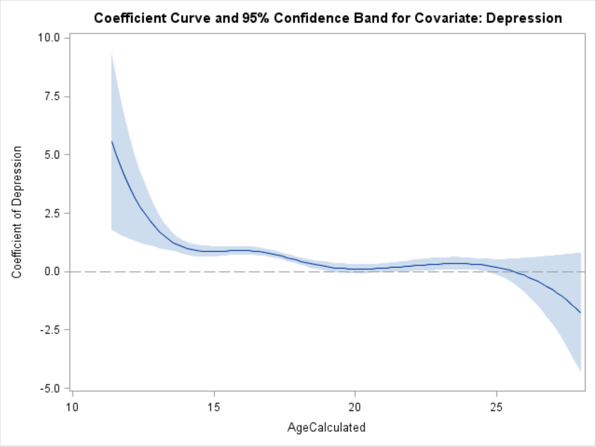
Fixed Effects Covariates

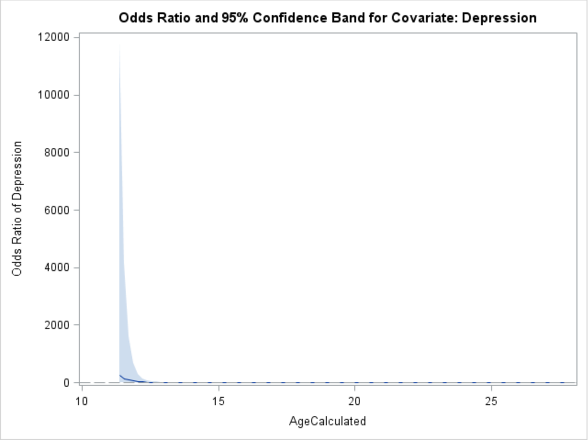
```
Obs Variable DF   Estimate StdErr WaldChiSq ProbChiSq   tValue ProbT
1   Male     131 0.1993   0.0624 10.2031   .001402060 3.19   0.0018
```





The plot above on the right shows that—all else being equal—participants' odds of smoking increased until about age 22, at which point the growth levels off. There may even be a decline in the odds of smoking after age 23 or so, although it is not statistically significant because the pointwise confidence intervals are very wide.





From the plot at left above, it appears that depression is positively related to smoking but only during the teen years, and not significantly related afterwards. The odds ratio plot at right above is not very helpful because the exponentiated confidence intervals are quite wide at the beginning and end of the interval of interest, due to sparse data. Restricting the data to, say, ages 14 to 24, might be useful in a follow-up analysis.

6 References

Chen, P., & Chantala, K. (2014). *Guidelines for analyzing Add Health data (March 2014).* Accessed online at http://www.cpc.unc.edu/projects/addhealth/data/guides/wt-guidelines.pdf

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010) *Applied survey data analysis*. Boca Raton, FL: CRC Press.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R. & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology, 147*, 694-703.

Li, R., Dziak, J. D., Tan, X., Huang, L., Wagner, A. T., & Yang, J. (2015). *TVEM (time-varying effect model) SAS macro users' guide (Version 3.1.0).* University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13-22.

Pepe, M. S. & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation, 23,* 939-951.

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.

TVEM SAS Macro (Version 3.1.0) [Software]. (2015). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Xu, C., Chen, J., & Mantel, H. (2013). Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data*. Survey Methodology, 39,* 303-321.