# Tutorial Example for Windows SAS MixTVEM

In order to use the MixTVEM macro, we have to use the SAS %INCLUDE command to tell SAS to read the macro code:

```
%INCLUDE "C:\Users\myself\Documents\Tvem_Mix_Normal.sas";
```

This reads the code for the macro into SAS. Of course, replace the path "C:\Users\myself\Documents\" with whatever folder (directory) on your computer contains the downloaded macro.

Of course, we also need data. The tutorial example consists of two files. They are simulated (fake) data created to have similar appearance to the (real) Shiffman et al (1996, 1997) dataset as analyzed by Dziak, Li, Tan, Shiffman, and Shiyko (2015), which described the experiences of smokers trying to quit, including their self-rated negative affect and urge to smoke. One of the data files, "MixTvemSampleObservationLevel.txt," has time-varying covariates for each subject listed in a longitudinal ("tall" or "stacked") format, and most of the subjects have more than one line of data.

```
ID    NegAffect   Time  Urge
1     1.4   0.9   3.79
1     1.57  3.72  6.07
1     1.27  4.76  3.35
...
200   3.47  2.97  10
```

This can be read in using PROC IMPORT.

```
PROC IMPORT OUT=SubjectLevelData
            DATAFILE= "
C:\Users\myself\Documents\MixTvemSampleSubjectLevel.txt"
            DBMS=TAB REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

Of course, you should replace "C:\Users\myself\Documents\" with the location of the downloaded data files on your own computer.

John Dziak (jjd264@psu.edu)

The other file, “MixTvemSampleSubjectLevel.txt,” has subject-level, baseline data, arranged with one line per subject.

```
 ID    MeanPrequitNegAffect    BaselineCigarettesPerDay
       MinutesToFirstCigarette RelapseAtOneMonth
 1     1     28    19    0
 2     1     23    13    0
 3     1     32    9     0
 4     2     30    14    1
```

It can be read in in the same way as the other file, although we won’t need it immediately.

```
PROC IMPORT OUT=ObservationLevelData
            DATAFILE= "C:\Users\jjd264\Documents\Sims-
MixTvem\MixTvemSampleObservationLevel.txt"
            DBMS=TAB REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

Some other data preparation will come in handy later: (1) merging the observation-level and subject-level data, (2) creating a column of ones to serve as the regressor for the intercept coefficient, and (3) centering the negative affect predictor as done in the paper. This can be done as follows:

```
DATA Combined;
      MERGE SubjectLevelData ObservationLevelData;
      BY ID;
RUN;

PROC MEANS DATA=Combined;
RUN;

DATA Combined;
      SET Combined;
      CenteredNA = NegAffect - 1.5688092;
      Intercept = 1;
RUN;
```

We can now fit MiXTVEM models. Below is code for an intercept-only model with two classes and six interior knots. Because multiple random starts are requested, it may take a few minutes to run. However, using multiple random starts is worthwhile because it helps us judge how likely it is that we have found the global maximum likelihood. In practice, we recommend at least 20 and preferably 50 starts.

```
%TVEM_Mix_Normal(mydata = Combined,
                      time = Time,
                      dep = urge,
                      id = ID ,
                      tcov = Intercept  ,
                      latent_classes = 2,
                      Knots = 6 ,
                      Num_Starts = 5,
                      Std_Err_Option  = yes );
```

The output is as follows. The macro first uses a BIC-like statistic to choose lambda, the tuning parameter for the smoothing penalty function, that is, a weight that determines how much to penalize roughness in a potential solution.

**RoughnessPenalty**

Using lambda = 96.974368

Then the macro completes the estimation, and summarizes the main results. First, fit statistics are provided.

MixTVEM Macro

```
Time variable name:            Time
Response variable name:        urge
Number of subjects:                  200
Total number of observations:       4963
Varying-Coefficient Covariates:      Intercept
Effect of time between knots treated as: cubic
Log-likelihood:                       -9746.222
Residual squared error (RSS) weighted by posterior probability:  15697.219
GCV:                                  3.1772966
AIC:                                  19515.038
BIC:                                  19552.299
Estimated proportion nugget:  0.687608
Estimated autocorrelation parameter rho: 0.4343808
Proportion of starting values giving approximately the best obtained log-
likelihood:
      1
Count number of parameters:            25
Smoothed number of parameters:  11.296727
```

The next results shown are the estimated proportion of the whole population comprised by each class, and the estimated marginal error standard deviation within each class.

The SAS System
Class Proportions

| Class1 | Class2 |
|---|---|
| 0.48424 | 0.51576 |

The SAS System
Standard Deviations

| Sigma_Class1 | Sigma_Class2 |
|---|---|
| 1.77184 | 1.78406 |

Finally, logistic regression output is shown. Because no subject-level covariates are specified using the scov command in this example, the logistic regression is intercept-only and not very interesting. The p-value here tests the hypothesis that the size of class 2 is the same as the size of the reference class 1. The null hypothesis is not rejected. This does not, of course, prove that the classes are of equal size; it only means that there is no conclusive evidence to argue that either one is larger than the other.

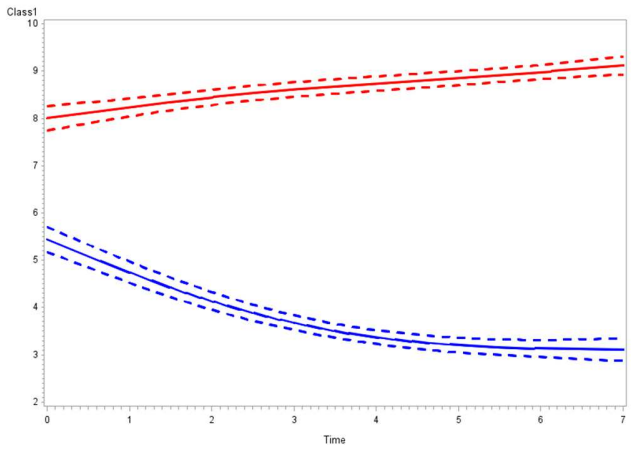
The decision to use six knots and two classes seems arbitrary. However, in practice you could use AIC or BIC, provided in the output, to try many possible choices and select the best.

The SAS System
Logistic Regression for Class Membership

| Class | SCOVNAME | Gamma | SE_Gamma | z | p |
|---|---|---|---|---|---|
| 1 | Intercept | 0.000000 | 0.00000 | . | . |
| 2 | Intercept | 0.063052 | 0.12839 | 0.49109 | 0.62336 |

We still do not have a picture of how the classes differ. We can get a plot of estimated coefficients with estimated pointwise 95% confidence intervals by graphing the contents of one of the automatically produced datasets.

```
SYMBOL1 COLOR="BLUE" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL2 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL3 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL4 COLOR="RED" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL5 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL6 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
PROC GPLOT DATA=MixTVEMGridFittedBeta1;
    PLOT Class1*Time Upper_Class1*Time Lower_Class1*Time
         Class2*Time Upper_Class2*Time Lower_Class2*Time
         / OVERLAY;
RUN;QUIT;
```




This plot shows a high, rising red curve and a low, falling blue curve. Each curve is surrounded by pointwise confidence bands represented by dotted lines. The upper curve represents class 2 and the lower represents class 1. The dotted lines are the estimated confidence interval limits.

We see that class 1 has a declining urge trajectory, while class 2 has a steady or rising one. The next most complex model we could fit would include a time-varying-effects intercept but a constant-effect predictor variable. The predictor variable itself may or may not be time-varying; the decision here is whether to model its effect as being different at different times. Let us used centered negative affect as a predictor in this way.

```
%TVEM_Mix_Normal(mydata = Combined,
                 time = Time,
                 dep = urge,
                 id = ID ,
                 tcov = Intercept  ,
                 cov = CenteredNA,
                 latent_classes = 2,
                 Knots = 6 ,
```

```
                     Num_Starts = 5,
                     Std_Err_Option  = yes );
SYMBOL1 COLOR="BLUE" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL2 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL3 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL4 COLOR="RED" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL5 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL6 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
PROC GPLOT DATA=MixTVEMGridFittedBeta1;
    PLOT Class1*Time Upper_Class1*Time Lower_Class1*Time
         Class2*Time Upper_Class2*Time Lower_Class2*Time
         / OVERLAY;
RUN;QUIT;
```

We get the following output:

The SAS System

Warning: Penalty_Max may need to be made smaller.

**RoughnessPenalty**

Using lambda = 0.000486

MixTVEM Macro

```
Time variable name:            Time
Response variable name:        urge
Number of subjects:                  200
Total number of observations:       4963
Constant-Coefficient Covariates:     CenteredNA
Varying-Coefficient Covariates:      Intercept
Effect of time between knots treated as: cubic
Log-likelihood:                      -9252.743
Residual squared error (RSS) weighted by posterior probability:   13087.82
GCV:                                  2.666007
AIC:                                 18559.486
BIC:                                  18648.54
Estimated proportion nugget: 0.7006774
Estimated autocorrelation parameter rho: 0.4858673
Proportion of starting values giving approximately the best obtained log-
likelihood:
      1
Count number of parameters:           27
Smoothed number of parameters:        27
```

The SAS System
Constant Regression Effects

| Class | VARNAME | Beta | SE_Beta | z | p |
|---|---|---|---|---|---|
| 1 | CenteredNA | 1.26757 | 0.062071 | 20.4213 | 0 |
| 2 | CenteredNA | 1.06199 | 0.063613 | 16.6945 | 0 |

The SAS System
Class Proportions

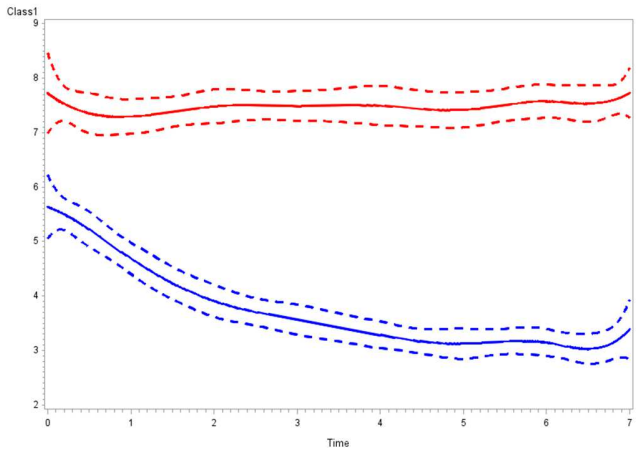
| Class1 | Class2 |
|---|---|
| 0.48904 | 0.51096 |

The SAS System
Standard Deviations

| Sigma_Class1 | Sigma_Class2 |
|---|---|
| 1.61702 | 1.62983 |

The SAS System
Logistic Regression for Class Membership

| Class | SCOVNAME | Gamma | SE_Gamma | z | p |
|---|---|---|---|---|---|
| 1 | Intercept | 0.000000 | 0.00000 | . | . |
| 2 | Intercept | 0.043845 | 0.14271 | 0.30723 | 0.75866 |




This plot shows a high, mostly steady red curve, and a low, declining blue curve.
Each curve has many small, nonsignificant wobbles. The upper curve represents class 2 and the lower represents class 1. The dotted lines are the estimated confidence interval limits.

The warning in the output indicates that the search for a tuning parameter here chose the smallest value in the range considered, which may be a sign that something is wrong. We could try exploring further by inputting different values for the optional macro parameter Penalty_Max. However, we do not worry about this in this illustrative example. The p-values for centered negative affect in each class are extremely small, strongly indicating that negative affect is a good predictor of urge, for members of both classes. Notice that the AIC and BIC values for the model with the predictor are much smaller (better) than for the model without.

We can go on to check whether the regression coefficient for negative affect should be time-varying. This entails putting CenteredNA in as a "tcov" instead of simply a "cov."

```
%TVEM_Mix_Normal(mydata = Combined,
                      time = Time,
                      dep = urge,
                      id = ID ,
                      deg = 3,
                      tcov = Intercept CenteredNA,
                      latent_classes = 2,
                      Knots = 6 6 ,
                      Num_Starts = 5,
                      Std_Err_Option  = yes );
```

The results are as follows.

**RoughnessPenalty**

Using lambda = 153.69402

MixTVEM Macro

```
Time variable name:            Time
Response variable name:        urge
Number of subjects:                 200
Total number of observations:      4963
Varying-Coefficient Covariates:     Intercept CenteredNA
Effect of time between knots treated as: cubic
Log-likelihood:                      -9251.347
Residual squared error (RSS) weighted by posterior probability:   13085.09
GCV:                                 2.6542326
AIC:                                 18535.854
BIC:                                 18590.539
Estimated proportion nugget: 0.6808266
Estimated autocorrelation parameter rho: 0.4377261
Proportion of starting values giving approximately the best obtained log-
likelihood:
      1
Count number of parameters:           45
Smoothed number of parameters:  16.579857
```

The SAS System
Class Proportions

| Class1 | Class2 |
|---|---|
| 0.51071 | 0.48929 |

The SAS System
Standard Deviations

| Sigma_Class1 | Sigma_Class2 |
|---|---|
| 1.62801 | 1.61878 |

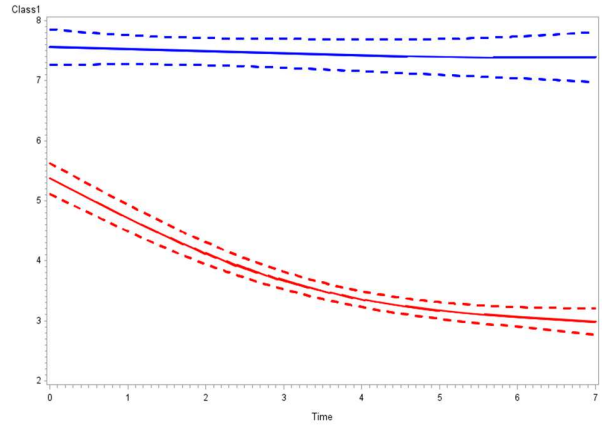
The SAS System
Logistic Regression for Class Membership

| Class | SCOVNAME | Gamma | SE_Gamma | z | p |
|---|---|---|---|---|---|
| 1 | Intercept | 0.000000 | 0.00000 | . | . |

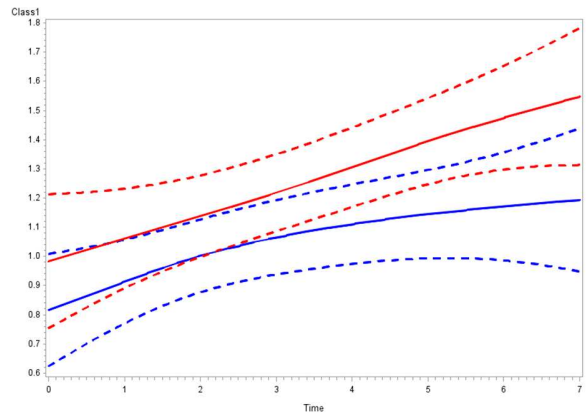
| Class | SCOVNAME | Gamma | SE_Gamma | z | p |
|---|---|---|---|---|---|
| 2 | Intercept | -0.042851 | 0.13744 | -0.31177 | 0.75522 |

The more interesting story is told by the plots. We plot the first regression coefficient (the intercept) and then the second (the time-varying effect of negative affect).

```
SYMBOL1 COLOR="BLUE" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL2 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL3 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL4 COLOR="RED" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL5 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL6 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
PROC GPLOT DATA=MixTVEMGridFittedBeta1;
    PLOT Class1*Time Upper_Class1*Time Lower_Class1*Time
         Class2*Time Upper_Class2*Time Lower_Class2*Time
         / OVERLAY;
RUN;QUIT;
PROC GPLOT DATA=MixTVEMGridFittedBeta2;
    PLOT Class1*Time Upper_Class1*Time Lower_Class1*Time
         Class2*Time Upper_Class2*Time Lower_Class2*Time
         / OVERLAY;
RUN;QUIT;
```




The plot at left represents the intercept function $\beta_{0c}$ (which the macro calls GridFittedBeta1). It shows a high, steady blue curve above a low, declining red curve. The high curve represents class 1 and the low curve represents class 2. The dotted lines are the estimated confidence interval limits.




The plot at left represents the intercept function $\beta_{1c}$ (which the macro calls GridFittedBeta2). It shows two rising curves, with the red curve slightly higher than the blue curve, although their pointwise confidence bands overlap. The higher curve represents class 2 and the lower curve represents class 1. The dotted lines are the estimated confidence interval limits.

Notice that the classes have switched labels relative to the previous model, so that the one with the higher intercept is now labeled as class 1 and the lower as class 2. This switch has no substantive interpretation, because the numeric class labels are arbitrary; the classes are defined by their parameters. Looking at the second plot, we can see that the individuals with a declining intercept function may have a higher coefficient for negative affect. However, the difference is not statistically significant; the pointwise confidence intervals always overlap. For both groups, it appears that negative affect becomes a stronger regression predictor over time.

We can also use subject-level covariates as predictors of what class an individual will belong to, and thus of what the regression coefficients and coefficient functions of the time-varying covariates will be. Let's try to predict class from two measures of prequit (baseline) smoking behavior: average minutes to first cigarette in the morning, and average number of cigarettes smoked per day.

```
%TVEM_Mix_Normal(mydata = Combined,
                      time = Time,
                      dep = urge,
                      id = ID ,
                      tcov = Intercept CenteredNA,
                      latent_classes = 2,
                      scov = BaselineCigarettesPerDay
                      MinutesToFirstCigarette,
                      Knots = 6 6 ,
                      Num_Starts = 5,
                      Std_Err_Option  = yes );
```

The output is similar to the previous analysis, except for the logistic regression part. We see that both covariates have p-values less than .05, especially minutes to first cigarette.

The SAS System
Logistic Regression for Class Membership

| Class | SCOVNAME | Gamma | SE_Gamma | z | p |
|---|---|---|---|---|---|
| 1 | Intercept | 0.00000 | 0.00000 | . | . |
| 1 | BaselineCigarettesPerDay | 0.00000 | 0.00000 | . | . |
| 1 | MinutesToFirstCigarette | 0.00000 | 0.00000 | . | . |
| 2 | Intercept | -3.98530 | 1.33772 | -2.9792 | 0.002890 |
| 2 | BaselineCigarettesPerDay | -0.09273 | 0.04563 | -2.0322 | 0.042129 |
| 2 | MinutesToFirstCigarette | 0.38502 | 0.02211 | 17.4127 | 0.000000 |

Smoking fewer cigarettes, and especially waiting longer in the morning before smoking them, predicts membership in the apparently favorable Class 2.

We could also explore whether the classes have different rates of relapse by including future relapse, measured at 1 month after quit date, as a covariate. We can use the following code:

```
%TVEM_Mix_Normal(mydata = Combined,
                      time = Time,
                      dep = urge,
                      id = ID ,
                      tcov = Intercept CenteredNA,
                      latent_classes = 2,
                              scov = RelapseAtOneMonth,
                      Knots = 6 6 ,
                      Num_Starts = 5,
                      Std_Err_Option  = yes );
SYMBOL1 COLOR="BLUE" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL2 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL3 COLOR="BLUE" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL4 COLOR="RED" INTERPOL=JOIN LINE=1 VALUE=NONE WIDTH=2;
SYMBOL5 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
SYMBOL6 COLOR="RED" INTERPOL=JOIN LINE=2 VALUE=NONE WIDTH=2;
PROC GPLOT DATA=MixTVEMGridFittedBeta1;
```

```
    PLOT Class1*Time Upper_Class1*Time Lower_Class1*Time
         Class2*Time Upper_Class2*Time Lower_Class2*Time
         / OVERLAY;
RUN;QUIT;
PROC GPLOT DATA=MixTVEMGridFittedBeta2;
    PLOT Class1*Time Upper_Class1*Time Lower_Class1*Time
         Class2*Time Upper_Class2*Time Lower_Class2*Time
         / OVERLAY;
RUN;QUIT;
```

The results are as follows (note that the labels switch again):

MixTVEM Macro

```
Time variable name:            Time
Response variable name:        urge
Number of subjects:                  200
Total number of observations:       4963
Varying-Coefficient Covariates:      Intercept CenteredNA
Subject-Level Trajectory Covariates: RelapseAtOneMonth
Effect of time between knots treated as: cubic
Log-likelihood:                       -9250.598
Residual squared error (RSS) weighted by posterior probability:  13083.261
GCV:                                  2.6549353
AIC:                                  18536.357
BIC:                                  18594.342
Estimated proportion nugget: 0.6807547
Estimated autocorrelation parameter rho:  0.437647
Proportion of starting values giving approximately the best obtained log-
likelihood:
      1
Count number of parameters:            46
Smoothed number of parameters:  17.580103
```

The SAS System
Class Proportions

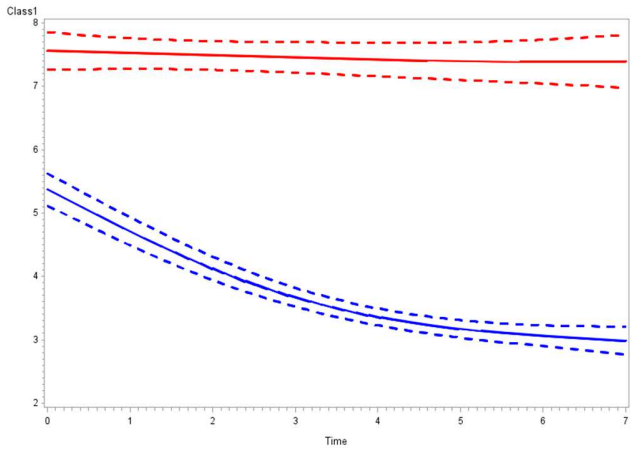
| Class1 | Class2 |
|---|---|
| 0.48942 | 0.51058 |

The SAS System
Standard Deviations

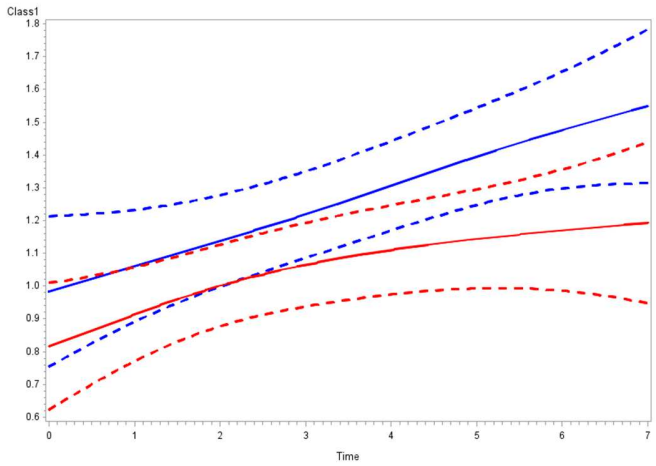
| Sigma_Class1 | Sigma_Class2 |
|---|---|
| 1.61906 | 1.62756 |

The SAS System
Logistic Regression for Class Membership

| Class | SCOVNAME | Gamma | SE_Gamma | z | p |
|---|---|---|---|---|---|
| 1 | Intercept | 0.00000 | 0.00000 | . | . |
| 1 | RelapseAtOneMonth | 0.00000 | 0.00000 | . | . |
| 2 | Intercept | -0.03068 | 0.15308 | -0.20042 | 0.84115 |
| 2 | RelapseAtOneMonth | 0.51254 | 0.20201 | 2.53719 | 0.01117 |




The plot at left, representing $\beta_{0c}$, shows a high, steady red curve and a low, declining blue curve. The high curve represents class 2 and the low curve represents class 1. The dotted lines are the estimated confidence interval limits.




The plot at left, representing $\beta_{1c}$, shows two rising curves. Their pointwise confidence intervals overlap. The higher curve is blue and represents class 1. The lower curve is red and represents class 2. The dotted lines are the estimated confidence interval limits.

From the logistic regression output, there appears to be a significant relationship between relapse and membership in the class 2 (high urge intercept) class. Thus it appears that even after adjusting for negative affect, participants with higher urge to smoke were more likely to relapse. Of course, this is not surprising. However, we get an estimated odds ratio here: relapsers are exp(0.51254) or about 1.67 times as likely to have been in the high class.

## References

Dziak, J. J., Li, R., Tan, X., Shiffman, S., & Shiyko, M. P. (2015). Modeling Intensive Longitudinal Data With Mixtures of Nonparametric Trajectories and Time-Varying Effects. Psychological Methods, in press.

Shiffman, S., Hickcox, M., Paty, J. A., Gnys, M., Kassel, J. D., & Richards, T. (1996). Progression from a smoking lapse to relapse: prediction from abstinence violation effects and nicotine dependence. Journal of Consulting and Clinical Psychology, 64, 993-1002.

Shiffman, S., Engberg, J., Paty, J. A., Perz, W., Gnys, M., & Kassel, J. D., & Hickcox, M. (1997). A day at a time: Predicting smoking lapse from daily urge. Journal of Abnormal Psychology, 106, 104-116.