Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

Markov Tutorial #3: Latent GOLD Longitudinal Analysis of Sparse Data

Overview

The goal of this tutorial is to show how Latent GOLD 5 can be used to estimate latent Markov models with

- sparse longitudinal (panel) data
- unequal time intervals
- time constant and time varying predictors

The Data

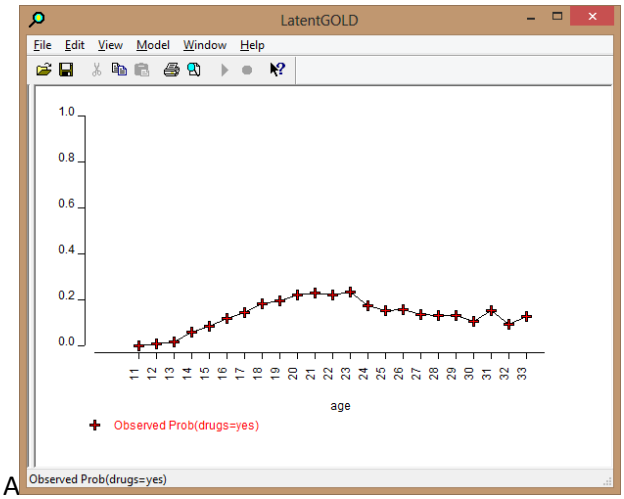
These data are from the 9-wave National Youth Survey (NYS)[1] and were analyzed previously using latent Markov models by Vermunt, Tran and Magidson (2008).

- Sample: N=1725 pupils who were of age 11-17 at the initial measurement occasion (in 1976)
- Survey conducted annually from 1976 to 1980 and at three year intervals after 1980
- To account for the unequal time intervals and to use age as the time scale, models are defined for 23 time points (T+1=23), where t=0 corresponds to age 11 and the last time point to age 33.
- For each subject, data is observed for at most 9 time points (the average is 7.93) which means that responses for the other time points are treated as missing. (See Figure 2 .)
- Dichotomous dependent variable – 'drugs' indicating whether respondent used hard drugs during the past year (1=yes; 0=no).
- Time-varying predictors are 'time' (t) and 'time_2' (t^2); time-constant predictors are 'male' and 'ethn4' (ethnicity).

As shown in Figure **1**A, the overall trend in drug usage during this period is non-linear, with zero usage reported for 11 year olds, increasing to a peak in the early 20s and then declining through age 33. Figure **1**B plots the results from a mixture latent Markov model suggesting that the population consists of 2 distinct segments with different growth rates, Class 2 consisting primarily of non-users.

[1] Elliot, D.S., Huizinga, D., and Menard, S. (1989). *Multiple problem youth: delinquency, substance use and mental health problems.* New York: Springer-Verlag.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

A 

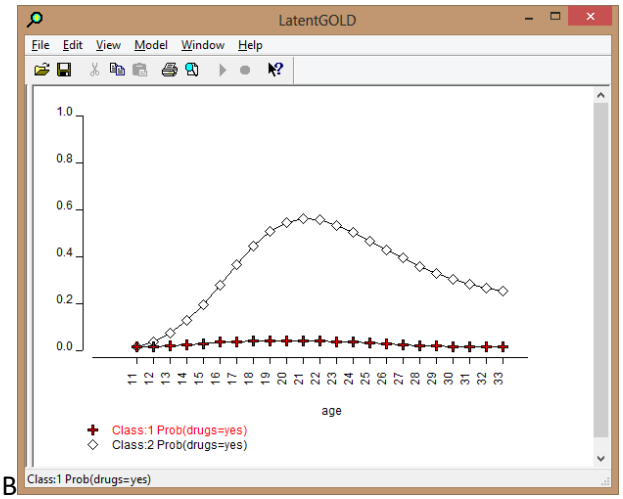
B 

Figure 1(A) Plot of sample proportions of hard drug users by age and (B) corresponding class-specific predicted probabilities for each class obtained from the 2-class mixture latent Markov model.

In this tutorial we will show how to use Latent GOLD 5 to set up the 2-class mixture latent Markov model (as well as a variety of other models) from these data where the variable 'id' is used to identify records associated with each of the 1,725 pupils in the sample.

Figure 2 shows the first 10 records for case id #1, corresponding to ages 11 – 20 (time = 0 - 9). Since this case was age 13 at the time of the initial interview (1976), the value of the dependent variable is blank for the first 2 records corresponding to ages 11 and 12. Responses for the first 5 years of the survey are highlighted, showing first use of hard drugs during year 1982 (time = 6).

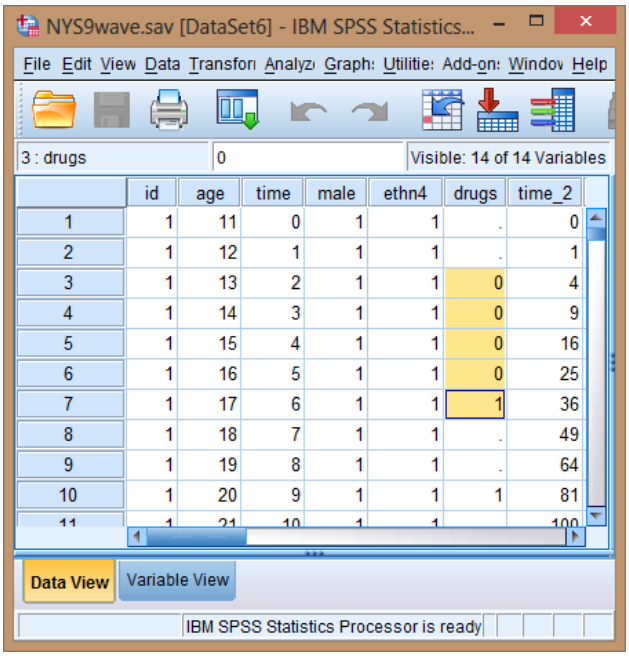


Figure 2. Data file showing the first 10 records for case id = 1**.**

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

Estimating the null model (unrestricted 1-class growth model)

As a starting point we estimate the unrestricted 1-class growth model, which assumes mutual independence between responses at each time point (null model):

- From the File menu, open file 'NYS9wave.sav'
- Right click on 'Model1' and select 'Regression'
- Move 'drugs' to the Dependent box, right click and select 'nominal'.
- Move 'time' to the Predictors box, right click and select 'nominal'.
- Move 'id' to the Case ID box
- In the Output tab, request the additional output 'Estimated Values-Regression'

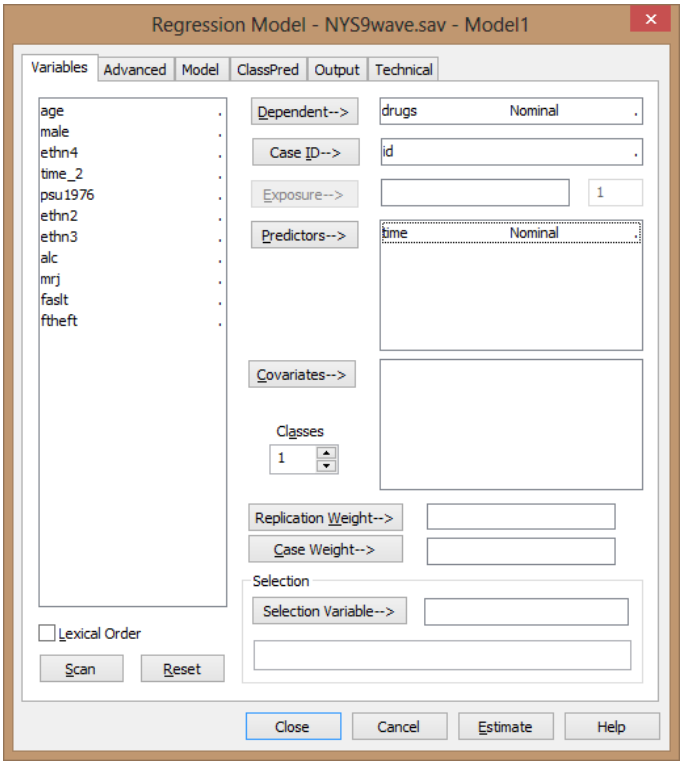


Figure 3. Specifying the null regression model in the Variables tab.

- Click Estimate
- Rename 'model1' to 'null'
- Click on EstimatedValues-Regression to display the predicted rate of drug usage

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

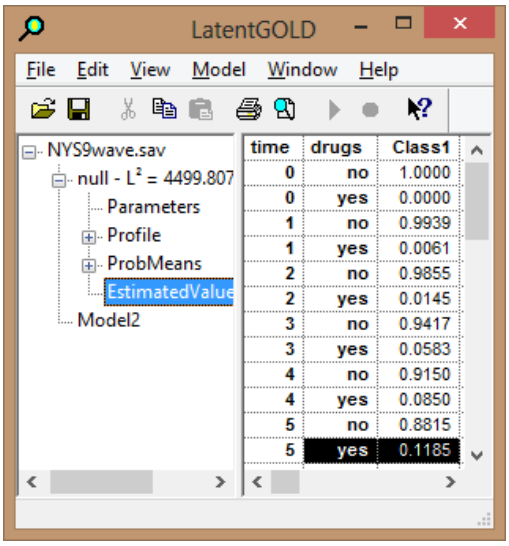


Figure 4. EstimatedValues-Regression output showing predicted rates of hard drug usage.

These predicted rates match the sample proportions for drugs = 'yes' plotted in Figure **1**A. For example, for 16 year olds (time = 5) the predicted percentage of hard drug users is 11.85% (see highlighted row[2] in Figure 4). However, this null model does not fit these data due to a large first order autocorrelation[3] which violates the mutual independence assumption. That is, it is not valid to conclude that all 16 year olds have a probability of .1185 of using hard drugs.

Estimating the 1- and 2-class restricted growth models

Given the shape of the curve in Figure **1**A, following Vermunt et al. (2008), we will assume that the logit(drug = 1) follows a quadratic function of time. To estimate 1- and 2-class logit models containing this quadratic restriction:

- Double click on the 'null ' model to reopen the Variables tab
- Right-click on the predictor 'time' and select 'numeric'
- Move the variable 'time_2' to the Predictors box
- In the Classes box, change '1' to '1-2'

[2] Persons aged 16 have value 'time' = 16 – 11 = 5.

[3]The longitudinal BVRs are Lag1(null) = 2282.1, Lag1(1-class reg) = 1552.2, Lag1(2-class reg) = 239.2, Lag2(null) = 1196.1, Lag2(1-class reg) = 682.1, and Lag2(2-class reg) = 65.8. These L-BVRs are not available as output when estimated from the GUI, but they can be requested as output from the syntax module.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

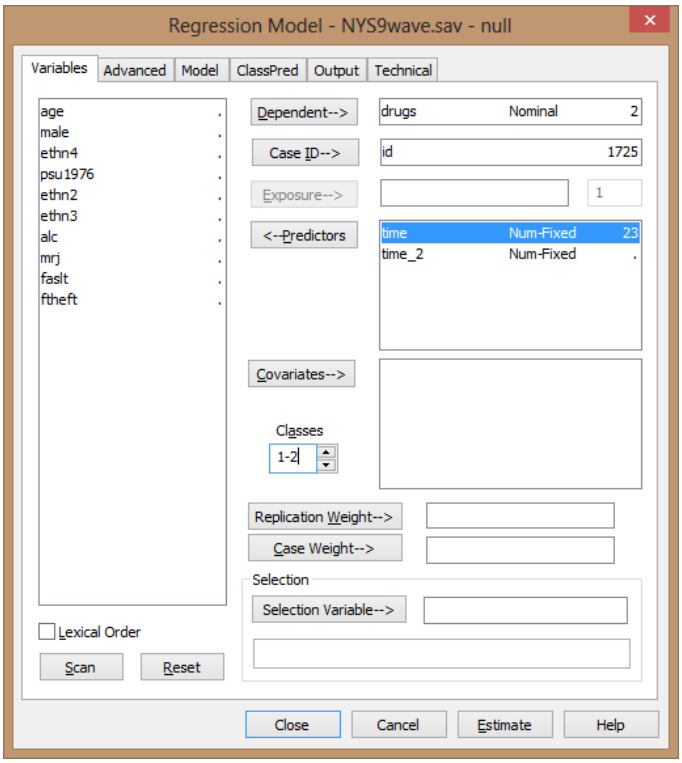


Figure 5. Specifying the latent class growth model in the Variables tab.

- Click Estimate
- Rename the 2 estimated models to '1-class regression' and '2-class regression'
- Click on the name of the data file 'NYS9wave.sav' to display the Model Summary output

Since we are estimating these models on sparse data, chi-square statistics don't apply. Thus,

- Right click in Summary Output pane to open the Model Summary Display, and use it to remove L^2, df, and p-value (see Figure 6).

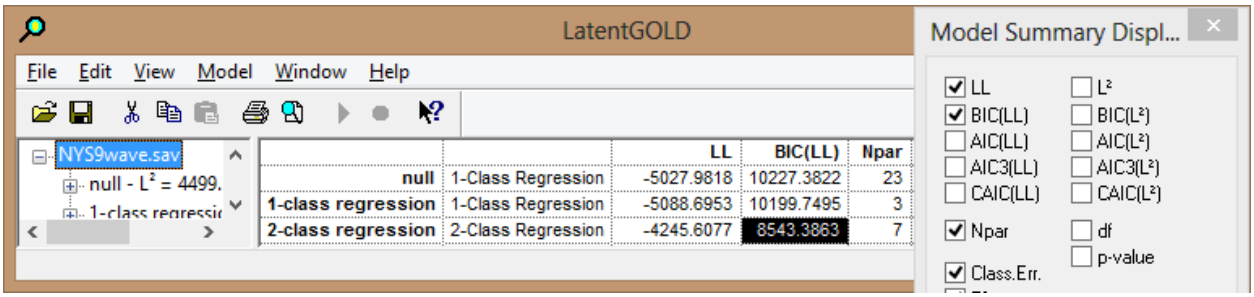


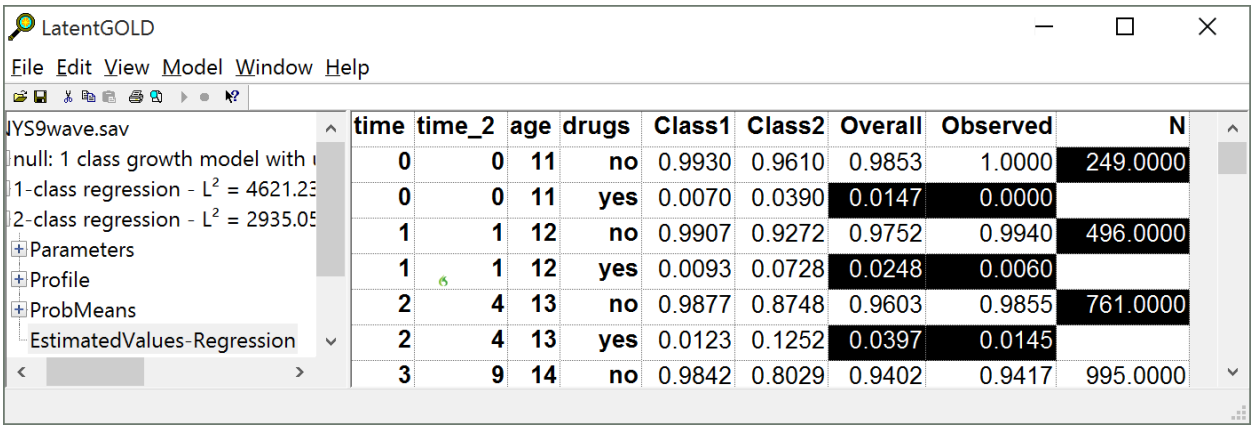
Figure 6. Model Summary output and Model Summary Display.

The BIC shows that among these 3 models the 2-class regression model (2-class restricted growth model) fits best.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

To compare the predicted and observed rates of drug use for this model:

- Click on 'EstimatedValues-Regression' Output



time	time_2	age	drugs	Class1	Class2	Overall	Observed	N
0	0	11	no	0.9930	0.9610	0.9853	1.0000	249.0000
0	0	11	yes	0.0070	0.0390	0.0147	0.0000	
1	1	12	no	0.9907	0.9272	0.9752	0.9940	496.0000
1	1	12	yes	0.0093	0.0728	0.0248	0.0060	
2	4	13	no	0.9877	0.8748	0.9603	0.9855	761.0000
2	4	13	yes	0.0123	0.1252	0.0397	0.0145	
3	9	14	no	0.9842	0.8029	0.9402	0.9417	995.0000

Figure 7. Estimated Values output for 2-class regression model

Of the 249 pupils who were age 11 ('**time'** = 0) at the time of their interview, the model predicts that overall 1.47% would say that they use drugs (0.7% among pupils in class 1 vs. 3.9% for class 2) compared to the actual observed rate of 0%. Of the 496 pupils who were age 12 at the time of the interview, the model predicts 2.48% say they are drug users compared to the observed rate of 0.6%.

Note: to include the variable '**age**' in this output, include '**age**' as one of the model predictors and use the Model tab to set its effects to zero.

We will now show that a comparable latent Markov model provides a better model fit.

Estimating the 2-state latent Markov model

Since the response variable 'drugs' has only 2 categories (1=user, 0=nonuser), we will estimate latent Markov models with 2 latent states, one representing 'true users', the other 'true nonusers'. The word 'latent' in the latent Markov model refers to the fact that the model allows for measurement error[4] in the response variable. Since true users (state 1) might be reluctant to acknowledge their drug use, we would expect more measurement error for responses associated with state 1 than state 2. Later, we will see that the results are consistent with this expectation, which makes us more sure that the latent states are meaningful.

- Right click on the new model name 'Model4' and select 'Markov'

[4] When the number of latent states equals the number of response categories, selecting 'Perfect' from the States box in the Advanced tab causes the measurement errors to be set to zero, reducing the *latent* Markov model to a Markov model.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

- Move ‘drugs’ to the Indicators box, right click and select ‘nominal’
- Request 2 states

Again, following Vermunt et al. (2008), we structure the time effect on the transitions to be quadratic:

- Move ‘time’ and ‘time_2’ to the Covariates box

We will also use ‘age’ to label the time points for the longitudinal plot:

- Move ‘age’ to the covariates box
- Right click on ‘age’ and select ‘Time Label’

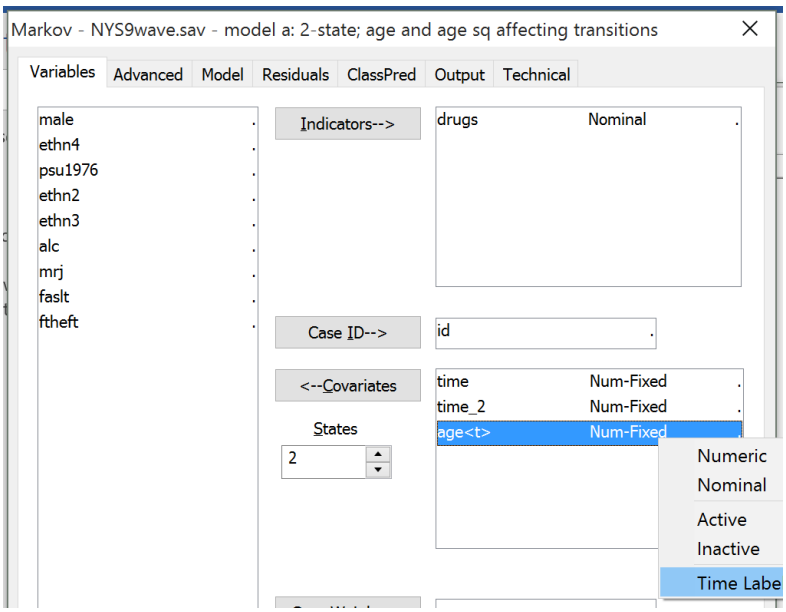


Figure 8. Latent GOLD setup for the 2-state latent Markov model.

- In the Model tab, uncheck the box for age in the Model for Transitions
- In the Output tab, request ‘Classification – Posterior’
- Click Estimate
- Rename the model ‘model a: 2-state; age and age-sq affecting transitions’
- Click on the name of the data file ‘NYS9wave.sav’ to display the Model Summary output

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

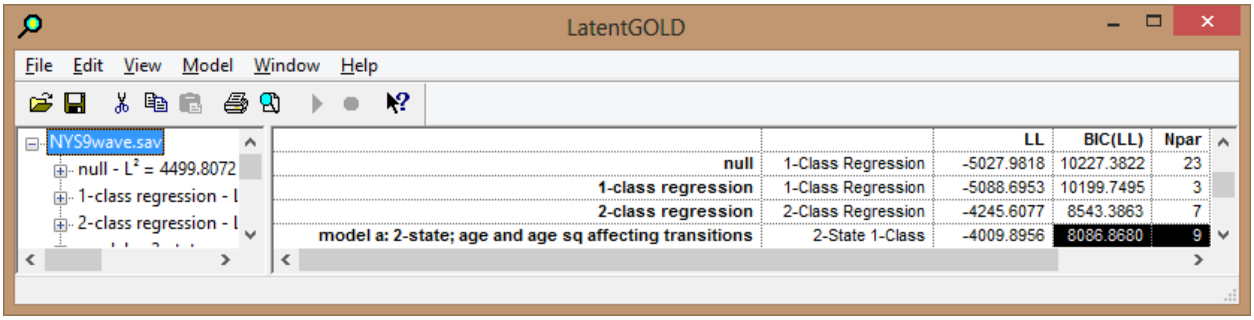


Figure 9. Model Summary output.

The lower BIC for the 2-state LM model confirms that the latent Markov model is a better fit for these data. Also, the longitudinal bivariate residuals (L-BVRs) for Lag1 and Lag2 are both less than 3.84, which means that they are not significantly higher than chance (see Figure 10).

- Click on the Bivariate Residuals output

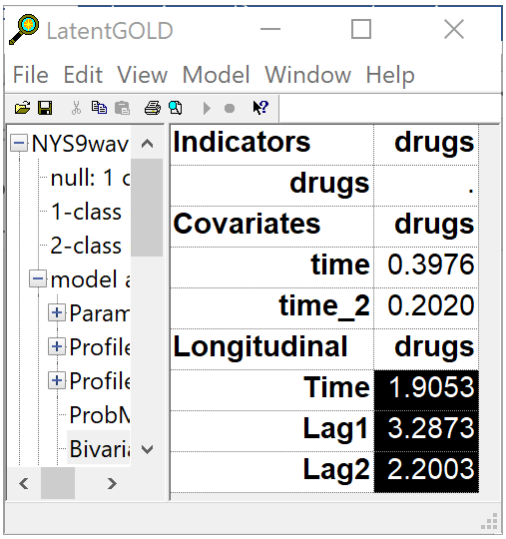


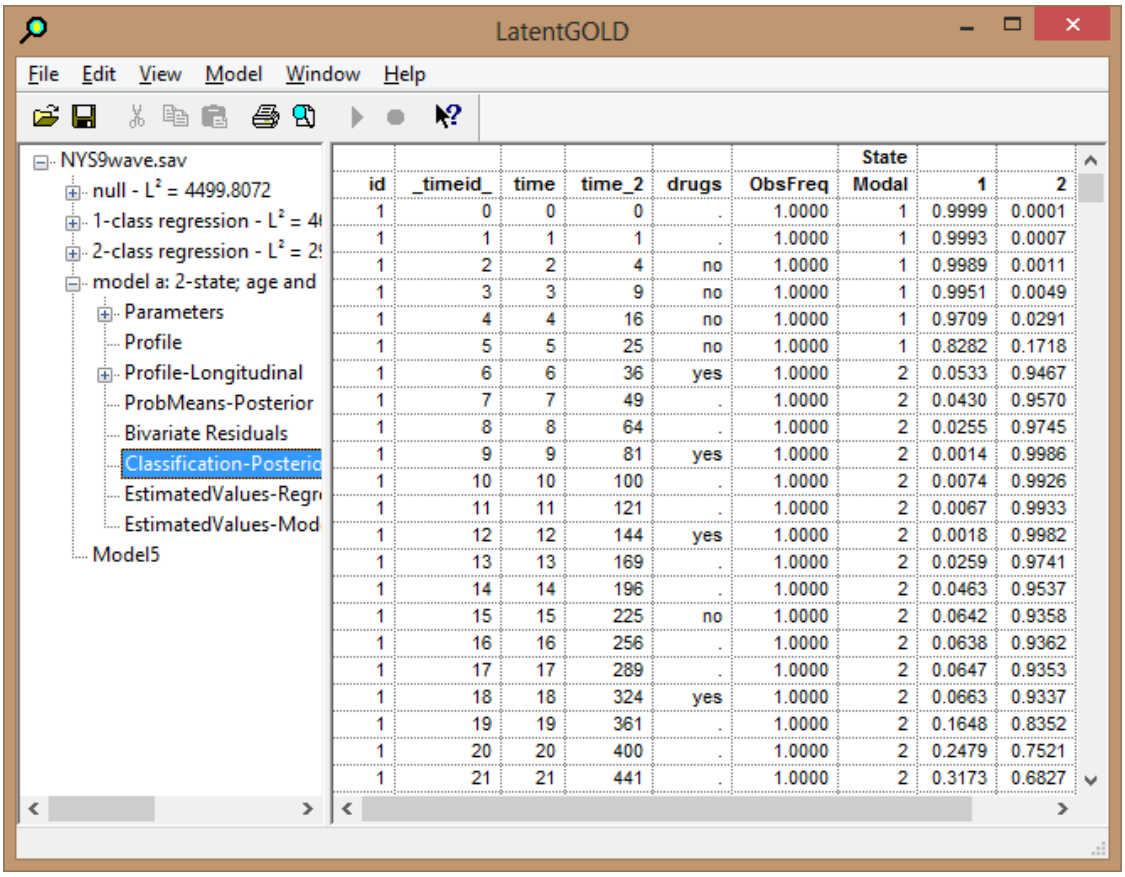
Figure 10. Longitudinal bivariate residuals for the latent Markov model.

Specifically, L-BVR (Lag1) = 3.2873 and L-BVR (Lag2) = 2.2003 corresponding to the first and second order autocorrelations, respectively. These represent a substantial improvement over the L-BVRs associated with the Null model, and the restricted growth models[5].

[5]The longitudinal BVRs are Lag1(null) = 329.4, Lag1(1-class reg) = 336.6, Lag1(2-class reg) = 27.4, Lag2(null) = 201.2, Lag2(1-class reg) = 205.3, and Lag2(2-class reg) = 4.1. These L-BVRs are not available as output when estimated from the Basic program (GUI), but they can be requested from the Syntax module using keyword 'bvrlongitudinal'.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

- Click on the Classification-Posterior output



						State		
id	_timeid_	time	time_2	drugs	ObsFreq	Modal	1	2
1	0	0	0	.	1.0000	1	0.9999	0.0001
1	1	1	1	.	1.0000	1	0.9993	0.0007
1	2	2	4	no	1.0000	1	0.9989	0.0011
1	3	3	9	no	1.0000	1	0.9951	0.0049
1	4	4	16	no	1.0000	1	0.9709	0.0291
1	5	5	25	no	1.0000	1	0.8282	0.1718
1	6	6	36	yes	1.0000	2	0.0533	0.9467
1	7	7	49	.	1.0000	2	0.0430	0.9570
1	8	8	64	.	1.0000	2	0.0255	0.9745
1	9	9	81	yes	1.0000	2	0.0014	0.9986
1	10	10	100	.	1.0000	2	0.0074	0.9926
1	11	11	121	.	1.0000	2	0.0067	0.9933
1	12	12	144	yes	1.0000	2	0.0018	0.9982
1	13	13	169	.	1.0000	2	0.0259	0.9741
1	14	14	196	.	1.0000	2	0.0463	0.9537
1	15	15	225	no	1.0000	2	0.0642	0.9358
1	16	16	256	.	1.0000	2	0.0638	0.9362
1	17	17	289	.	1.0000	2	0.0647	0.9353
1	18	18	324	yes	1.0000	2	0.0663	0.9337
1	19	19	361	.	1.0000	2	0.1648	0.8352
1	20	20	400	.	1.0000	2	0.2479	0.7521
1	21	21	441	.	1.0000	2	0.3173	0.6827

Figure 11. Classification-Posterior output for id=1.

For example, Figure 11 shows that case id = 1 is predicted to be in the non-user state (modal state = 1) until age 17 (**time** = 6 in Figure 11) at which time they switch to the user state (modal state = 2) and then remains in that state.

To compare graphically the predicted and observed rates of drug use for this model, display the longitudinal plot:

- Click on ' Longitudinal-Plot' for this model
- Right click in the plot to open the Interactive Plot Control
- From the plot control uncheck the boxes for State: 1 and State: 2
- Check the boxes for Overall and Observed
- At the bottom of the plot select Overall Prob to highlight the model predictions in red

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

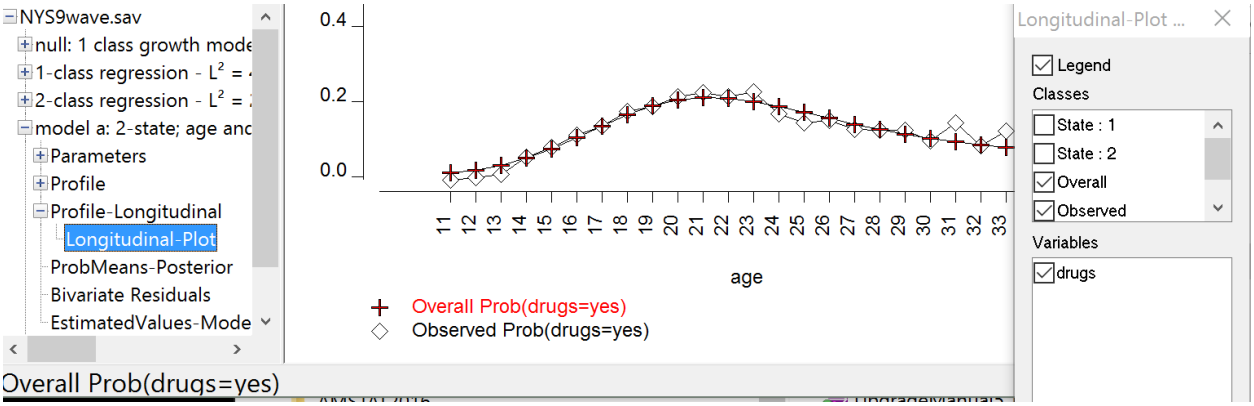


Figure 12. Longitudinal plot for two state latent Markov model

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

Estimating a 2-class mixture latent Markov Model

Next, we will test whether a 2-class model reduces the L-BVRs to non-significant levels. We will allow the 2 classes differ with respect to 1) their initial state probabilities and 2) their transition probabilities.

- Double click on 'model a: 2-state; age and age-sq affecting transitions'
- In the Advanced tab, request 2 classes

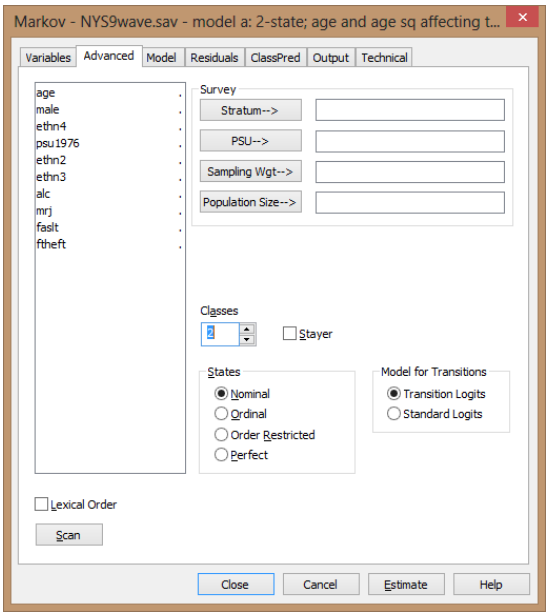


Figure 13. Requesting 2 classes in the Advanced tab.

- Click Estimate
- Rename the model 'model b: model a + 2 class'
- Click on the name of the data file 'NYS9wave.sav' to display the Model Summary output

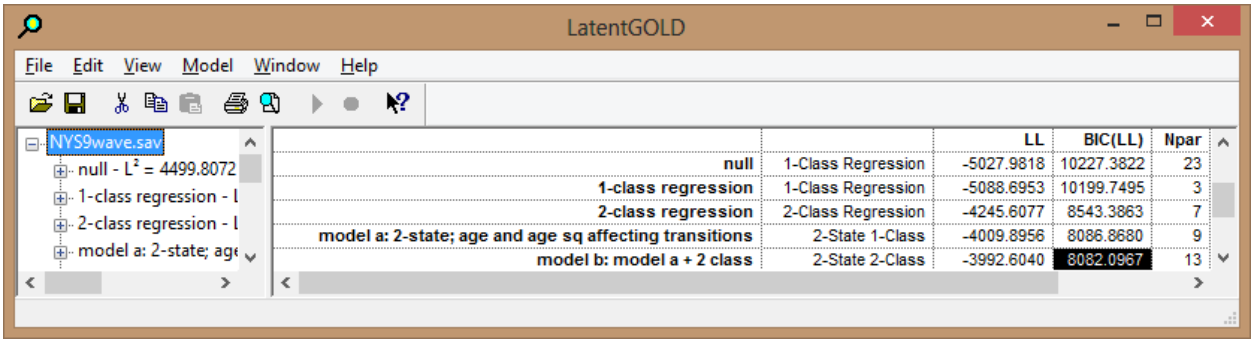


Figure 14. Model Summary output.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

Using 2 classes, we see some improvement in the BIC (Figure 14). Additionally, each L-BVR < 3.84, the critical value for chi-square with 1 degree of freedom, indicating a good model fit – L-BVR (Lag1) = .06 and L-BVR(Lag2) = .005 (see Figure 15).

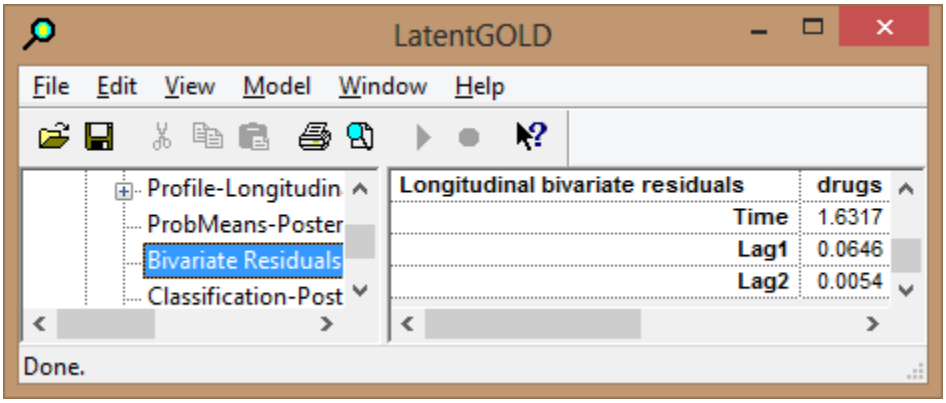


Figure 15. L-BVRs for 2-class mixture latent Markov model.

As shown in the EstimatedValues-Model output (Figure 16):

Class size: 65.86% of the cases are in Class 1

Initial state: Over 99% of class1 and class 2 pupils begin in the non-user state (State 1)

Transition probabilities: Users (State 2) are more likely to change to become non-users in the next year if they are in class 1 than class 2 (prob = .0496 compared to .0012).

Measurement Error: Pupils are less likely to admit that they are users (86.21%) than non-users (98.86%), consistent with the two states being meaningful – i.e, consistent with our expectations.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

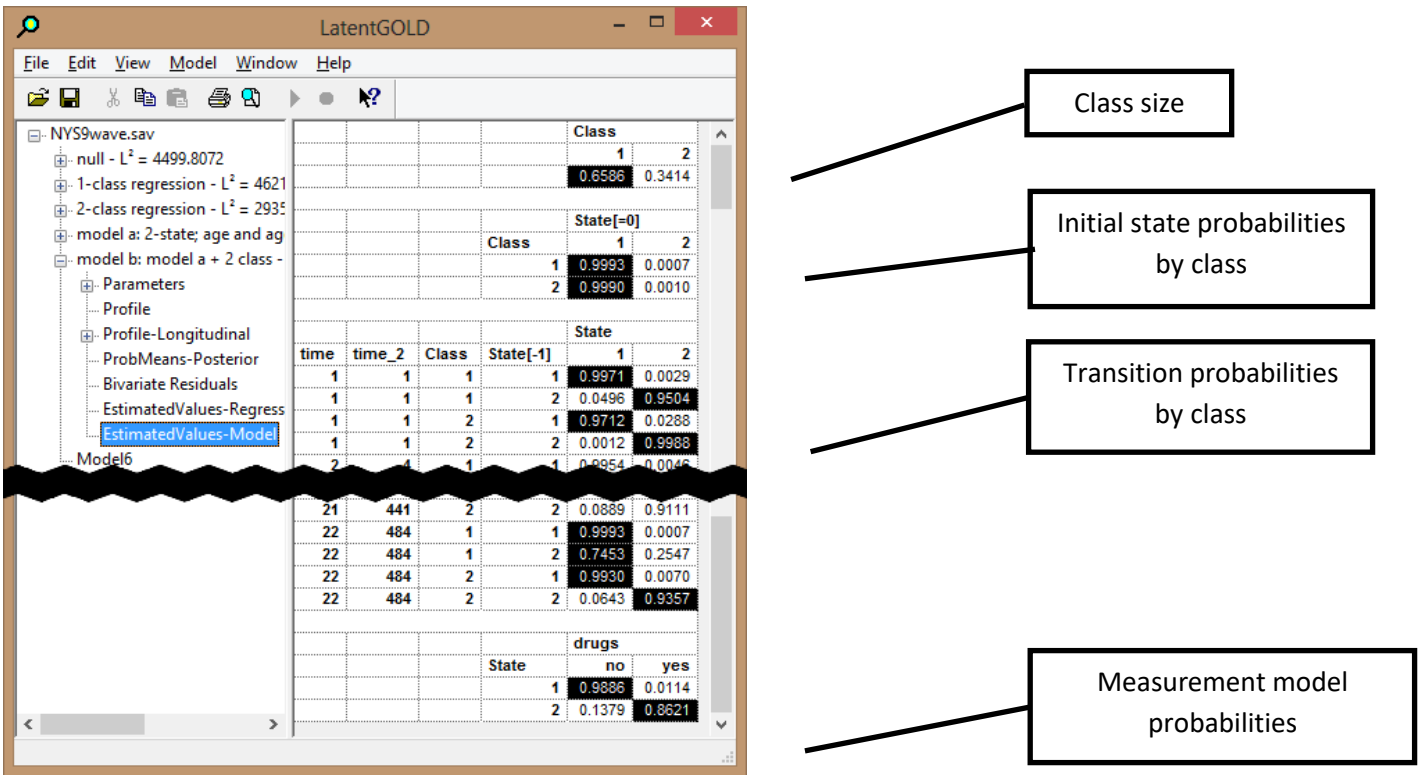


Figure 16. EstimatedValues-Model output for model b.

Including gender and ethnicity as covariates in model

Now we will investigate whether Classes 1 and 2 can be distinguished based on gender and/or ethnicity.

- Double click on 'model b: model a + 2-class'
- Move 'male' and 'ethn4' to the Covariates box, right click and make them 'Nominal'

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

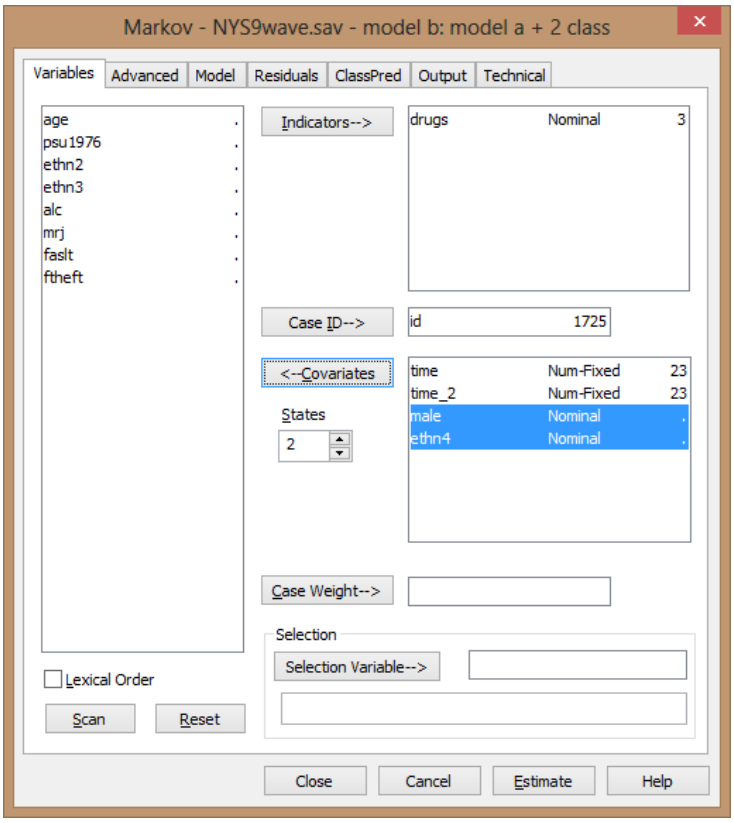


Figure 17. Adding gender and ethnicity as covariates in the Latent Markov Variables tab.

- Click on the 'Model' tab
- Uncheck boxes for 'male' and 'ethn4' to remove these effects from the model for Initial States
- Uncheck boxes for 'male' and 'ethn4' to remove these effects from the model for Transitions

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

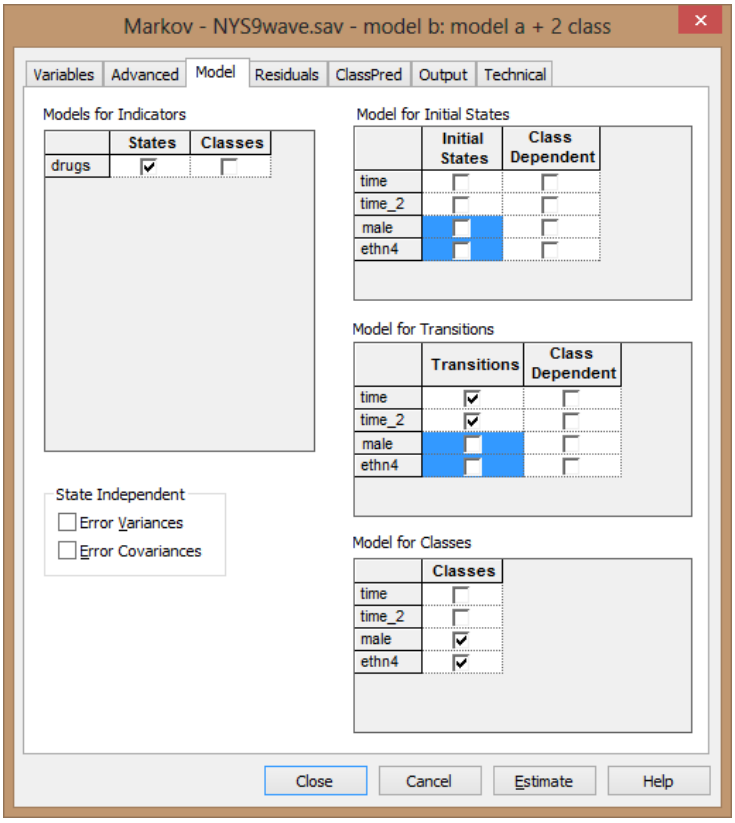
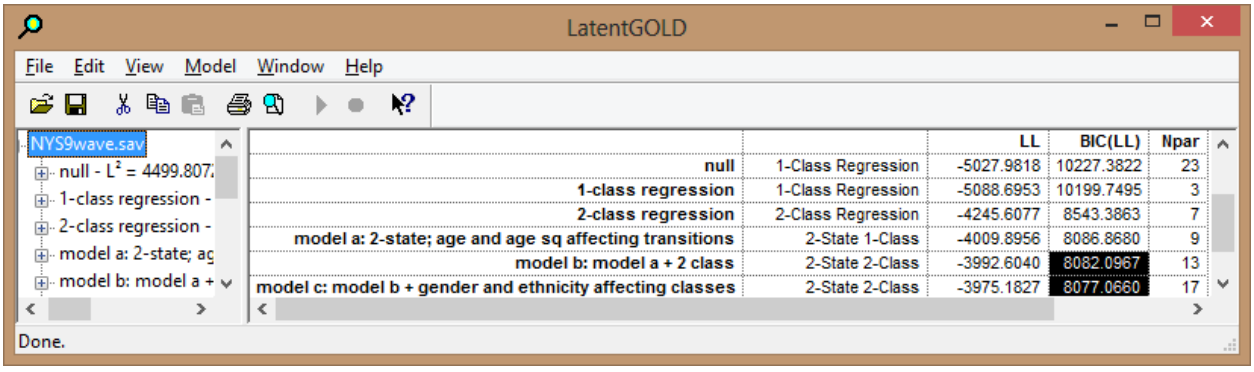


Figure 18. Model tab including male and ethn4 covariate effects (in the model for Classes).

- Click Estimate
- Rename the model 'model c: model b + gender and ethnicity affecting classes'
- Click on the name of the data file 'NYS9wave.sav' to display the Model Summary output



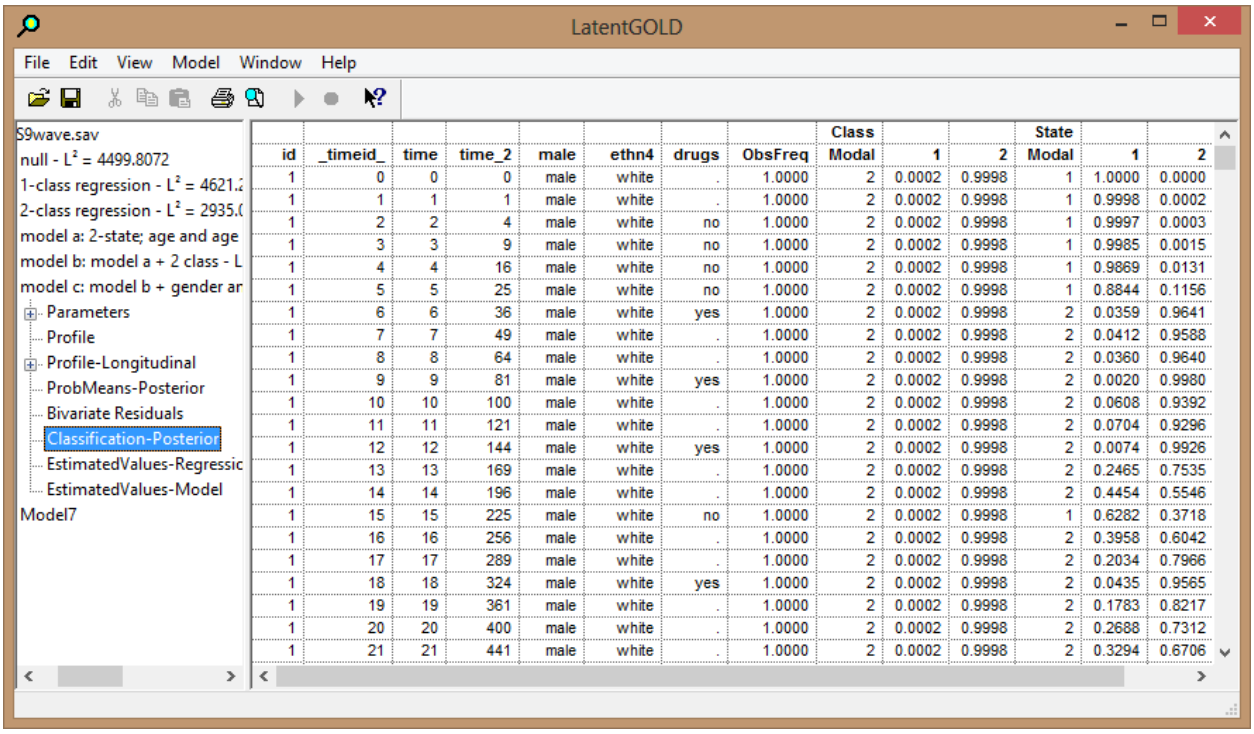
		LL	BIC(LL)	Npar
null	1-Class Regression	-5027.9818	10227.3822	23
1-class regression	1-Class Regression	-5088.6953	10199.7495	3
2-class regression	2-Class Regression	-4245.6077	8543.3863	7
model a: 2-state; age and age sq affecting transitions	2-State 1-Class	-4009.8956	8086.8680	9
model b: model a + 2 class	2-State 2-Class	-3992.6040	8082.0967	13
model c: model b + gender and ethnicity affecting classes	2-State 2-Class	-3975.1827	8077.0660	17

Figure 19. Model Summary output, showing that adding gender and ethnicity improves the fit.

As we can see in Figure 19, model c is preferred according to the BIC.

- For model c, click on the Classification-Posterior output.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

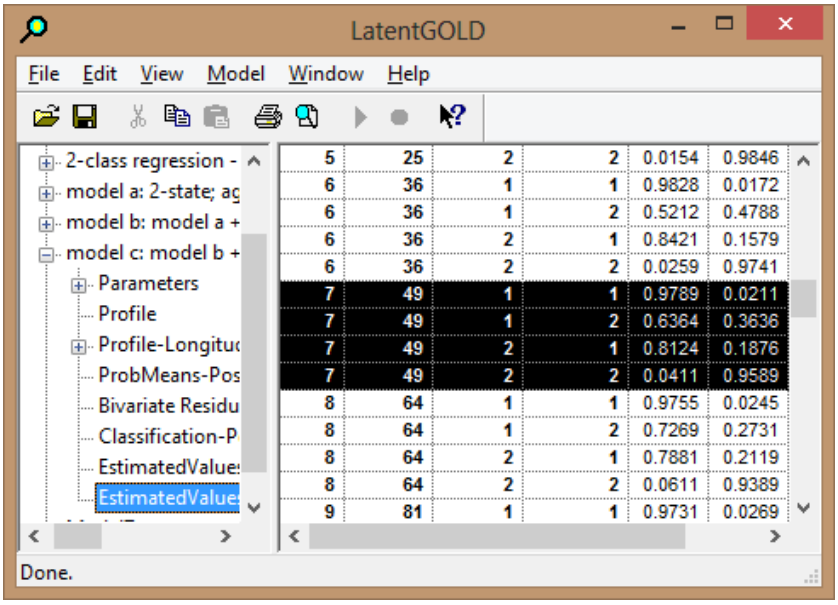


								Class			State		
id	_timeid_	time	time_2	male	ethn4	drugs	ObsFreq	Modal	1	2	Modal	1	2
1	0	0	0	male	white	.	1.0000	2	0.0002	0.9998	1	1.0000	0.0000
1	1	1	1	male	white	.	1.0000	2	0.0002	0.9998	1	0.9998	0.0002
1	2	2	4	male	white	no	1.0000	2	0.0002	0.9998	1	0.9997	0.0003
1	3	3	9	male	white	no	1.0000	2	0.0002	0.9998	1	0.9985	0.0015
1	4	4	16	male	white	no	1.0000	2	0.0002	0.9998	1	0.9869	0.0131
1	5	5	25	male	white	no	1.0000	2	0.0002	0.9998	1	0.8844	0.1156
1	6	6	36	male	white	yes	1.0000	2	0.0002	0.9998	2	0.0359	0.9641
1	7	7	49	male	white	.	1.0000	2	0.0002	0.9998	2	0.0412	0.9588
1	8	8	64	male	white	.	1.0000	2	0.0002	0.9998	2	0.0360	0.9640
1	9	9	81	male	white	yes	1.0000	2	0.0002	0.9998	2	0.0020	0.9980
1	10	10	100	male	white	.	1.0000	2	0.0002	0.9998	2	0.0608	0.9392
1	11	11	121	male	white	.	1.0000	2	0.0002	0.9998	2	0.0704	0.9296
1	12	12	144	male	white	yes	1.0000	2	0.0002	0.9998	2	0.0074	0.9926
1	13	13	169	male	white	.	1.0000	2	0.0002	0.9998	2	0.2465	0.7535
1	14	14	196	male	white	.	1.0000	2	0.0002	0.9998	2	0.4454	0.5546
1	15	15	225	male	white	no	1.0000	2	0.0002	0.9998	1	0.6282	0.3718
1	16	16	256	male	white	.	1.0000	2	0.0002	0.9998	2	0.3958	0.6042
1	17	17	289	male	white	.	1.0000	2	0.0002	0.9998	2	0.2034	0.7966
1	18	18	324	male	white	yes	1.0000	2	0.0002	0.9998	2	0.0435	0.9565
1	19	19	361	male	white	.	1.0000	2	0.0002	0.9998	2	0.1783	0.8217
1	20	20	400	male	white	.	1.0000	2	0.0002	0.9998	2	0.2688	0.7312
1	21	21	441	male	white	.	1.0000	2	0.0002	0.9998	2	0.3294	0.6706

Figure 20. Classification-Posterior output for id=1

Figure 20 shows that case id = 1 is classified into Class 2. To understand how the 2 classes differ in terms of their transition probabilities,

- Click on the EstimatedValues-Model output.



5	25	2	2	0.0154	0.9846
6	36	1	1	0.9828	0.0172
6	36	1	2	0.5212	0.4788
6	36	2	1	0.8421	0.1579
6	36	2	2	0.0259	0.9741
7	49	1	1	0.9789	0.0211
7	49	1	2	0.6364	0.3636
7	49	2	1	0.8124	0.1876
7	49	2	2	0.0411	0.9589
8	64	1	1	0.9755	0.0245
8	64	1	2	0.7269	0.2731
8	64	2	1	0.7881	0.2119
8	64	2	2	0.0611	0.9389
9	81	1	1	0.9731	0.0269

Figure 21. EstimatedValues-Model output, highlighting 18 year olds

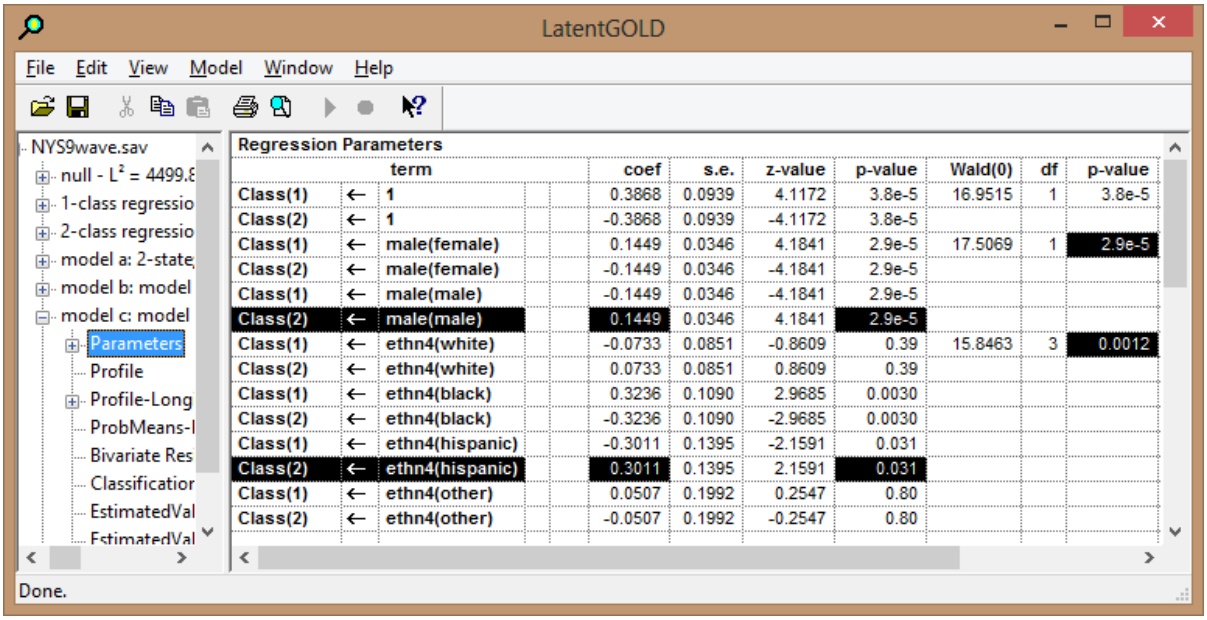
Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

For concreteness we will focus on 18 year olds (see highlighted cells for time = 7 in Figure 21). We see that 18 year olds who were in the lower usage state (State 1) at age 17 have a probability of .1876 of switching to the higher usage state (State 2) if they are in Class 2, compared to a probability of only .0211 of switching if they were in Class 1. In addition, if they were in the higher use state (State 2) at age 17, they have a probability of .9589 of remaining in that state, compared to only .3636 if they were in Class 1. Thus, based on these different transition probabilities we see that Class 2 is more likely to move to and remain in a higher drug usage state than Class 1.

To see how gender and ethnicity relate to the classes,

- Click on the Parameters output.

Note that the p-values for gender and ethnicity are both significant (see right-most p-values highlighted in Figure 22). In particular, males and Hispanics are significantly more likely to be in Class 2.



		term	coef	s.e.	z-value	p-value	Wald(0)	df	p-value
Class(1)	←	1	0.3868	0.0939	4.1172	3.8e-5	16.9515	1	3.8e-5
Class(2)	←	1	-0.3868	0.0939	-4.1172	3.8e-5			
Class(1)	←	male(female)	0.1449	0.0346	4.1841	2.9e-5	17.5069	1	2.9e-5
Class(2)	←	male(female)	-0.1449	0.0346	-4.1841	2.9e-5			
Class(1)	←	male(male)	-0.1449	0.0346	-4.1841	2.9e-5			
Class(2)	←	male(male)	0.1449	0.0346	4.1841	2.9e-5			
Class(1)	←	ethn4(white)	-0.0733	0.0851	-0.8609	0.39	15.8463	3	0.0012
Class(2)	←	ethn4(white)	0.0733	0.0851	0.8609	0.39			
Class(1)	←	ethn4(black)	0.3236	0.1090	2.9685	0.0030			
Class(2)	←	ethn4(black)	-0.3236	0.1090	-2.9685	0.0030			
Class(1)	←	ethn4(hispanic)	-0.3011	0.1395	-2.1591	0.031			
Class(2)	←	ethn4(hispanic)	0.3011	0.1395	2.1591	0.031			
Class(1)	←	ethn4(other)	0.0507	0.1992	0.2547	0.80			
Class(2)	←	ethn4(other)	-0.0507	0.1992	-0.2547	0.80			

Figure 22. Parameters output for model c, showing that age and ethnicity are significant.

To create the plot shown in Figure **1**B with the variable labels for 'age' shown on the x-axis,

- Double click on model c to open the variables tab
- Move 'age' to the Covariates box, right click and select 'Inactive'
- Right click on 'age' again and select 'Time Label'

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

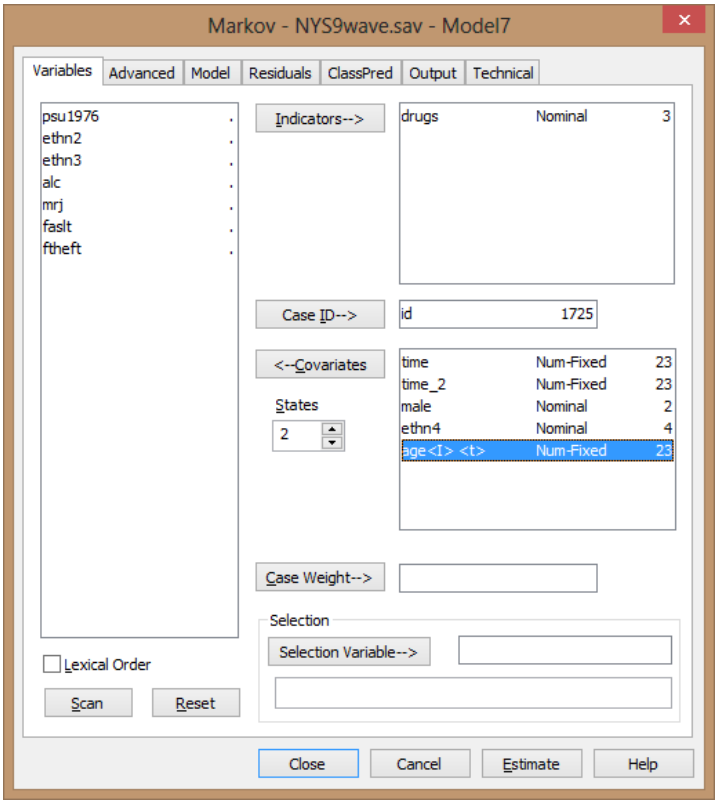


Figure 23. Variables tab showing 'age' specified as a time label.

- Click Estimate
- Click on the "+" next to 'Profile-Longitudinal
- Click on 'Longitudinal-Plot'

To alter this plot to compare the overall predictions from the 2-class model with the corresponding observed proportions,

- Right click on the plot to open the Longitudinal-Plot Display
- Uncheck 'Class : 1' and 'Class : 2' to remove the current plot lines
- Scroll down and check 'Overall' and 'Observed'

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

As shown in Figure 24, the predictions are very close to the observed proportions except for a minor blip at age 31.

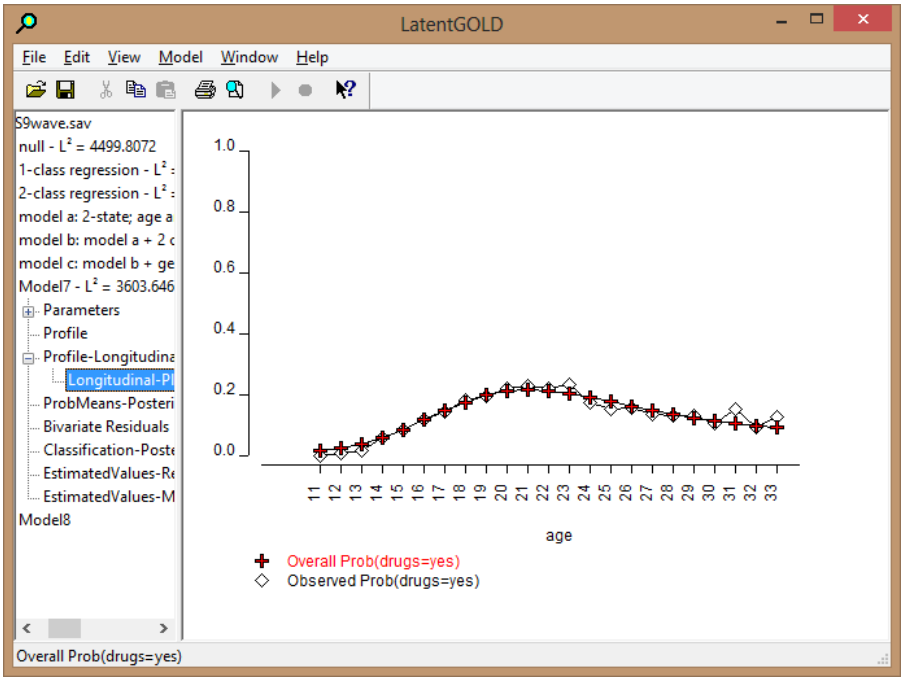


Figure 24. Plot of predicted probabilities vs. observed proportions for 2-class latent Markov model.

Extended Latent Markov Models with Multiple Indicators

Figure 25 shows a similar plot based on a 4-state latent Markov model where the states are ordered and based on 3 drug use indicators -- hard drugs (as plotted above), alcohol, and marijuana.

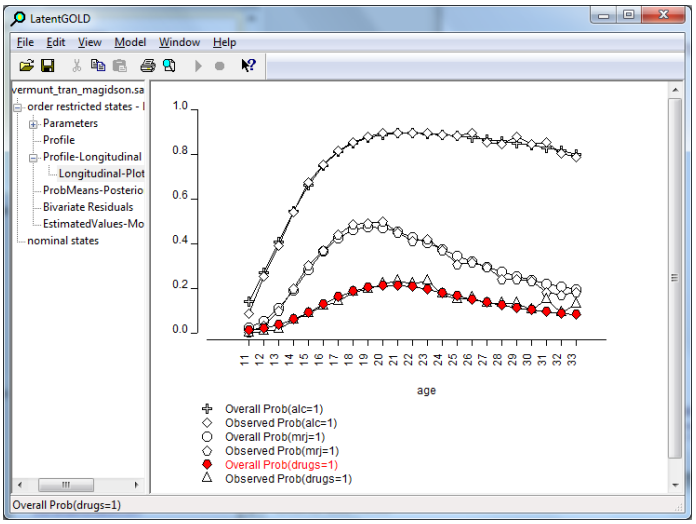


Figure 25. Longitudinal plots for latent Markov model based on 3 indicators – alcohol, marijuana and hard drug usage.

Copyright © 2018 by Statistical Innovations Inc. All rights reserved. (version: 1/31/2018)

The following profile plot shows the overall means (probability of usage) of the three drugs associated with each of the four ordered states.

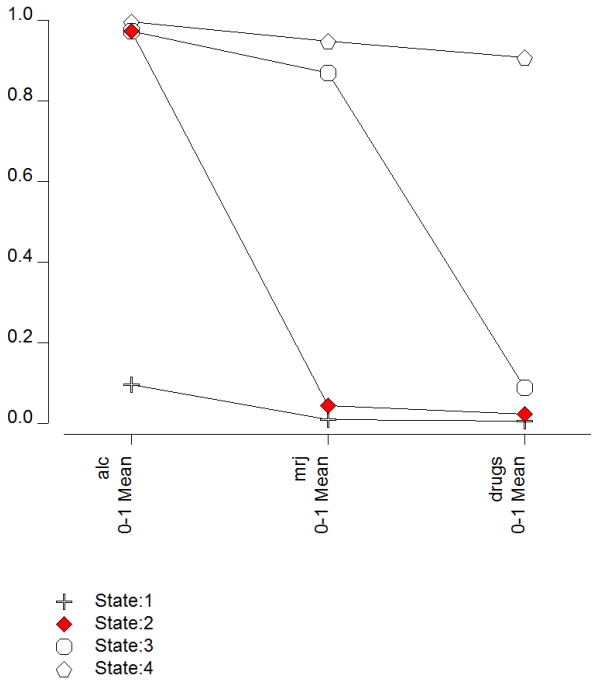


Figure 26. Profile plot for 3-variable 4-state model