Latent Structure Analysis Procedures in SAS®

Deanna Schreiber-Gregory, National University, Moorhead, MN

ABSTRACT

The current study looks at several ways to investigate latent variables in longitudinal surveys and their use in regression models. Three different analyses for latent variable discovery will be briefly reviewed and explored. The latent analysis procedures explored in this paper are PROC LCA, PROC LTA, PROC TRAJ, and PROC CALIS. The latent variables will then be included in separate regression models. The effect of the latent variables on the fit and use of the regression model compared to a similar model using observed data will be briefly reviewed. The data used for this study was obtained via the National Longitudinal Study of Adolescent Health, a study distributed and collected by Add Health. Data was analyzed using SAS® 9.4. This paper is intended for any level of SAS user. This paper is also written to an audience with a background in behavioral science and/or statistics.

INTRODUCTION TO LATENT VARIABLES

When being introduced to statistics, students are usually presented with a group of variables to analyze. One or two variables are identified as being the subjects or dependent variables of the analysis, while the others are identified as the independent variables or contributing factors. A specific statistical model is then ran by following the directions outlined in the particular problem and a result is produced. This process enables students to develop a grasp on how variables in the world interact and affect each other. However, as most of us know, there is much more going on the interactions of objects and people in daily life than can be either simply observed or explained by any single identified variable. Therefore, how can we obtain a grasp on those more subtle, unobserved aspects of cause/effect and interaction that are otherwise too difficult to measure?

One way to look at "unobserved" variables is through latent variable modeling. Through this type of modeling, a statistician is able to view the impact of variables that are not able to be directly observed during the course of a study. Latent variable modeling has a long history and dates back as early as 1904 (SAS, 2014). Latent variables are included in many different kinds of regression models and are more formally referred to as "systematic unmeasured variables;" however, their more widely referred to as factors (SAS, 2014).

Latent variable modeling is quite common and has proved quite useful in the social and behavioral sciences. It has been used for things such as personality assessment, marketing research, disorder pathology, analysis of contributing factors to a particular issue (disorder, epidemic, crisis), and development of treatment and preventative programs. According to the SAS support website, another interesting use that latent variables can serve is the purification of predictors within a regression analysis. The premise behind this view of "purification" is the consideration of the common assumption that linear regression models use predictors that are measured without confounding errors. In other words, a linear regression model is assumed to be measured without the detrimental effect of error inclusion. This can be represented by the following equation:

$$y = \alpha + \beta x + \varepsilon$$

Even considering this assumption, it is important to note that errors can occur. If, for example, *x* had been somehow contaminated and therefore contains measurement errors, then the estimate of β could end up becoming severely biased and inexplicably result in a masking of the true relationship between the *x* and *y* variables. A solution, therefore, needs to be developed in order to address this potential issue. One way to address this problem is through the use of a measurement model for *x.* In this model, we let F_x represent a purified version of the *x* variable. This new model can be included in the equation, which then transforms *x* into something like this:

$$x = F_x + \delta$$

In the above equation δ is present to represent a random measurment error term while F_x represents the measure of *x* without the counfound of containing measurement errors itself. The above definition of *x* can then be entered back into the initial equation to creat the following:

$$y = \alpha + \beta F_x + \varepsilon$$

With F_x (latent variable) being free from measurement errors, δ representing the possibility of measurent errors, and the overall equatin taking these errors into effect, we can safely say that the estimation of β that we were so concerned with earlier can be confidently labeled as unbiased and thus resulting in a model that reflects the true relationship between the variables.

Considering the positive effect on model quality that can result from the inclusion of latent variables, their nature and application certainly warrants further exploration. However, not all latent variables can be calculated the same. Whether the variable is created from categorical data, continuous data, or data represented across time, different latent variable analyses must be taken into consideration. For data that takes on a categorical nature, a latent class analyses would be used to help identify latent class variables with this type of format. For data that it represented in a continuous format, a latent profile analysis would be the appropriate application. And lastly, for data that represents points across time, a latent transition analysis or trajectory analysis are necessary to explore the latent variables that appear over time. Each of these types of latent variables are discussed and explored in this paper.

INTRODUCTION TO THE DATA SET

The Add Health research team was formed in the early 1990s as a direct response to a United States Congress mandate to fund a study that explored adolescent health in America. The Add Health team submitted their proposal for the National Longitudinal Study of Adolescent Health and received the first set of their three program project grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) in time for their 1994 project launch. Funding for this project was also provided from 23 other federal agencies and foundations, all with the desire to support the largest, most comprehensive longitudinal study of health in adolescents that has ever been conducted. Since its creation, Add Health has provided the opportunity for individuals from a wide range of disciplines (public health, human development, biomedical sciences, and other such fields from the social/behavioral sciences and theoretical pools) to explore their research questions and publish their work. Through their work, Add Health has developed the reputation as being a valuable resource for over 10,000 researchers, has assisted in the process of obtaining over 400 independently funded research grants, and has assisted in the publication of over 1,900 research articles for inclusion in a multitude of different disciplinary journals and research outlets.

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-95 school year. The Add Health cohort was then followed into young adulthood in order to gain a more comprehensive picture of how adolescent health affects health and life outcomes of adulthood. This study was conducted through four in-home interviews, the most recent in 2008, when the sample was aged 24-32 (52 respondents were 33-34 years old at the time of the Wave IV interview). Add Health combines longitudinal survey data on respondents' social, economic, psychological and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships, providing unique opportunities to study how social environments and behaviors in adolescence are linked to health and achievement outcomes in young adulthood. As an added resource, the fourth wave of interviews even expanded the collection of biological data in Add Health to understand the social, behavioral, and biological linkages in health trajectories as the Add Health cohort aged through adulthood.

ALTERNATIVE R^2 AND MODEL FIT STATISTICS

In this study, the max-rescaled r-square statistic (adjusted Cox-Snell) provided by SAS® as an option in the model statement of PROC SURVEYLOGISTIC is used as the main reference point in the explanation of how much each final model explains the occurrence of the dependent variable (and is therefore the means to which we evaluate the impact of the latent variables on the explanatory power of the model), however, there is some debate as to whether this is the most appropriately calculated statistic for the job. Paul D. Allison, in his talk on model fit statistics at SAS Global Forum 2014, touched on the possible preferential use of McFadden and Tjur tests. Allison argues that that the Cox-Snell R^2 has appeal as it is able to be naturally extended to regression models other than logistic, such as negative binomial regression and Weibull regression; however, the main limitation of the Cox-Snell test, and thus the reason we are exploring other options of estimating R^2, is its less-than-desirable short upper bound. The Cox Snell R^2 upper bound is less than 1.0. In fact, the upper bound for this test can oftentimes be a lot less than 1.0 depending on *p*, the marginal proportion of cases with events. Allison goes further to provide these examples of this gross deviance: if p=.5, the upper bound reaches a maximum of .75, but if p=.9 (or .1), the upper bound is a mere .48. This is why the max re-scaled R^2 is provided and used in this analysis, as it divides the original Cox-Snell R^2 by its upper bound, thus helping thus helping fix the problem of the "lower" upper bound. However, given that this deviation does exist, it would be beneficial to explore and record other R^2 alternatives when completing an in-depth analysis, however, for the sake of time and consistency, the Cox-Snell R^2 will continue to be reported in the analyses within this paper.

For model fit, the surveylogistic procedure provides three different model fit statistics: Akaike's Information Criterion (AIC), Schwarz Criterion (SC), and the maximized value of the logarithm of the likelihood function multiplied by -2 (-2 Log L). When interpreting these model fit statistics, it is useful to note that lower values of each of these statistics indicates better fit; however, these statistics are open to interpretation and should be considered carefully as they are highly dependent on the structure of the model and sensitive to the number of variables and interactions that are included. These provided statistics are what this paper is primarily using for model fit; however, there are several ways to measure model fit in a logistic regression model. Paul D. Allison, in the same SAS Global Forum 2014

presentation on model fit statistics mentioned above, covers five alternative goodness-of-fit measures for logistic regression models: Hosmer-Lemeshow test, standardized Pearson sum of squared residuals, Stukel's test, and the information matrix test. For the sake of consistency and time, we will continue to look to the model fit statistics provided by SURVEYLOGISTIC; but it is worth noting to explore these other alternatives when conducting a more thorough analysis. For anyone who would like to explore these alternative statistics, a GOFLOGIT macro is available at https://github.com/friendly/SAS-macros/blob/master/goflogit.sas was created as a comprehensive evaluation of the statistics available above (except for Hosmer-Lemeshow, which is simply indicating the lackfit option in model statement after a forward slash); however, Allison warns that a very fundamental problem in the application of a couple of these statistics is included in the model. In short, this macro is available to use if one so wishes but extreme caution should be taken in its interpretation. Please refer to Allison's paper mentioned above for a more in depth explanation as to the limitation of the available macro.

METHODS

Data Cleaning Procedure

Given that the data was collected via a survey format, an overview of data cleaning procedures would be beneficial to understand how the variables were evaluated and controlled within both the latent analyses and final regression analysis. This overview is provided below:

After inputting all the data for each questionnaire into SAS®, the next step was to clean up the data and narrow it down to only the variables that were relevant to this study. Since this study involved multiple questionnaires, the first step in data cleaning was to comb through all the questionnaires and organize the different questions and variety of answers into a reference document. This enabled the author to see exactly when each question appeared, where it appeared, how it was asked, when it was asked (year), if it was asked to participants as adolescents or adults, and the varying answers that could be given. At this time as well, each variable was classified as either categorical or numerical (for use in either a class or profile analysis respectively) and whether it was a question asked across the years and available for longitudinal analyses. Given that the survey was given in paper format, it seemed necessary to use a Microsoft Excel® workbook for this initial step. This way, the different steps in the process of organizing and renaming the data were organized in a linear pattern and could be easily understood, added to, and redone at a later date if necessary. A lot of footwork was necessary to manually read through each survey and pull the appropriate questions. Furthermore, in some cases, it was in the best interest of this research project to combine or distinguish between different answers to a question (for example: creating a binary yes/no answer from a likert scale or creating an ordinal variable from continuous data. In these cases, reorganization of the data was necessary; therefore, for these variables if/then/else statements were used and documented within the SAS code. When this happened, the original and new questions were both provided and the process of transformation was recorded. This way, if an error occurred, it would be easier to backtrack and identify the possible cause for a more efficient fix. Each of these steps (and more) were outlined in the reference document in order to assist with future data use and more efficient troubleshooting.

Controlling for Skewness, Kurtosis, and Missing Values

Other very important aspects of the data exploration phase are the identification and control of skewness, kurtosis, and missing values. Especially when considering these possible aspects of our data set, it is important to attempt to understand and identify where skewness, kurtosis, and missing values exist, why they exist, and possible implications of their existence. Not only would this help us create a more robust model as it would make sure that our data best represents the population we are trying to study, but it would also assist in our understanding of that population and the robustness of the resulting latent analyses.

Descriptive Statistics

To begin the analysis, the researcher used PROC SURVEYFREQ and PROC MEANS to get an idea of the data distribution and other descriptive statistics. Frequencies for demographics, risk behaviors, social attitudes, family life, and mental health variables were all reviewed. When viewing the output generated by these procedures, one must consider the current debate as to the appropriateness of weighting variables included in survey analyses of this size. Some research says that weighting variables is not error proof and can contribute to excluding important factors that would otherwise have shown significant in a nonweighted analysis. This could lead to losing some insight into significant contributing factors and therefore negatively affect the integrity and robustness of the model itself. Other research suggests that weighting the variables helps exclude variables with borderline significance that could muddy the significance and generalizability of the model. The appropriateness of weighting the variables involved in the model was explored using the results and concluded to be an appropriate addition to the analyses based on the fact that the data is from a national sample. An example of the code used is provided below:

```
proc means data=Add_Health;
run;
```

```
proc surveyfreq data=Add_Health;
        strata stratum;
        weight weight;
        by  AlcoholLife1 AlcoholDay2 AlcoholDaySP3 AlcoholBinge4 AlcoholGet5;
run;
```

Categorical Data: Approaching the Difference Between Likert and Binary Scales

Since this study uses character data in both binary and Likert scale formats, it is worth exploration the resulting limitations that arise from these conflicting variable formats. Since binary data contains only two data points and Likert scale data contains more than two points, magnitudes of correlations between these variables shrink due to range restriction. In order to control for this a polychoric correlation matrix was needed. SAS® provides such a matrix in a macro available http://support.sas.com/kb/25/addl/fusion25010_1_polychor.sas.txt. The polychoric correlation matrix from SAS® can be implemented in two steps: (1) by first initializing the macro and computing the polychoric correlation matrix and (2) submitting the computed matrix to PROC FACTOR for factor extraction. An example of the coding is provided below:

```
data Add_Health_FA;
        set    AddHealthWaveI AddHealthWaveII AddHealthWaveIII AddHealthWaveIV;
run;
%polychor(data=Add_Health_FA, var=AlcoholLife1 AlcoholDay2 AlcoholDaySP3
AlcoholBinge4 AlcoholGet5 DrugsMarLife1 DrugsMarDay2 DrugsMarDaySP3
DrugsCocaLife1 DrugsCocaDay2 DrugsInhaLife1 DrugsInhaDay2 DrugsHeroLife1
DrugsMethLife1 DrugsSteroLife1 DrugsInjectLife1 DrugsEcstaLife1 DrugsPrescLife1
DrugsComboLife1 ExerHardActive1 ExerHardActive2 ExerSoftActive1 ExerStrength4
ExerStretch5 ExerTeam6 HealthDoctor1 HealthDentist2 MoodDep1 MoodConsiderS2
MoodPlanS3 MoodAttemptS4 MoodSeriousS5 SexForce1 SexHist2 SexAge3 SexNumLife4
SexNumMonth4 SexSub5 SexProtect7 SexPregnant8 SexSTD9 TobacTry1 TobacDaily2
TobacQuit3 TobacDays4 TobacDaysSP4 TobacAmount5 TobacGet6 TobacChew7
TobacChewSP8 TobacCigar9 VehicleHelmet1 VehicleOtherSB2 VehicleSelfSB2
VehicleSelfSB2 VehicleOtherD3 VehicleSelfD3 ViolMultWeap1 ViolMultWeapSP1
ViolGun1 ViolUnsafe2 ViolThreatSP2 ViolDamageSP2 ViolFight3 ViolFightSP3
ViolInjury3 ViolSigOth4 ViolWhom4 WeightTry1 WeightThink1 WeightDietExer2
WeightFast2 WeightSupp2 WeightPurge2,out=Add_Health, type=corr);
```

METHODS FOR LATENT VARIABLES

Latent Profile Analysis

As a means to explore which latent profile variables could be found in our data set, a factor analysis was performed. The factor analysis was done in order to test the correlations between the different variables and to check for underlying dimensions of related variables (Child, 1990). The variables chosen for each factor analysis were chosen based on their base similarities (such as social constructs, family relationships, health, etc.). A correlation was first performed first in order to weed out any variables that were too closely related (pearson correlations greater than .9). After appropriate pruning was finished, the factor analysis was performed on the resulting group of variables.

```
proc corr data=Add_Health nocorr alpha nomiss;
        var AlcoholLife1 AlcoholDay2 AlcoholDaySP3 AlcoholBinge4 AlcoholGet5;
run;


proc factor data=Add_Health
        method=prinit
        priors=smc
        scree
```

```
        residuals
        rotate=promax
        corr
        heywood;
        var AlcoholLife1 AlcoholDay2 AlcoholDaySP3 AlcoholBinge4 AlcoholGet5;
run;
```

As seen in this sample code, proc factor for the alcohol variable was invoked using `method=prinit`, `priors=smc`, `scree`, `residuals`, `rotate=promax`, `corr`, and `heywood`. The option `method=prinit` requests that an iterated principal factor analysis be used. The option `priors=smc` requests that squared multiple correlations between a given input variable and the other variables in the model be used to estimate the variable's prior communality. The option `scree` requests that a scree plot of the eigenvalues be displayed in the output. The option `corr` requests that both a correlation matrix and partial correlation matrix be displayed in the output. The option `residuals` requests that a residual correlation matrix and associated partial correlation matrix be displayed for the factor analysis in the output as well. The option `rotate=promax` requests that an orthogonal promax rotation be performed on the resulting factors. This was chosen based on the fact that after the initial factor extraction, orthogonal transformation, and varimax transformation, common factors were found to remain uncorrelated with each other and therefore required a promax rotation to ensure that a given variable would only have a high loading on one factor and a near zero loading on other factors. Given the diversity and complexity of the data used, this was a necessary request. Finally, the option `Heywood` requests that any communality greater than 1, be set to 1, allowing iterations to proceed.

Some minor adjustments to the factor analysis for the different variable groups were needed based on an individual basis. For some factor groups, `priors` was set to `max` instead of `smc` based on the need for the prior communality estimate for each of the variables within these groups to be set to its maximum absolute correlation with any other variable (an error appeared in the log of the initial run which identified this need and this was the solution to address it). For other factor groups, `maxiter` was set to `100` or higher based on this variable groups need to limit the maximum number of iterations for factor extraction, as it was exceeded when using the default of 30 .

Latent Class Analysis

The concept of a latent class analysis is used widely in the clinical sciences as it enables researchers to explore the relationship between observed (measured and/or discrete) variables and suggested latent variables that can be derived by the interactions of existing observed variables. There is no procedure within SAS that explores this type of analysis. However, in response to the continuing need for an LCA-specific procedure, the Methodology Center at PennState University set to work on creating a more user friendly procedure to execute both a latent class analysis as well a latent transition analysis (covered next). The PennState LCA – LTA program can be included as an add-on to Base SAS® 9.x and associated SAS® Enterprise Guide. This program is available at http://methodology.psu.edu with instructions as to how to install and run it. The code used ends up looking something like this:

```
proc lca data=YRBS_Total_LCA;
        nclass 2;
        items MoodDep1 MoodConsiderS2 MoodPlanS3;
        categories 2 2 2;
        seed 861551;
run;
```

The PennState LCA-LTA add-on program and subsequent code is what the author used to identify the latent classes used the final logistic regression models. Latent classes that exist within the identified categorical variables of the Add Health data set will be discussed and reviewed during the presentation.

Latent Transition analysis

Another important and interesting form of latent analyses is those applied to longitudinal data. This type of analyses is important as we are able to see unobserved interactions between variables that occur over time. This type of research has helped shed light and explain many important phenomena that have occurred over time and can assist in the identification and prevention of issues that may arise in the future. Latent analyses for longitudinal data can be performed through two different means. One way to explore this type of data is through a latent transition analysis. This type of analysis is provided via the same program talked about earlier that explores latent class analyses. An

example of latent transition analysis code is available below:

```
PROC LTA DATA=Add_Health OUTPOST=Add_Health_Result;
      NSTATUS 5;
      NTIMES 3;
      ITEMS AlcoholLife1 AlcoholDay2 AlcoholDaySP3 AlcoholBinge4 AlcoholGet5;
      CATEGORIES 3 2 3 2;
      GROUPS gender;
      GROUPNAMES male female;
      MEASUREMENT TIMES GROUPS;
      COVARIATES1 AlcoholLife1 AlcoholDay2 AlcoholDaySP3 AlcoholBinge4
      AlcoholGet5;
      REFERENCE1 1;
      SEED 409621;
RUN;
```

Another way to explore the effect of unobserved variables over times is through a trajectory analysis. As with the concepts of the latent class and transition analyses mentioned earlier, a researcher has also developed the methodlogy for a trajectory analysis and has made it available to the public via his site (Jones, 2007). An example of PROC TRAJ code is located below:

```
proc traj data=Add_Health out=Add_Health_Result outstat=healthstat
      outplot=healthplot ci95m;
      var AlcoholDay1 AlcoholBinge4;
      indep d1-d14;
      model zip;
      ngroups 4;
      start -5 -.5 0 0 0 0 0 .5 0 0 70 10 10 10;
      order 0 2 2 2;

      %trajplotnew (healthplot,healthstat, 'Daily Alcohol Use',
      'Alcohol Binge')
run;
```

Both the applications of PROC TRAJ and PROC LTA were applied to the Add Health data.

Structure Equation Models

Another way to approach Latent Structure Analyses is through employment of Structural Equation Modeling techniques and the CALIS procedure in SAS. Structural equation modeling is a type of statistical process that includes the analysis of covariance and mean structures alongside the fitting of groups and systems of linear structural equations, factor analyses, and path analyses. It is interesting to note that in terms of mathematical and statistical techniques, the various types of analyses mentioned above are considered to be interchangeable given their common underlying methodology, however, the inclusion of each of these techniques helps emphasize each of the different aspects of the final analysis. SAS offers PROC CALIS as a specialized procedure specifically designed to approach this type of analysis.

LOGISTIC REGRESSION MODEL

A logistic analysis was conducted in order to test how much of the variability in the dependent variables could be explained by the chosen latent and observed variables. The logistic analysis was written in a manner so that a multiple regression analysis could be performed, given that some of the particular variables used were categorical. Also, given that the variables used are in a complex survey format, PROC SURVEYLOGISTIC was a necessary procedure to employ for this analysis as it accounts for complex survey designs.

```
proc surveylogistic data=Add_Health;
      class MoodConsider2 DemoAge1 DemoSex2 DemoEth3;
```

```
        cluster psu;
        strata stratum;
        model MoodConsider2 = DemoAge1 DemoSex2 DemoEth3 / rsq;
        weight weight;
run;
```

Nesting options were used within PROC SURVEYLOGISTIC throughout this paper in order to account for the combination of the different data sets and to account for the structure used within these data sets. Since the data sets used are for the same region (national) and differ mainly in the years that they were given, the nesting options available remain consistent across the years and do not need to be readjusted. The nesting option `cluster` was used in order to account for survey degrees of freedom. According to the CDC (2014), SAS® considers survey degrees of freedom to be the difference between the number of PSUs and the number of first stage sampling strata among strata and PSUs that contain at least one observation with a value for the variable(s) within the analysis. Considering the possible occurrence in subpopulation analysis (an occurrence that is potentially increased in probability with multiple datasets) that an analytic variable is missing for all survey respondents in one or more PSU or stratum, then an alternate definition of survey degrees of freedom is needed in order to avoid overestimation. This alternative definition has in fact been defined and recommended by Korn and Graubard (1999) and is used by the CDC as the PSU variable. Therefore, in defining `cluster`, the programmer is enabling SAS® to correctly calculate the degrees of freedom without the threat of overestimation. As an added bonus to this option, the log will indicate if there were empty clusters for a variable and how many of the clusters available were included in the analysis. The `strata` statement is also used in order to indicate the name of the stratification variable needed for this study (stratum). Lastly, the `weight` option is used in order to indicate the name of the weight variable to be used in the analysis (weight). All of these nested options assist with ensuring an appropriately coded and rounded data set for analysis.

It is also useful to note that max-rescaled r-square estimates were used in this analysis to approximate model fit. However, there is some thought as to the relevance of using this estimate, which an available option through the SURVEYLOGISTIC procedure in SAS® by indicating `rsq` at the end of a forward slash in model statement, as it is not viewed as the most accurate estimate of model fit. Alternatives to this model were presented in a SAS® Global Forum 2014 paper by Paul D. Allison discussed earlier in this paper.

It is worthy to note that the application of latent structure analyses can be applied to pretty much any type of regression model as long as it does not violate the fundamental assumptions of that model. For the purposes of this paper and to save time, logistic regression application is the chosen regression type covered in this example.

RESULTS AND CONCLUSION

In conclusion, latent variable modeling can have a significant impact on the overall power and significance of any model if applied appropriately to the specific types of data available. Depending on the data available to the researcher, one can conduct a latent variable analysis through use of the PROC LCA, PROC LTA, PROC TRAJ, and PROC CALIS procedures. With these many different procedures available through SAS to approach this very unique and robust addition to regression analyses, there is no excuse to apply it to everyday problems throughout business and social environments.

Please contact the author for review of the specific variables and latent variables used, results specific to the example covered in this paper, and variations of the final model structure.

REFERENCES

About Add Health (2010). Retrieved June 8th, 2014, from http://www.cpc.unc.edu/projects/addhealth/about.

Allison, Paul D. 2012. *Logistic Regression Using SAS® : Theory and Application, Second Edition,* Cary, NC: SAS® Institute Inc.

Allison, Paul D. (2014, March). *Measures of Fit for Logistic Regression.* Paper presented at SAS® Global Forum 2014, Washington, D.C

Center for Disease Control and Prevention (2014). Combining YRBS data across years and sites. From http://www.cdc.gove/yrbs (accessed June, 2014).

Child, D. (1990). The essentials of factor analysis, second edition. London: Cassel Educational Limited.

Christensen, K. B., Nielsen, M. L., and Smith-Hansen, L. (2003). *Latent Covariates in Generalized Linear Models.* Retrieved from: http://publichealth.ku.dk/sections/biostatistics/reports/2003/rr-03-12.pdf.

Field, A., & Miles, J. (2012). *Discovering Statistics Using SAS*® , Thousand Oaks, CA: Sage Publications.

Introduction to SAS® . UCLA: Academic Technolgy Services, Statistical Consulting Group. From http://www.ats.ucla.edu/stat/sas/notes2/ (accessed August, 2012).

Jones, B. L., & Nagin, D. S. (2007). *Advances in group-based trajectory modeling and an SAS procedure for estimating them.* Sociological Methods and Research. 35 (4): 542-571.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). *A SAS procedure based on mixture models for estimating developmental trajectories.* Sociological Methods and Research. 29 (3): 374-393.

Korn, EL and Graubard, BI (1999). Analysis of Health Surveys. John Wiley & Sons, New York. p. 209-211.

Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A., & Collins, L. M. (2013). *PROC LCA & PROC LTA users' guide* (Version 1.3.0). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Latent Variable Models (2014). SAS® Institute Inc. Support. Retrieved June 10th, 2014, from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_intromod_a0000000339.htm.

National Longitudinal Study of Adolescent Health (Add Health), 1994-2008 (ICPSR 21600) [Public Use Data]. (2014). Inter-university Consortium for Political and Social Research (ICPSR): The University of Michigan. Retrieved from http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600?archive=ICPSR&q=21600.

O'Rourke, N., & Hatcher, L. 2013. *A Step-by-Step Approach to Using SAS® for Factor Analysis and Structural Equation Modeling, Second Edition*, Cary, NC: SAS® Institute Inc.

PROC LCA & PROC LTA (Version 1.3.0) [Software]. (2013). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

PROC TRAJ [Software]. (2012). Carnegie Mellon University. Retrieved from http://www.andrew.cmu.edu/user/bjones/index.htm.

SAS® Institute Inc. 2008. SAS /STAT® 9.2 *User's Guide*. Cary, NC: SAS® Institute Inc.

Thompson, D. M. (2006). *Performing Latent Class Analysis Using the CATMOD Procedure*. Paper presented at SUGI 31, San Fransisco, CA.

ACKNOWLEDGMENTS

This author would like to acknowledge the Add Health research team and thank them for their incredible efforts in collecting and compiling such a comprehensive study of adolescent and adult health. This author would also like to thank Add Health team for making their work and data available to the public for research and educational purposes.

CONTACT INFORMATION

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

Deanna Schreiber-Gregory

National University

Department of Community Health

La Jolla, CA

E-mail: d.n.schreibergregory@gmail.com