2012

Using generalized estimating equations to analyze repeated measures binary data from the young adolescent crowd study

Lauren Ashley Beacham
Louisiana State University and Agricultural and Mechanical College

USING GENERALIZED ESTIMATING EQUATIONS TO ANALYZE REPEATED MEASURES BINARY DATA FROM THE YOUNG ADOLESCENT CROWD STUDY

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Applied Statistics

in

The Department of Experimental Statistics

by
Lauren Beacham
B.A., Agnes Scott College, 2009
August 2012

Acknowledgments

I would like to thank my family: Karon and Orlando Martinez, Renee and Steve Phillips, and my brother Christopher for their love and support. To my grandparents Virginia and Bill Cooper who helped make it possible for me to attend college, I would not be here without you. To my mom, Sara, thank you for always encouraging me in all my dreams. To my Dad, who passed away before I came to LSU, I miss you always and I hope you can see me and are proud. I love you all and am so lucky to have such an amazing support system.

I would also like to thank all of my professors at LSU. In particular I want to thank Dr. Brian Marx for agreeing to take me on as his student, letting me use this project as my thesis, and inspiring my interest in modeling analysis while here at LSU. I would also like to thank my committee members: Dr. James Geaghan for all his encouragement and advice, and Dr. Luis Escobar for always pushing me to see how much I can learn and accomplish.

A very special thanks to Aruna Lakshmanan whose special project paper on the YACS study provided the impetus for this thesis as well as a helpful description of the study.

Finally, I want to thank Sylvia and Elaine at the Department of Experimental Statistics for all of their help throughout my graduate career.

Table of Contents

List of Tables

Abstract

The young adolescent crowd study (YACS) was conducted in order to look at the influence of various factors on use of controlled substances by middle school students. The contributing factors investigated were demographics (gender and race), self-esteem in different modalities such as school or athletic performance, and the peer group students belong to. Each student has a binary response for whether they have used alcohol, marijuana or cigarettes which was recorded in both seventh and eighth grade. Since the data has a binary repeated measures response, generalized estimating equations (GEE) in a logistic regression setting is a good way to model the data. The theory and method of GEE is explained in detail followed by results, issues encountered and a discussion of how the model worked with the data set.

Chapter 1. The Young Adult Crowd Study

Substance abuse is a well documented issue with teenagers and use can start as early as middle school (Dolcini and Adler, 1994). Factors that influence substance use is a popular field of study with researchers looking at possible issues such as familial interaction, participation in extracurricular activities, self esteem, and peer pressure and participation in the school social structure (Lakon and Valente, 2012; Jones and Heaven, 1998; Dolcini and Adler, 1994; Smith and Williams, 1993; Selnow and Crano, 1986). To date results of these studies have been conflicting, especially for the factor of crowd participation. For example, Selnow and Crano (1986) and Smith and Williams (1993) found that students who are members of the social structure are less likely to use drugs than students who are outsiders or outcasts. Conversely, Lakon and Valente (2012) found that popular students, or students high up in the social structure were more likely to use drugs than their peers, La Greca, Prinstein, & Fetter (2001) found both outcasts and popular students were more likely to use than smart or normal students, while Jones and Heaven (1998) found no effect of peer groups at all. Despite differences in findings on influences there is a general consensus that finding these factors are important in predicting at risk students. If these at risk students can be identified early on such as in middle school, then intervention methods can be implemented, decreasing the odds of use later on in life.

The Young Adolescent Crowd Study (YACS) is a survey that was conducted from 1993 to 1997 and aims to look at the influence of various factors on use of controlled substances, specifically alcohol, marijuana, and cigarettes, by inner city middle school students. The objective was to look at factors that promoted these behaviors. The contributing factors investigated were demographics (gender and race), self-esteem in different modalities such as school or athletic performance, and the peer group students were identified as being a part of. Students filled out the same questionnaires during both the seventh and eighth grade years, therefore the response is a repeated measure. Because of this structure the data cannot be modeled by simple analysis such as logistic regression. The purpose of the current thesis is to look at Generalized Estimating Equations as an appropriate model than can handle the repeated measures aspect of the data to get an accurate analysis of the factors involved in the odds of students drinking, smoking, and using marijuana.

Chapter 2. Data Collection

2.1 Subjects

The data came from two subsamples of the Young Adolescent Crowd Study (YACS) which was conducted for four years from 1993 to 1997. Each sample is composed of seventh graders who were followed into their eight grade year. Students in YACS sample 3-4 were in seventh grade when YACS 2-3 students were in eighth. All students came from an inner city school in Oakland, California. There were 303 students total, 165 from the YACS 2-3 sample and 138 from the YACS 3-4 sample. Fifty-seven percent of the total sample was female. Forty-three percent of the sample identified as Asian, 37% as Black, and 20% as other. Data was collected by having students fill out two surveys, the self-perception profile and a teen health survey. Students filled out both surveys in the seventh grade year and again in the eighth grade. Crowd affiliation was determined by having students report which crowd they believed the other students belonged to. The possible options were the popular crowd, the smart crowd, the normal crowd, or none of the above. Students who were reported to belong in multiple groups were classified as multiple crowd and students not reported to belong in any group were classified as outsiders.

2.2 Instruments

Self-Perception Profile

The Self-Perception Profile (Harter, 1985) is a 36 item survey of self-perception. The items are divided into six subscales: scholastic competence, social acceptance, athletic competence, physical appearance, behavioral conduct, and global self-worth. Scholastic measures perceived ability in academics, social looks at perceived popularity with their peers, physical is satisfaction with how they look , behavioral is belief in their ability to meet adult expectations, and finally global considers general self-worth outside of any specific domain.

The measure uses a structured alternative format. For each item the student is presented with two opposing statements and asked to decide which one best describes him or her. For example, an item from the Scholastic Competence scale states, “Some kids often forget what they learn BUT other kids can remember things easily.” The student then decides whether the chosen statement is “really” or “sort of” true for him or her. Each item is scored from 1 to 4 with higher scores reflecting a more positive self perception.

Teen Health Survey

Students were asked to fill out a self-report survey on a variety of student behaviors and demographic information. Relevant to this study are the background section and the drug use sections. Background questions included gender and ethnicity the students identified as, as well as average grades and their parents' education levels. The drug use section contained questions about students interactions with alcohol, cigarettes, and marijuana. Questions asked include whether students had ever used the substance, at what age they first used, how often they used in the last year, how often they used with friends, and if they had ever been offered the substance at school.

2.3 Variables Used in the Analysis

Demographics The two demographic vaiables used in the analysis are gender and race. Gender indicates whether the individual identifies as male(1) or female(2) as reported in the seventh grade. Race indicates what ethnicity the student identifies as as reported in the seventh grade. It is coded as Black(1), Asian(2), or other(3).

Crowd Variable Crowd indicates what social group the student was reported as being in in the seventh grade using dummy variables: Pop/Jock(1,) Smart(2,) Normal(3), Multiple Crowd(4), or Outsider(5). For the Marijuana GEE group 5 (outsider) was put into Smart group (2) due to convergence issues that will be discussed further in Chapter 4.

Self-Perception Variables These are the continuous variables resulting from the Self-perception Profile. Larger numbers indicate a better self-perception. The six variables are: Scholastic, Athletic, Social, Physical, Behavior, Global which measure self-perception of academic performance, peer popularity, ability at sports and outdoor games, physical attractiveness, ability of meet adult expectations, and an overall sense of self, respectively.

Substance Use Variables Alcohol, Cigarettes, and Marijuana are the response variables. Each is a repeated measures binary variable of whether a student has drank, smoked cigarettes, or used marijuana in the past year. Each student has a response for seventh and eighth grade. The coding is 1 if yes, 0 if no.

Table 1: Variables Used in the Model

Variable	Description
Gender	Male or Female
Race	Black, Asian, or Other
Crowd	Pop/Jock, Smart, Normal, Multiple Crowd, or Outsider
Scholastic	Perception of academic performance
Athletic	Perception of ability at sports and outdoor games
Social	Perception popularity among peers
Physical	Perception of physical attractiveness
Behavior	Perception of ability to meet adult expectations
Global	Overall self-perception
Alcohol	Drank alcohol in the past year?
Cigarettes	Smoked cigarettes in th past year?
Marijuana	Used marijuana in the past year?

Chapter 3. Statistical Methods

3.1 Generalized Linear Modeling

Generalized linear models (GLM) are a generalization of standard linear regression that allow the response variables to have a distribution other than the Gaussian. GLMs have two assumptions about the distribution of the responses. First, given the x_i, responses y_i are assumed to be independent of one another. Secondly, the distribution of the y_i must belong the exponential family. To be a member of the exponential family the probability density function must be able to be written as

$$f(y_i|\theta_i, \phi, w_i) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi}w_i + c(y_i, \phi, w_i)\right\}, \tag{1}$$

where θ_i is the natural parameter, ϕ is a scaler, w_i is a weight depending on whether the data is grouped, and b and c are well known functions relating to the type of exponential family. An example of functions b and c can be found in Fahrmeir and Tutz.

General linear models also have structural assumptions. The expected value $E(y_j|x_j) = \mu_i/\phi$. The expected value μ_i is related the the linear predictor $\eta_i = z_i'\beta$ through the function $\mu_i = h(\eta_i) = h(z_i'\beta$ $\eta_i) = g(\mu_i)$ where h is a known one-to-one function, β is a vector of unknown parameters, z_i is a design vector determined as an appropriate function of the covariates, and g is the link or inverse function of h. So, a GLM can be characterized by three parts, the member of the exponential family being used, the link function, and the design vector.

3.2 Logistic Regression

The data for the YACS study is a set of Bernoulli trials which is a special case of the binomial distribution. The values for the response y are 1 if there is a 'success', and 0 otherwise. For this data set a 'success' would be if the student has used the substance within the last year. The distribution of the binomial is

$$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}. \tag{2}$$

Thus, there exists some probability π that an observation will be a 'success'. When data follow a Bernoulli or binomial distribution, an appropriate GLM is logistic regression. The scale parameter ϕ is set to one and the link function η will equal $log(\frac{\mu}{1-\mu})$. In logistic regression the unknown probability π is estimated using a set of regressors. The basic form of a logistic regression is

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta' x; \tag{3}$$

with binary data since $\pi = \mu$. Probability π can be achieved be rearranging the equation to get

$$\pi = h(\eta) = \frac{exp(\eta)}{1 + exp(\eta)}, \tag{4}$$

where η is equal to $\alpha + \beta' x$. The right hand side of (3) is a standard linear predictor setting, where α is an estimate of the intercept term, β is an estimate vector of k coefficients, and x is a vector of data that corresponds to values for each of the k regressor variables. The left hand side of (3) is a log odds, which measures the weight of the probability of success π over nonsuccess $1 - \pi$. For the YACS data this would be the probability of using the substance based on the demographic, crowd, and self-perception variables. If π were known, the odds of a success to nonsuccess can be calculated as

$$\frac{\pi}{1 - \pi} = m \qquad \text{where} \qquad m \geq 0 \qquad . \tag{5}$$

This odds can be estimated from the equation of the logistic regression with the set of values x. Using x the estimated odds is

$$\frac{\pi(x)}{1 - \pi(x)} = exp\left(\alpha + \beta' x\right) \qquad . \tag{6}$$

3.3 Generalized Estimating Equations

Generalized linear models including logistic regression work under the assumption that the data are independent, however this is not always the case. Often data are clustered, for example observations taken from trees growing near each other, the two eyes of an individual, or multiple responses from the same individual over time. These clustered data are potentially correlated within cluster which must be taken into account during analysis. There are several extensions of generalized linear modeling that can handle correlated data. One of these is Generalized Estimating Equations (GEE). GEE's are a quasi-likelihood method meaning the assumption of being a member of the exponential family is discarded and only the first and second moments need to be specified.

There are two types of GEEs: subject specific and population averaged or marginal models. Subject specific models do not assume a common correlation of the repeated measures across subjects, rather the correlation is allowed to vary so that each subject may have a good fit for their own responses. This model is typically used for data such as dose response curves or growth curves where the relationship of responses for each individual is of interest. In contrast, population averaged or marginal models do assume a common correlation of the repeated response measures. This model is used where multiple responses are taken on individuals but each individual's pattern of responses is not of interest.

In the YACS study we are not interested in the individual patterns of response, but rather the overall effect of the regressors on the response so marginal models are the most appropriate analysis for the data.

As explained in Section 3.2 the probability of success can be estimated by the expected value of y given the explanatory data x. We have

$$P(y_j = 1|x_j) = \pi_j(\beta_j) = E(y_{ij}|x_{ij}), \tag{7}$$

for $j = 1, \ldots, m$.

This can be written as $h(z'_{ij}\beta)$, where h is an inverse logit function and $z_{ij} = z_{ij}x_{ij}$ is a design vector for discrete and continuous variables. The variance $v(\pi_j)$ is then $\pi_j(1-\pi_j)$. The covariance within cluster, (y_j, y_k), is a function of the marginal mean plus an additional parameter α, that is $cov(y_j, y_k) = c(\pi_j, \pi_k; \alpha)$.

If $y_i = (y_{i1}, \ldots, y_{im_i})$ and $x_i = (x_{i1}, \ldots, x_{im_i})$ are the responses and observations for a cluster i, i from $1, \ldots, n$, and $\pi_i(\beta) = \pi_{i1}(\beta_1), \ldots, \pi_{im_i}(\beta_{m_1}), \Sigma_i(\beta, \alpha)$ are the marginal means vectors and working covariance matrices, then for fixed α the GEE equation for vector β is

$$S_\beta(\beta, \alpha) = \sum_{i=1}^{n} Z'_i D_i(\beta) \Sigma_i^{-1}(\beta, \alpha)(y_i - \pi_i) = 0 \tag{8}$$

where Z'_i is a design matrix $(z_{i1}, \ldots, z_{im_i})$, and $D_i(\beta)$ is a diagonal matrix $\text{diag}\{D_{ij}(\beta_j)\} = \partial h_j/\partial\eta_{ij}$ evaluated at $\eta_{ij} = z'_{ij}\beta_j$.

The second GEE to estimate α must then be added. If for each cluster the vector $w_i = (y_{i1} \cdot y_{i2}, y_{i1} \cdot y_{i3}, \ldots, y_{im_{i-1}} \cdot y_{im_i})$ and $\theta_i = E(w_i)$ denotes the vector of second moments $\theta_{ijk} = E(y_{ij}y_{ik}) = P(y_{ij} = y_{ik} = 1)$ then the GEE for α is

$$S_\alpha(\beta, \alpha) = \sum_{i=1}^{n} \left(\frac{\partial\theta_i}{\partial\alpha}\right) (C_i)^{-1}(w_i - \theta_i) = 0 \tag{9}$$

with working covariance matrix C_i for $\text{cov}(w_i)$. Once the equations for β and α are specified, estimates are obtained using Fisher scoring equations for S_β and S_α by alternating between the two parameters.

Correlation Structure

The estimator for α is seen in (8) where w_i is the vector of pairwise products $y_{ij}y_{ik}$, C_i is a working covariance matrix for w_i, and θ_i is the expected value of w_i.

The simplest correlation model is the independence model. This assumes no correlation within clusters, adds no additional parameters to the estimating equations, and has the identity matrix as the C_i working covariance matrix. The autroregressive structure AR(1) assumes a temporal dependence within clusters. The elements of the covariance matrix take the form $y_{ij}y_{ik} = \alpha^{|k-j|}$. This structure imposes only one parameter α and the level of correlation depends on the distance between times. As α is a number between zero and one, as it is raised to higher powers the correlation will decrease. Measurements taken closely together in time are assumed to be more correlated than measurements take further apart. For the YACS data there are only two time points in each cluster, therefore only one correlation, however this structure leaves open the option to easily add in more years of data without having to alter the model.

QIC

The Akaike Information Criteria (AIC) is a goodness of fit measure for likelihood based models. It is defined as $AIC = -2L + 2p$ where L is the log likelihood and p is the number of parameters in the model. An extension of AIC is QIC, the quasi-likelihood under the independence model information criteria. This measure is more appropriate for GEE which is a quasi-likelihood method. Like the AIC, in QIC lower numbers are better. QIC is defined as

$$QIC(R) = -2Q(g^{-1}(x\beta_R)) + 2trace(A_I^{-1}V_{MS,R}). \tag{10}$$

The first half of the sum, $-2Q(g^{-1}(x\beta_R))$ is the value of the quasi-likelihood calculated with the proposed correlation structure R and $g^{-1}(x\beta_R = \hat{\pi}$ where g^{-1} is the inverse link function for the model, a logit for this model. In the second half of the sum, we can define A_I as the variance matrix under the independence model. We define $V_{MS,R}$ as the sandwich estimate of variance under the hypothesized correlation structure R.

Parameter Estimates

Results are given in terms of log odds. The odds ratio for a variable can be calculated by exponentiating its individual β. As seen in equation 11, for continuous variables the odds ratio is the multiplicative effect on odds of success if the value of the variable increases, holding all other variables constant. For categorical variables it is a ratio for two levels of the variable. This is then interpreted as the odds of success over failure for one level of the variable as compared to the other, holding all other variables constant.

$$\frac{\left.\frac{\pi(x)}{1-\pi(x)}\right|_{x_i+1}}{\left.\frac{\pi(x)}{1-\pi(x)}\right|_{x_i}} = \exp(\beta_i) \quad . \tag{11}$$

YACS Model

For the YACS data each student is a cluster with two measurements, 1 each from seventh and eighth grade. The fully specified model for the alcohol and cigarettes is:

$$\begin{aligned}\log\left(\frac{\pi_1}{1-\pi_1}\right) = \log\left(\frac{\pi_2}{1-\pi_2}\right) &= \beta_0 + \beta_1 Race(1) + \beta_2 Race(2) + \beta_3 Sex + \beta_4 Crowd(1) + \beta_5 Crowd(2)\\ &+ \beta_6 Crowd(3) + \beta_7 Crowd(4) + \beta_8 Scholastic + \beta_9 Social + \beta_{10} Athletic\\ &+ \beta_{11} Physical + \beta_{12} Behavior + \beta_{13} Global\end{aligned}$$

The fully specified model for marijuana is:

$$\begin{aligned}\log\left(\frac{\pi_1}{1-\pi_1}\right) = \log\left(\frac{\pi_2}{1-\pi_2}\right) &= \beta_0 + \beta_1 Race(1) + \beta_2 Race(2) + \beta_3 Sex + \beta_4 Crowd(1) + \beta_5 Crowd(2)\\ &+ \beta_6 Crowd(3) + \beta_7 Scholastic + \beta_8 Social + \beta_9 Athletic + \beta_{10} Physical\\ &+ \beta_{11} Behavior + \beta_{12} Global\end{aligned}$$

Chapter 4. Results

Analysis was completed in SAS with *Proc Genmod* using the repeated statement and correlation type option. Full code can be found in Appendix A. The GEE analysis converged for both the alcohol and cigarette models. As mentioned, before the GEE did not converge for the marijuana model, since there were not any observations of success for the students in the smart crowd, creating a zero count for that cell. One solution to this problem is to merge the Smart group with another group. The outsider crowd had not shown any significant differences from the Smart in the other two models and both groups had similar response patterns. Thus, for the marijuana model the smart and outsider groups were merged. The GEE converged after this alteration.

For all three models there were no strong differences between the Independent and AR(1) correlation structures. QIC for the Independence Alcohol model was 499.43 and for the AR(1) was 499.33, a difference of only 0.1. Differences for the Cigarette and Marijuana QICs were similar. As AR(1) is the more theoretically appropriate structure for this model all estimates mentioned hereafter are from the AR(1). Discussion of significant parameters is below. Tables showing full GEE results can be found in Appendices B1, B2, and B3 for alcohol, cigarettes, and marijuana respectively.

4.1 Alcohol Results

The following table contains the significant parameters from the Alcohol GEE analysis:

Table 2: Significant Parameters from the Alcohol GEE Analysis

Effect	Beta Estimate	Std. Error	95% CI	Odds Ratio
Black-Asian	1.5826	0.3269	(0.9419, 2.2233)	4.8676
Asian-Other	-1.3692	0.3520	(-2.0591, -0.6793)	0.2543
Pop/Jock-Smart	2.3786	0.8335	(0.7449, 4.0122)	10.7898
Pop/Jock- Normal	2.1231	0.5280	(1.0881, 3.1580)	8.3570
Pop/Jock-Multiple Crowd	1.8098	0.4915	(0.8465, 2.7731)	6.1092
Pop/Jock-Other	1.3788	0.5333	(0.3336, 2.4241)	3.9701
Normal-Other	-0.7442	0.3665	(-1.4625, -0.0260)	0.4751
Behavior	-0.8973	0.2476	(-1.3825, -0.4121)	0.4077

For the demographic variables there was a significant effect for race. Both Black and other students are significantly more likely to drink than Asian students. For Black over Asian students the odds ratio is 4.87 indicating the odds of having drank alcohol in the past year are 4 times greater if a student is Black rather than Asian, controlling for other variables. The crowd variable shows a significant effect for Pop/Jock group against all other groups. Pop/Jocks are significantly more likely to drink than all of the other crowds, controlling for other variables. Specific odds ratios can be seen in Table 2. There is also a significant crowd effect for students in the Outsider crowd, showing they are more likely to drink than students in the normal crowd. Finally, there was a significant effect for the behavior self-perception variable. Students are less likely to drink when they have a higher perception of their ability to meet expectations.

4.2 Cigarette Results

The significant parameters from the Cigarette GEE analysis can be seen in Table 3.

Table 3: Significant Parameters from the Cigarette GEE Analysis

Effect	Beta Estimate	Std. Error	95% CI	Odds Ratio
Smart-Normal	-1.5431	0.8750	(-3.2581, 0.1719)	0.2137
Social	0.8111	0.3086	(0.2063, 1.4158)	2.2504
Behavior	-0.9902	0.2581	(-1.4960, -0.4844)	0.3715

The cigarette GEE analysis showed no significant effects for any of the demographic or the crowd variables. There was a marginally significant effect with normal students being more likely to smoke that smart students controlling for other variables. The self-perception profile did have two significant variables. Students have a multiplicative increase of 2 in odds of smoking as their perception of how peers see them increases. Also, mirroring the alcohol results, there is a negative effect of behavioral self-perception.

4.3 Marijuana Results

Table 4: Significant Parameters from the Marijuana GEE Analysis

Effect	Beta Estimate	Std. Error	95% CI	Odds Ratio
Black-Other	0.8686	0.3994	(0.0859,1.6513)	2.3836
Black-Asian	2.6595	0.4843	(1.7102,3.6088)	14.2891
Asian-Other	-1.7909	0.5255	(-2.8209, -0.7610)	0.1668
Pop/Jock-Smart/Outsider	1.9232	0.6386	(0.6715, 3.1749)	6.8428
Pop/Jock- Normal	1.7036	0.6142	(0.4999, 2.9073)	5.4937
Pop/Jock-Multiple Crowd	1.9537	0.5819	(0.8133, 3.0941)	7.0547
Behavior	-0.7737	0.3131	(-1.3875, -0.1600)	0.4613
Global	-0.6443	0.3560	(-1.3420, -0.0534	0.5250

GEE analysis for marijuana showed significance for all race comparisons. Black students are more likely to have used marijuana then either Asian or other students. Asian students are also less likely to use marijuana then the other students. The Black to Asian odds ratio is the most striking. Black students have odds 14.26 times higher than Asian students for using marijuana, controlling for other variables. For the crowd variable there is again a significant effect for Pop/Jock students compared to all other crowds, with Pop/Jocks being having odds of at least 5 times higher than the comparison crowd, again controlling for other variables. Interestingly the behavior variable has a significant negative effect, making it the only variable to have an effect on all three models. Global self-perception also has a negative effect, so students are less likely to use marijuana when they have a higher overall perception of themselves.

Chapter 5. Discussion

Generalized Estimating Equations appears to be a reasonable model for the YACS data as it can handle repeated measures binary data. Although there were issues with convergence, they were able to be solved without impeding the intent of analysis to identify important factors in middle school drug use. The analysis did find significant factors to drug use. The most important factor to decrease overall odds for all drugs is the behavioral measure. Students who feel that they are capable of meeting adult expectations were less likely to smoke, drink, or use marijuana. Conversely being in the popular crowd significantly increases odds of drinking and using marijuana. There were no differences shown for the two correlation structures in this data analysis. In the future it would be interesting to see how adding more years of data would affect this conclusion. Other models such as conditional models could also be pursued here.

References

Dolcini, M.M. and Adler, N.E. (1994)“Perceived Competencies, Peer Group Affiliation, and Risk Behavior Among Early Adolescents” *Health Psychology,13*(6), 496-506.

Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modeling Based on Generalized Linear Models (2nd ed.).* New York, New York: Springer-Verlag.

Jones, S.P. and Heaven, P.C.L. (1998) “Psychosocial Correlates of Adolescent Drug-taking Behaviour” *Journal of Adolescence, 21*(2), 127-134.

Hardin, J.W and Hilbe, J.M. (2003) *Generalized Estimating Equations.* Boca Raton, Florida: CRC Press LLC.

La Greca, A.M., Prinstein, M.J., & Fetter, M.D. (2001)“Adolescent Peer Crowd Affiliation: Linkages With Health-Risk Behaviors and Close Friendships” *Journal of Pediatric Psychology, 26*(3), 131-143.

Lakon, C.M. and Valente, T.W. (2012) “Social integration in friendship networks: The synergy of network structure and peer influence in relation to cigarette smoking among high risk adolescents” *Social Science & Medicine, 74*(9), 1407-1417.

Liang, K. and Zeger, S.L. (1986) “Longitudinal Data Analysis using Generalized Linear Models” *Biometika,73*(1), 13-22.

Selnow, G.W., Crano, WD. (1986) “Formal vs informal group affiliations: Implications for alcohol and drug use among adolescents” *Journal of Studies on Alcohol, 47*(1), 48-52.

Williams J.G., Smith J.P. (1993)“Alcohol and other drug use among adolescents: Family and peer influences’ *Journal of Substance Abuse, 5*(3), 289-294.

Zeger, S.L., Liang, K., and Albert, P.S. (1988) “Models for Longitudinal Data: A Generalized Estimating Approach” *Biometrics, 44*, 1049-1060.

Appendix A: SAS Code

```
proc genmod data=thesis.Newy descending;
class race sex Crowd idd;
model Alcohol= race sex Crowd Scholastic Social Athletic Physical Behavior Global/ d=bin itprint;
repeated subject=idd / corrw type =ar(1) modelse;
Title1 'GEE for alcohol use with AR1 correlation structure';
run;

proc genmod data=thesis.Newy descending;
class race sex allcrwd idd;
model Alcohol=race sex Crowd Scholastic Social Athletic Physical Behavior Global/ d=bin itprint;
repeated subject=idd / corrw type=ind modelse;
Title1 'GEE for alcohol use with Independent correlation structure';
run;

proc genmod data=thesis.Newypot descending;
class race sex Crowd idd;
model Marijuana= race sex Crowd Scholastic Social Athletic Physical Behavior Global/ d=bin itprint;
repeated subject=idd / corrw type =ar(1) modelse;
Title1 'GEE for Pot use with AR1 correlation structure';
run;

proc genmod data=thesis.Newypot descending;
class race sex Crowd idd;
model Marijuana= race sex Crowd Scholastic Social Athletic Physical Behavior Global/ d=bin itprint;
repeated subject=idd / corrw type=ind modelse;
Title1 'GEE for pot use with Independent correlation structure';
run;

proc genmod data=thesis.Newy descending;
class race sex Crowd idd;
model Smoked= race sex Crowd Scholastic Social Athletic Physical Behavior Global/ d=bin itprint;
repeated subject=idd / corrw type =ar(1) modelse;
Title1 'GEE for cigarette use with AR1 correlation structure';
run;

proc genmod data=thesis.Newy descending;
class race sex Crowd idd;
model Smoked= race sex Crowd Scholastic Social Athletic Physical Behavior Global/ d=bin itprint;
repeated subject=idd / corrw type=ind modelse;
Title1 'GEE for cigarette use with Independent correlation structure';
run;
```

Appendix B: Full GEE Results Tables

Parameters significant with $p < 0.05$ are indicated with **
Parameters with $p < 0.1$ are indicated with *

B.1 Alcohol

Table 5: Demographic Results from Alcohol GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Black-Other	Independent	0.2273	0.2978	(-0.3564, 0.8110)	1.2552
	AR(1)	0.2134	0.3223	(-0.4183, 0.8452)	1.2379
Black-Asian	Independent	1.5771 **	0.2919	(1.0050, 2.1493)	4.8409
	AR(1)	1.5826 **	0.3269	(0.9419, 2.2233)	4.8676
Asian-Other	Independent	-1.3498 **	0.3247	(-1.9863, -0.7133)	0.2593
	AR(1)	-1.3692 **	0.3520	(-2.0591, -0.6793)	0.2543
Male-Female	Independent	0.3808	0.2530	(-0.1151, 0.8767)	1.4635
	AR(1)	0.3774	0.2747	(-0.1609, 0.9157)	1.4585

Table 6: Crowd Results from Alcohol GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Pop/Jock-Smart	Independent	2.4125 **	0.7001	(1.0403, 3.7848)	11.1618
	AR(1)	2.3786 **	0.8335	(0.7449,4.0122)	10.7898
Pop/Jock- Normal	Independent	2.1576 **	0.4903	(1.1967, 3.1185)	8.6504
	AR(1)	2.1231 **	0.5280	(1.0881,3.1580)	8.3570
Pop/Jock-Multiple Crowd	Independent	1.8260 **	0.4566	(0.9311, 2.7210)	6.2090
	AR(1)	1.8098 **	0.4915	(0.8465, 2.7731)	6.1092
Pop/Jock-Outsider	Independent	1.4083 **	0.4956	(0.4370,2.3795)	4.0890
	AR(1)	1.3788 **	0.5333	(0.3336, 2.4241)	3.9701
Smart -Normal	Independent	-0.2549	0.6804	(-1.5884,1.0786)	0.7750
	AR(1)	-0.2555	0.7407	(-1.7073, 1.1963)	0.7745
Smart- Multiple Crowd	Independent	-0.5865	0.6603	(-1.8807, 0.7077)	0.5563
	AR(1)	-0.5687	0.7193	(-1.9786, 0.8411)	0.5663
Smart-Outsider	Independent	-1.0043	0.6732	(-2.3237, 0.3152)	0.3663
	AR(1)	-0.9997	0.7334	(-2.4371, 0.4377)	0.3680
Normal-Multiple Crowd	Independent	-0.3316	0.3033	(-0.9260, 0.2628)	0.7178
	AR(1)	-0.3133	0.3292	(-0.9586, 0.3321)	0.7310
Normal-Outsider	Independent	-0.7493 **	0.3378	(-1.4115, -0.0872)	0.4727
	AR(1)	-0.7442 **	0.3665	(-1.4625, -0.0260)	0.4751
Multiple Crowd-Outsider	Independent	-0.4178	0.3087	(-1.0228, 0.1873)	0.6585
	AR(1)	-0.4310	0.3353	(-1.0882, 0.2262)	0.6499

Table 7: Self Perception Results from Alcohol GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Scholastic	Independent	0.4589	0.2431	(-0.0177, 0.9354)	1.5823
	AR(1)	0.4646 *	0.2639	(-0.0526, 0.9817)	1.5914
Social	Independent	0.4410 *	0.2605	(-0.0696, 0.9516)	1.5543
	AR(1)	0.4467	0.2825	(-0.1070, 1.0003)	1.5631
Athletic	Independent	0.0082 **	0.2202	(-0.4234, 0.4399)	1.0082
	AR(1)	0.0117	0.2387	(-0.4561, 0.4795)	1.0118
Physical	Independent	-0.0948	0.2521	(-0.5890, 0.3994)	0.9096
	AR(1)	-0.0901	0.2735	(-0.6260, 0.4459)	0.9138
Behavior	Independent	-0.8870 **	0.2279	(-1.3336, -0.4404)	0.4119
	AR(1)	-0.8973 **	0.2476	(-1.3825, -0.4121)	0.4077
Global	Independent	-0.3235	0.2511	(-0.8157, 0.1686)	0.7236
	AR(1)	-0.3277	0.2725	(-0.8618, 0.2065)	0.7206

B.2 Cigarettes

Table 8: Demographic Results from Cigarette GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Black-Other	Independent	-0.1209	0.3314	(-0.7703 0.5286)	0.8861
	AR(1)	-0.1214	0.3602	(-0.8274 0.5847)	0.8857
Black-Asian	Independent	0.1790	0.3006	(-0.4101, 0.7680)	1.1960
	AR(1)	0.1847	0.3643	(-0.4079, 1.0200)	1.2029
Asian-Other	Independent	-0.2998	0.3347	(-0.9559, 0.3562)	0.7410
	AR(1)	-0.3061	0.3643	(-1.0200, 0.4079)	0.7363
Male-Female	Independent	-0.2998	0.3347	(-0.9559, 0.3562)	0.7410
	AR(1)	-0.3061	0.3643	(-1.0200, 0.4079)	0.7363

Table 9: Crowd Results from Cigarette GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Pop/Jock-Smart	Independent	1.0431	0.8961	(-0.7132, 2.7993)	2.8380
	AR(1)	1.0353	0.9769	(-0.8795, 2.9500)	2.8160
Pop/Jock- Normal	Independent	-0.4915	0.5014	(-1.4743, 0.491)	0.6117
	AR(1)	-0.5078	0.5460	(-1.5779, 0.5623)	0.6018
Pop/Jock-Multiple Crowd	Independent	-0.0734	0.4819	(-1.0178, 0.8710)	0.9292
	AR(1)	-0.0766	0.5251	(-1.1057, 0.9526)	0.9263
Pop/Jock-Outsider	Independent	-0.0505	0.5197	(-1.0690, 0.9680)	0.9508
	AR(1)	-0.0509	0.5662	(-1.1605, 1.0588)	0.9504
Smart -Normal	Independent	-1.5346 *	0.8024	(-3.1073, 0.0381)	0.2155
	AR(1)	-1.5431 *	0.8750	(-3.2581, 0.1719)	0.2137
Smart- Multiple Crowd	Independent	-1.1165	0.7938	(-2.6723, 0.4393)	0.3274
	AR(1)	-1.1118	0.8659	(-2.8089, 0.5852)	0.3290
Smart-Outsider	Independent	-1.0936	0.8131	(-2.6873, 0.5002)	0.3350
	AR(1)	-1.0861	0.8872	(-2.8249, 0.6526)	0.3375
Normal-Multiple Crowd	Independent	0.4181	0.3113	(-0.1920, 1.0283)	1.5191
	AR(1)	0.4312	0.3387	(-0.2326, 1.0951)	1.5391
Normal-Outsider	Independent	0.4411	0.3583	(-0.2612, 1.1434)	1.5544
	AR(1)	0.4569	0.3900	(-0.3075, 1.2214)	1.5792
Multiple Crowd-Outsider	Independent	0.0229	0.3485	(-0.6601, 0.7060)	1.0232
	AR(1)	0.0257	0.3800	(-0.7191, 0.7705)	1.0260

Table 10: Self Perception Results from Cigarette GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Scholastic	Independent	0.4049	0.2620	(-0.1087, 0.9185)	1.4991
	AR(1)	0.3944 *	0.2851	(-0.1643, 0.9532)	1.4835
Social	Independent	0.8088 **	0.2837	(0.2528, 1.3649)	2.2452
	AR(1)	0.8111 **	0.3086	(0.2063, 1.4158)	2.2504
Athletic	Independent	0.1962	0.2471	(-0.2880, 0.6805)	1.2168
	AR(1)	0.2045	0.2689	(-0.3225, 0.7314)	1.2269
Physical	Independent	-0.4783*	0.2726	(-1.0126, 0.0560)	0.6198
	AR(1)	-0.4856	0.2968	(-1.0673, 0.0961)	0.6153
Behavior	Independent	-0.9865 **	0.2370	(-1.4509, -0.5220)	0.3729
	AR(1)	-0.9902 **	0.2581	(-1.4960, -0.4844)	0.3715
Global	Independent	-0.4496*	0.2623	(-0.9637, 0.0645)	0.6379
	AR(1)	-0.4498	0.2857	(-1.0098, 0.1103)	0.6378

B.3 Marijuana

Table 11: Demographic Results from Marijuana GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Black-Other	Independent	0.8921 **	0.3157	(0.2733, 1.510)	2.4402
	AR(1)	0.8686 **	0.3994	(0.0859, 1.6513)	2.3836
Black-Asian	Independent	2.6544 **	0.3794	(1.9108, 3.3980)	14.2165
	AR(1)	2.6595 **	0.4843	(1.7102, 3.6088)	14.2891
Asian-Other	Independent	-1.7623 **	0.4134	(-2.5726, -0.9520)	0.1717
	AR(1)	-1.7909 **	0.5255	(-2.8209, -0.7610)	0.1668
Male-Female	Independent	0.4499	0.2856	(-0.1100, 1.0097)	1.5682
	AR(1)	0.4153	0.3641	(-0.2983, 1.1290)	1.5148

Table 12: Crowd Results from Marijuana GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Pop/Jock-Smart/Outsider	Independent	1.8855 **	0.5016	(0.9025, 2.8685)	6.5896
	AR(1)	1.9232 **	0.6386	(0.6715, 3.1749)	6.8428
Pop/Jock- Normal	Independent	1.7249 **	0.4807	(0.7827, 2.6671)	5.6120
	AR(1)	1.7036 **	0.6142	(0.4999, 2.9073)	5.4937
Pop/Jock-Multiple Crowd	Independent	1.9906 **	0.4560	(1.0969, 2.8843)	7.3199
	AR(1)	1.9537 **	0.5819	(0.8133, 3.0941)	7.0547
Smart/Outsider -Normal	Independent	-0.1606	0.3711	(-0.8879, 0.5667)	0.8516
	AR(1)	-0.2196	0.4708	(-1.1424, 0.7032)	0.8028
Smart/Outsider- Multiple Crowd	Independent	0.1051	0.3696	(-0.6193, 0.8295)	1.1108
	AR(1)	0.0305	0.4691	(-0.8889, 0.9499)	1.0310
Normal-Multiple Crowd	Independent	0.2657	0.3334	(-0.3878, 0.9191)	1.3043
	AR(1)	0.2501	0.4231	(-0.5792, 1.0794)	1.2842

Table 13: Self Perception Results from Marijuana GEE

Effect	WCS	Beta estimate	Std. Error	95% CL	Odds Ratio
Scholastic	Independent	-0.0363	0.2659	(-0.5576, 0.4849)	0.9644
	AR(1)	-0.0270	0.3377	(-0.6890, 0.6349)	0.9734
Social	Independent	0.5297 *	0.2941	(-0.0467, 1.1062)	1.6984
	AR(1)	0.5467	0.3740	(-0.1862, 1.2797)	1.7275
Athletic	Independent	-0.0091	0.2481	(-0.4954, 0.4773)	0.9909
	AR(1)	-0.0476	0.3131	(-0.6612, 0.5659)	0.9535
Physical	Independent	0.0855	0.2688	(-0.4413, 0.6124)	1.0893
	AR(1)	0.1004	0.3412	(-0.5682, 0.7691)	1.1056
Behavior	Independent	-0.7614 **	0.2453	(-1.2422, -0.2806)	0.4670
	AR(1)	-0.7737 **	0.3131	(-1.3875, -0.1600)	0.4613
Global	Independent	-0.6194 **	0.2786	(-1.1655, -0.0733)	0.5383
	AR(1)	-0.6443 **	0.3560	(-1.3420, 0.0534)	0.5250

Vita

Lauren Ashley Beacham was born to Bobby and Sara in Decatur, Ga. She has one older brother, Christopher. She grew up in nearby Lawrenceville, Ga with a typical childhood of ballet lessons, neighborhood friends, and her brother's baseball games. Lauren was always the bookish type and decided to attend Agnes Scott College, a women's liberal arts institution. While at college she majored in psychology and worked as a research assistant at a nearby hospital. She received her degree in 2009 and continued her research work which she enjoyed but decided data was her real passion. She applied to LSU for it's program as well as it's proximity to delicious Cajun food. For two years Baton Rouge has provided a great education as well as many fond memories and crawfish boils. After graduation Lauren plans to find work in Atlanta so she can return home to her friends and family.