8-2002

Models and Graphics in the Analysis of Categorical Variables: The Case of the Youth Tobacco Survey.

Deborah Susan Hosler
East Tennessee State University

Models and Graphics in the Analysis of Categorical Variables: The Case of the Youth Tobacco Survey

---

A Thesis

Presented to the Faculty of the Department of Mathematics

East Tennessee State University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Mathematical Sciences

---

by

Deborah Susan Hosler

August, 2002

Edith Seier, Ph.D., Chair

Robert Price, Ph.D.

T. Henry Jablonski, Jr., M.P.H.

Keywords: Youth Tobacco Survey, Logistic Regression, Log-linear Modeling, Correspondence Analysis, Multivariate Analysis, Odds Ratio

ABSTRACT

Models and Graphics in the Analysis of Categorical Variables: The Case of the Youth Tobacco Survey

by

Deborah Susan Hosler

Youth Tobacco Surveys have been conducted in several states in the U.S. in recent years in order to design policies with the goal of reducing tobacco use among young people. Some primary analysis of those surveys has been done, but few analyses include modeling, and the study of independence has been addressed, mainly, in the bivariate context.

In this work contemporary methods, which are of relative recent appearance in categorical data analysis, will be examined, including logistic and log-linear modeling as well as graphical displays and correspondence analysis. These methods will be applied to data from the 2000 Tennessee Youth Tobacco Survey.

The objective is to demonstrate that methods of multivariate categorical data analysis can provide fresh insight about the behavior of adolescents with respect to tobacco use. The ultimate purpose of this work is to recommend methodology that goes beyond that which is currently published.

DEDICATION

To my mentors

Kate Hosler,

Jeanne Munns,

Karen Dempsey,

Doris and Bernice Allen,

Edith Hosler,

and

Maude Melton.

ACKNOWLEDGMENTS

I would like to use this space to express my heartfelt appreciation of Dr. Edith Seier for her intelligence, knowledge, and infinite patience. I also give my thanks to the staff and faculty of the Department of Mathematics at ETSU. Last, but not least, I am grateful for my husband, Bill Dobbins, who made sure this learning mind never went hungry.

Contents

List of Tables

List of Figures

1 INTRODUCTION

In consideration of the facts that, nationally, tobacco use is deemed the leading preventable cause of human deaths and that nearly eighty percent of habitual users begin using tobacco prior to age eighteen, the United States Centers for Disease Control and Prevention (CDC), made a recommendation in 1998 *that states establish and maintain comprehensive tobacco control programs to reduce tobacco use among youth* [40] (p.2). The primary methods by which this task will be accomplished are surveillance and evaluation. Surveillance is initially needed to determine current patterns of tobacco use as it relates to various elements of the adolescent physical and social environments. This basis of knowledge can then be used in the effort to implement well designed tobacco control programs, which will, in turn, require information provided by further surveillance to aid in effective evaluation of these programs. To this end the CDC formally created the Youth Tobacco Surveillance and Evaluation System (YTSES), two components of which are the National Youth Tobacco Survey (NYTS) and the state Youth Tobacco Survey (YTS). These relatively large-scale surveys are specifically designed to gather the data necessary to facilitate the states' development and subsequent evaluation of such programs.

The year 2000 version of the YTS consists of a 63–item questionnaire [41] to which participating states are allowed to add questions according to individual need. The core questionnaire requests sampled students' responses for *questions about tobacco use, exposure to environmental tobacco smoke, smoking cessation, school curriculum, minor's ability to purchase or otherwise obtain tobacco products, knowledge and attitude about tobacco, and familiarity with pro-tobacco and anti-tobacco media messages*

[40] (p.4). Generally, the variables measured by such questions are either categorical in nature, including nominal and ordinal types, or discrete. Forty-five state youth tobacco surveys were administered from years 1998 through 2000, with sample sizes ranging from 452 to 33,586 students [40] [41]. The resulting data sets tend to be rather large since they include measurements of many categorical variables for large numbers of respondents. The goal of this research is to appraise methods currently employed by states to analyze the data produced in youth tobacco surveys, as well as to explore and showcase contemporary methods of categorical data analysis that might also be used by states to obtain from such data the relevant information required for design of enlightened control programs.

Three methods of data analysis that could provide more sophisticated means of surveillance and evaluation, but which to this point have been mostly overlooked, are examined in the sections that follow. Data from the 2000 Tennessee Youth Tobacco Survey among high school students will be employed in examples of analysis using these methods. The benefits of studying odds ratios obtained from logistic regression models for binary response variables are discussed in section three. Loglinear models for counts given in two- and multi-way tables will be examined in section four. Section five includes a fascinating, but relatively underused, method of exploratory data analysis for categorical variables called correspondence analysis. However, before proceeding with analysis, further description of the YTS and examples of documented statistical analysis are given in the next section.

2 YOUTH TOBACCO SURVEYS

2.1 Youth Tobacco Surveys

The CDC's state Youth Tobacco Survey is an instrument whose use is increasing nationwide. The first states to administer the YTS were Florida, Mississippi, and Texas in 1998. The YTS was conducted in thirteen states during the next year. Florida and Mississippi repeated the survey in 1999. Missouri, Nebraska, and South Dakota gave the survey only to children attending middle schools; however, Arkansas, Georgia, Kansas, New Jersey, North Carolina, Oklahoma, Tennessee, and Texas administered the survey both to middle and high school students in their respective states [40]. In the year 2000 twenty-nine states participated in the YTSES. Most states intend to repeat the YTS on either an annual or biennial basis [41]. Clearly, these surveys will generate virtual mountains of data for subsequent analysis.

The 2000 YTS, having a rather broad scope, attempts to measure many of the physical, social, and environmental aspects related to adolescent tobacco use. Thus, the 63 questions in the core questionnaire fall into different categories according to the type of information they evoke. The first five questions involve demographics such as age, gender, grade and ethnicity. Next, a block of 21 questions deals with the student's tobacco related behaviors including lifetime use of cigarettes and other assorted products, daily use, age at first use, intensity of use, and preferences (i.e., brand, menthol). Eight or so questions ask for the student's tobacco related intentions, perceptions, and opinions. The questions addressing whether or not a student will try smoking a cigarette in the next year, does the student think tobacco is ad-

dictive, do young smokers look cool, and is smoking or second-hand smoke harmful, belong to this category. There are categories of questions relating to tobacco exposure in the student's physical environment (smoke in a room or in a car for example), home, community, and among peers. There are also questions that explore pro-tobacco stimuli from actors, athletes, advertising, and apparel, as well as questions about anti-tobacco messages from such sources as doctors, dentists, school, ads, and community programs. Finally, a group of six questions addresses intentions about and attempts at cessation. Each question in the survey represents a variable; that is, a characteristic or quality that may vary from one respondent to the next.

2.2 The 2000 Tennessee Youth Tobacco Survey

The Tennessee Youth Tobacco Survey (TnYTS), conducted first in 1999, was repeated in year 2000. The stated purpose of the survey is to enable surveillance and evaluation of progress made over time by tobacco control programs at state and regional levels. Design of the study deserves further description since the data resulting from the 2000 TnYTS will be used here to illustrate various methods of categorical data analysis. Generally, the 2000 version of the Tennessee Youth Tobacco Survey will be referred to as TnYTS in the discussions that follow.

The Tennessee Department of Health and the CDC's Office on Smoking and Health worked together to develop and validate the 76–item TnYTS, which has 13 more questions compared to the YTS core questionnaire. The survey was administered to 10,779 students attending 119 middle schools and 9959 students from 105 high schools in the spring of 2000 through cooperative efforts by the Tennessee Department of Health,

the Tennessee Department of Education, and the CDC, along with assistance from health organizations and school districts at the local level. The sampling method followed a two-stage cluster design in order to select particular schools (public only) and then the classes of respondents to which the survey would be given. The probability distribution for the selection of schools was proportional to enrollment, and, within each school, second period classes were selected randomly. All students in the selected classes were given the questionnaire. The overall response rate is given to be 72.5% of middle school and 68.9% of high school students surveyed, according to the Tennessee Department of Health. The CDC is given credit for the statistical analysis of the data, weighted to account for nonresponse by both school and student. The analysis was performed on computer using SAS (Statistical Analysis System) to determine point estimates and using SUDAAN (Software for the Statistical Analysis of Correlated Data) for computation of the standard errors of estimates and 95% confidence intervals for the true proportions of a given response, as described in Report 1 of the 2000 TnYTS [38].

2.3 Overview of the Published Analysis of Youth Tobacco Surveys

Publications containing results of state youth tobacco surveys are generally easier to locate on the world-wide-web than in libraries. A large amount of the statistical information gleaned from the YTS, conducted in various states of the union, is located only in documents produced by state health departments for public record. For

instance, see publications from the states of Florida [6], Georgia [10], Kansas [16], Minnesota [21], Mississippi [22], New Jersey [24], Oklahoma [25], South Dakota [33], Tennessee [38], and Texas [39], among others. Occasionally the information is also reported in pamphlets and news-letters, published by public health or educational organizations, like these examples from Arkansas [19], Georgia [9], and South Dakota [34]. A few research articles discussing results of youth tobacco surveys appear in journals such as the Journal of the American Medical Association [2], Minnesota Medicine [28], Preventive Medicine [32], the North Carolina Medical Journal [5], and the American Journal of Drug and Alcohol Abuse [12]. Perhaps more of the latter will be evident in the future as more complex analyses are undertaken for YTS data.

The majority of participating states have adapted the YTS for use, but some states, like Missouri [23] and Massachusetts [32], administer a form of the CDC's Youth Risk Behavior Survey (YRBS). The YRBS is designed to monitor a number of risky adolescent behaviors, one of which involves tobacco use, but the YTS specifically provides measurements on a greater number of variables related to perception and use of tobacco than does the YRBS [40]. Furthermore, most states have sampling strategies and methods of statistical analysis, under guidance of the CDC, similar to those for the TnYTS. This thread of commonality weaves the states together so that valid comparisons can be made at either the state or national level [22]. As a result, the findings reported from one state to the next appear to be more or less the same, depending on the amount of information a particular state chooses to include. The survey results documented by the state of Tennessee are fairly typical of the information included in reports from other participating states, and therefore reports

obtained for Tennessee are used as examples.

Three reports concerning survey data collected in Tennessee are considered. The first to be discussed is Report 1 of the 2000 TnYTS pertaining to tobacco use prevalence [38]. The statistics given in this document primarily consist of conditional distributions for region, gender, grade, and ethnicity, stated in terms of relative frequency, obtained from two-way contingency tables for middle school and high school data (separately). For instance, the report gives weighted percentages of middle school students (or high school students) who report *current tobacco use* (used some tobacco product on at least one of the past 30 days) per region of the state, per gender, per race, and per grade. The same comparison is carried out for *current cigarette use*, *current smokeless tobacco use*, *current cigar/cigarillo use*, and *current bidis use*. Each of these topics is displayed graphically with a bar chart that shows point estimates for the weighted percentages as well as 95% confidence intervals to facilitate visual comparison between categories.

The subject of Report 2 of the 1999 TnYts is students' exposure to environmental tobacco smoke [36]. As always, identical analyses of middle and high school data are carried out separately and then compared in the report. Graphics include pie charts and bar charts. The document gives relative frequencies for a number of variables including the following: students who were in a car or in a room with someone smoking in the past week, students having at least one friend who smokes cigarettes, and students who think that second hand smoke is harmful. The conditional distribution of grade given that a student was exposed tobacco smoke in the last seven days is shown in a bar chart. Prevalence is examined with proportions noted for *lifetime use*

of tobacco (ever tried any tobacco product), *lifetime use of cigarettes* (ever tried a cigarette), *current cigarette use*, and *frequent cigarette use* (smoked on at least 20 of the past 30 days). There is a section of the report addressing parental awareness and acceptance that gives percentages of students responding positively to the following three items. Firstly, those claiming that their parents know they currently smoke cigarettes; secondly, those claiming that their parents approve that they currently smoke cigarettes; and thirdly, those claiming that they currently live with someone that smokes cigarettes. The percentage of high school and middle school students that reported smoking cigarettes on school property in the past month closes the publication.

The third example is Report 3 of the 1999 TnYTS, the main topics of which are the social influences related to adolescents and tobacco [37]. Cigarette brand preference is examined among current smokers and then broken down per ethnicity and gender for the first section. Bar charts enable visual comparison of brand preference among different groups. Section two looks at methods of obtaining cigarettes, refusal of tobacco sale due to age, and whether proof of age was required for purchase of tobacco for current users. Section three gives statistics pertaining to media influences including percentage of current users who own apparel sporting tobacco logos, percentages of current users who would be willing to wear clothing with tobacco logos, and percentage of students (not necessarily current users) who would be willing to wear such items. Also given is the percentage of students who see actors smoking cigarettes on TV or in movies, and the percentage of those who are exposed to TV or Internet that see either actors and athletes using tobacco or advertising for tobacco.

The final sections address friends' behavior and risk perception. The former provides percentages of students with at least one cigarette smoking friend, and percentages of current smokers and of nonsmokers that responded positively that smoking makes teenagers look cool. A line-and-dot plot shows two conditional distributions of grade. One is given that a student is a current smoker claiming that cigarette smokers have more friends and the other is with respect to nonsmokers who think that smokers have more friends. The latter section delivers the proportion of current smokers that think a pack-a-day (or more) cigarette habit may be harmful to their health.

It is obvious from these three reports that the survey contains many variables of interest that, taken two or three at a time, open up many tiny windows with which to view the complex and intertwining issues related to teen tobacco use. The biggest problem with examining only pairwise associations between variables in the YTS is that it ignores the fact that the data set describes a great many variables that do not exist in isolation. Why not use modeling techniques that take advantage of multiple associations and interactions between variables? While this perfunctory analysis may be an appropriate first glance at the survey data, perhaps it should not be the only glance.

Although the previous examples are typical of the published baseline results of youth tobacco surveys around the nation, another mission of the YTSES is to evaluate progress made by state sponsored tobacco control programs. To this end, some states that have conducted the YTS more than once include comparisons between different years. The FYTS results put out by the Florida Department of Health gives such a comparison [6]. Florida administered the FYTS in 1998, 1999, and 2000. Relative

frequencies pertaining to each year are reported and differences discussed for variables describing *current cigarette use*, *current cigar use*, *current smokeless tobacco use*, *tobacco use by grade*, and *tobacco use by region*. Bar chart "triplets" are used to compare the value of statistics obtained over the three years for each variable. It is notable that, in Florida, percentages for most variables declined over the three years of surveillance, a few significantly so. Another nice feature of the report is information about Florida's Tobacco Pilot Program, in operation since 1998. Numerous goals are outlined that are aimed to support an overall effort to decrease lifetime morbidity and mortality related to tobacco exposure. Statistics computed from the YTS provide evaluation data that is used to identify both strengths and weaknesses of the program which will grow and change accordingly. Over time, as elements of state tobacco control programs requiring change are identified, data analysis that takes multiple variables into account has potential to broaden understanding about the myriad issues surrounding tobacco. Whether the purpose is to establish benchmarks or to improve effectiveness of tobacco prevention and control, we suggest that methods of categorical data analysis incorporating statistical modeling and multivariate procedures may help to open relative picture-windows of deeper insight into the world of adolescent tobacco use and exposure.

3 LOGISTIC MODELS AND THE STUDY OF ODDS RATIOS

3.1 Logistic Regression Models

Logistic regression is a mainstay of modern categorical data analysis. Its popularity may rest on the fact that this method is not only easy to use and understand, but provides a way to closely examine the relationship between a single explanatory variable and a binary response variable while controlling for the effects of other explanatory variables. The explanatory variables included in a logistic regression model can be either continuous or categorical (ordinal or nominal) [35]. Logistic models are not uncommon in tobacco related studies. Here are a few examples of tobacco use analyses that include logistic regression.

- Horn et. al. [12] employ logistic regression with 1997 YRBS data to examine 14 risk variables for a comparison of cigarette and smokeless tobacco use in West Virginia.

- Huang, Unger and Rohrbach [13] work with data obtained from the 1995–1996 Independent Evaluation of the California Tobacco Control Prevention and Education Program. The researchers construct logistic regression models to evaluate the relationship between program exposure, perceived usefulness of the program, and susceptibility to tobacco use.

- Rigotti, Lee, and Weschler [27] look at tobacco use among a slightly older population, college students. They analyze data from the Harvard College Alcohol

Study which was conducted at 4-year colleges in 1993, 1997, and 1999, and build separate multiple logistic models to explain cigarette, cigar, smokeless tobacco, and overall tobacco use.

- In a study among students attending multiethnic middle schools, Carvajal et. al. also construct logistic models to explore associations between predictor variables and smoking status, from data measured in a battery of questionnaires. The information gained from the regression is intended to assist in the *design of comprehensive multiethnic interventions by testing the most important factors of initiation and escalation of smoking across various subgroups* [4] (p.255).

There is, therefore, some precedent in use of these procedures in surveillance of tobacco use and evaluation of control programs, yet logistic regression is generally ignored in primary analysis of the YTS. As a result tobacco control programs are being designed and implemented without the beneficial information provided by this type of analysis. Perhaps if its merits were better understood and appreciated logistic modeling would become more commonplace in the analysis of such surveys. Hence, a summary of logistic regression is in order at his time.

The logistic regression model is one member belonging to the family of Generalized Linear Models (GLIM), and is essentially a probability model in which the mean value of the response variable is given as a relation to one or more predictor variables in a regression equation. In the case of logistic regression, the response variable is a Bernoulli variable (or binary); that is, it takes only the values 1 representing *success* or 0 for *failure*, and so the mean response is equal to the proportion of successes. Thus, let π represent the probability of success in the population. Then the link

function for the GLIM is $g(\mu) = logit(\pi) = \log[\pi/(1-\pi)]$; the natural logarithm of the odds of success [18]. The regression equation has the form

$$logit(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

which is linear in the parameters (i.e., the β_i) although nonlinear within the link function per Ramsey and Schafer [26]. For the inverse, let $logit(\pi) = \alpha$. Then the probability of success can be estimated as

$$\hat{\pi} = e^{\hat{\alpha}}/(1 + e^{\hat{\alpha}}).$$

For the purposes of categorical data analysis, it is useful to make comparisons of the response variable for different levels of an explanatory variable. The odds and the odds ratio are the tools of choice in this regard. The odds that a particular event occurs are defined as $\pi/(1-\pi)$, where π is the probability of the event. Conveniently, the odds can be estimated by exponentiation of the logit. Thus, the predicted odds that the binary response has value equal to one (i.e., the odds of success) is given by

$$\hat{\pi}/(1-\hat{\pi}) = e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p}.$$

The odds ratio, on the other hand, is a ratio formed by the odds that an event occurs for two different groups. Let ω represent the odds of success. From the logistic model, the predicted odds ratio pertaining to the response variable for $x_n = K$ versus $x_n = L$ (fixing all other x's in the model) is estimated as

$$\hat{\omega}_K/\hat{\omega}_L = e^{[\hat{\beta}_n(K-L)]}.$$

It follows that $e^{[\hat{\beta}_n]}$ is the estimated odds ratio when the explanatory variable x_n increases incrementally (in the case of dummy variables, from one category to the next). Notice that the predictors may be continuous, discrete, or dummy (nominal) variables, making logistic regression paritcularly suited to analysis of large surveys such as the YTS.

Logistic regression is most easily carried out with the help of a computer and a statistical software package. We have chosen to employ SAS [30] for construction of logistic models using the 2000 TnYTS data. The SAS procedure *Proc Logistic* is designed for this purpose. SAS estimates parameters for logistic models by the method of maximum likelihood. An internal test of the null hypothesis that a parameter has value equal to zero yields a χ^2 statistic and corresponding p-value along with the output for each parameter estimate in the model. Our preferred measure of model fit in SAS output is the percent of concordant pairs. SAS computes this measure by examining every pair of observations, students in this case, where one member of the pair is a *success* and the other member is a *failure* with regard to the response variable. A pair is *concordant* if the student having greater probability of success according to the model is actually a success per observation of the response variable. Thus, percent concordance refers to the percent of all such pairs of students that are concordant with the model. These are the measures that will be referenced in the discussion of logistic regression models for current cigarette use that follow.

3.2 Logistic Models to Explain Current Cigarette Use

The YTS was conducted in the state of North Carolina in 1999 [5]. Some results of that study were reported as follows: a significant difference was noted in current tobacco use, including smoked and smokeless types, between males (44%) and females (32%). Current tobacco use increased with grade also. It was observed that 35% of ninth graders reported current tobacco use while 45% of twelfth graders did so. In addition, 43% of white students were current tobacco users, a significantly higher percentage compared to the 29% of black students that reported current tobacco use. As in the published analysis of the TnYTS, the data is weighted and SUDAAN employed to give 95% confidence intervals for the true proportions of a given response. Statistical significance of differences between subgroups is assumed to be present if there is no overlap of corresponding confidence intervals.

A comparable analysis may also be performed on the year 2000 data collected in Tennessee, using students' answers to the question concerning smoking cigarettes in the past 30 days as an indicator of current cigarette use. Students are considered current smokers if they report smoking on any of the past thirty days. A two-way contingency table for current smoker versus gender shows that nearly 36% of male respondents reported having smoked cigarettes in the past 30 days, but only 33% of females claimed the same. A chi-square test of independence indicates the existence of a significant association between current smoking and gender (p = 0.0222). Of ninth graders, 30% admitted to smoking in the past 30 days whereas 39% of twelfth graders were current smokers. The association between grade and current smoking is significant per the χ^2 test of independence ($p < 0.0001$). Along ethnic lines, 36%

of white students reported smoking in the past month while only 27% of other ethnicities did so. Once again a significant association is found ($p < 0.0001$). Note, we are working with non-weighted data for modeling purposes. SUDAAN was used to calculate the following statistics from weighted data, as reported in the 2000 TnYTS [38]. Students reporting current cigarette use: 33.4% of males, 31.3% of females, 27% of 9th graders, 39.3% of 12th graders, and 35.1% of Caucasians.

While this type of analysis offers a quick initial look at the isolated pairwise association between current cigarette use and gender, grade, or ethnicity, logistic regression will allow an exploration of current smoking as it relates to either gender, grade, or ethnicity while controlling for the effects of the other two variables. Extensive recoding of the data was necessary to enable modeling of the variables measured. For instance, binary variables were recoded and renamed so that the variable name specified indicated that $1 = yes$ and $0 = no$. A good example is the variable named *Male*, for which a response of yes corresponds to a measured value of one and no (a female) equals zero. Data recoding, as well as modeling, is accomplished easily using SAS [30]. The SAS programming code used to recode variables is displayed in appendix A.1. Specifically for the year 2000 Tennessee high school data, a logistic model is obtained for current cigarette use as explained by gender, grade , and ethnicity using the statistical software SAS and SAS code in appendix A.2. That model is

$$\widehat{Logit}(\pi) = -2.509 + 0.0905\,Male + 0.1386\,Grade + 0.4554\,White, \qquad (1)$$

where π represents the probability that a random student claims to be a current smoker. For a comparison of white males in the twelfth grade to white males in the

ninth grade, the estimated odds ratio is given by

$$e^{[0.1386(12-9)]} = 1.5156.$$

So, the odds that a twelfth grade white male is a current smoker are nearly 1.5 times that of a ninth grader with the same gender and ethnicity. Comparing a male (Male = 1) with a female (Male = 0) of the same grade and ethnicity, the odds ratio is estimated to be

$$e^{[0.0905(1-0)]} = 1.0947;$$

that is, the odds of the male being a current smoker are almost 1.1 times that of the female student. Finally, in a comparison of ethnicity, the estimated odds ratio given by the model stated above is

$$e^{[0.4554(1-0)]} = 1.5768.$$

This indicates that for fixed gender and grade, the white student is about 1.6 times more likely to be a current smoker than a student who reports some other ethnicity.

So far we have examined the odds ratios for current smoking for different categories of the same explanatory variable. However, the model also allows a broader comparison. For example, the model can be used to compare the odds of current smoking for a white male in the senior class to that of a female freshman of some other ethnicity. For the first group we substitute values of 1, 12, and 1, and for the second group substitute 0, 9, and 0, into the model for the explanatory variables in order to calculate the odds for each group. Accordingly, the odds of current smoking for a white male senior is 0.7409 and the odds of current smoking for a female

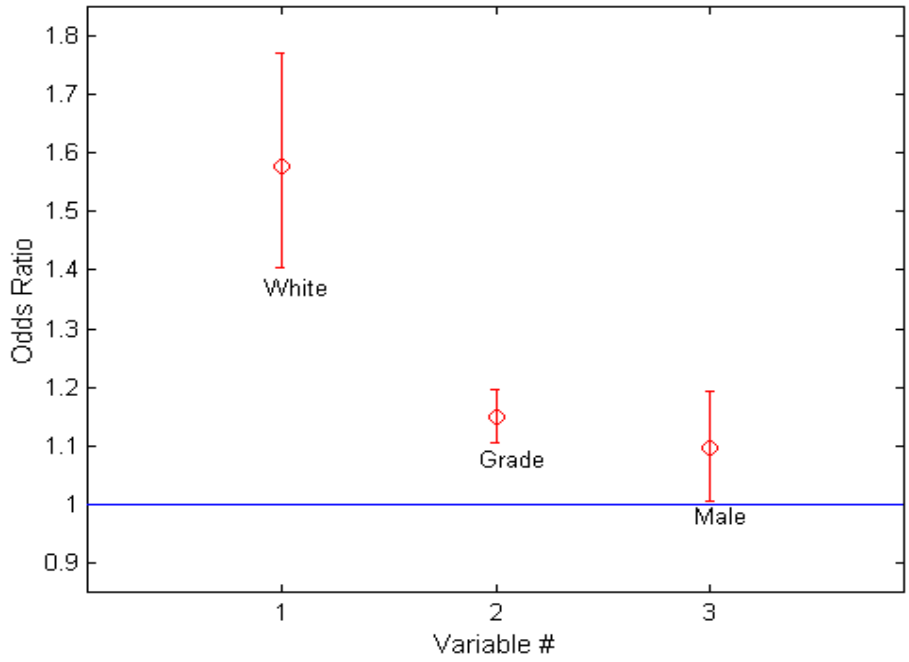

Figure 1: Odds ratio plot for current smoking per model (1)

freshman of nonwhite ethnicity is 0.2832. Taking the ratio of these odds yields 2.6; that is, a white male in the 12th grade is about 2.6 times more likely to be a current smoker than is a ninth grade nonwhite female. Although these statistics provide a bit more insight into adolescent smoking behavior than do the two-way tables, these three variables, gender, grade, and ethnicity, may not be the most meaningful predictors of current smoking. As a measure of association between predicted probabilities and observed responses, model (1) provides only 52.3% concordant pairs. Other variables in the survey may better explain the odds of being a current smoker.

In order to provide a more informative model to explain the probability that a Tennessee youth reports smoking a cigarette in the past month, the following variables are included in a logistic model:

- **Male**: Gender. 1 = male, 0 = female.
- **Grade**: Grade in school. 9 = freshman, 10 = sophomore, 11 = junior, 12 = senior.
- **Firstage**: Age at which student first smoked a whole cigarette. 8 = at most 8 yrs, 9 = 9 or 10 yrs, 11 = 11 or 12 yrs, 13 = 13 or 14 yrs, 15 = 15 or 16 yrs, 17 = at least 17 yrs of age.
- **Safe**: Does student think it is safe for a person to smoke for only 1 to 2 years? 1 = yes, 0 = no.
- **Car**: Has student ridden in a car with cigarette smoke in the past week (ipw)? 1 = yes, 0 = no.
- **Approve**: Do parents or guardians approve of the student smoking? 1 = yes, 0 = no.
- **Loose**: Does student have knowledge of places selling loose cigarettes in their community? 1 = yes, 0 = no.

Using these explanatory variables to predict the probability that a student reports current smoking gives rise to the model,

$$\begin{aligned}\widehat{Logit}(\pi) &= -0.8344 + 0.0126\,Male + 0.2246\,Grade \\ &\quad - 0.1236\,Firstage + 0.6057\,Safe + 1.6327\,Car \\ &\quad + 1.0777\,Approve - 0.7149\,Loose. \end{aligned} \tag{2}$$

This model yields 76.2% concordance, which is an improvement over model (1). Furthermore, gender is no longer a significant predictor when these new variables are taken into consideration ($p = 0.9258$).

There are other noteworthy features of model (2). Controlling for all other variables in the model, the following items come to light:

- A student who first smoked at age 9 or 10 years of age is approximately three times as likely to have smoked in the past month than one who first smoked a cigarette at age 17 years or older.

- The odds that a student is a current smoker are more than 5 times greater for those who have recently ridden in a car with cigarette smoke compared to those who have not.

- Surprisingly, the knowledge of whether loose cigarettes are for sale in a student's environment appears to have a protective effect. A student who knows where to buy loose cigarettes has half the odds of being a current smoker compared to one who does not.

- Approval by parents of their student's smoking behavior is significantly associated with current smoking ($p < 0.0001$), with the odds of current smoking almost three times greater when parents approve of it.

Yet, although the concordance rate is high, model (2) has serious defects. It is a good illustration that caution and common sense are also of great value in the appraisal of logistic regression models.

Firstly, the negative association between smoking cigarettes in the past month and knowledge of "any places that cell single or loose cigarettes" in the area of a student's school or residence is unexpected. For one thing, it might be argued that in some cases people, and not places, sell loose cigarettes. Further, two primary groups comprise the body of students who claim not to have smoked in the past 30 days - those who smoked before but not in the past 30 days, and those who have never smoked. Of the 9257 students responding to the question concerning loose cigarettes, 3183 claimed to have smoked within the past 30 days, 3260 indicated that they had smoked previously but not in the past 30 days, and 2814 students maintained that they never smoked cigarettes. It is doubtful that students who have never smoked would be cognizant of places where they could buy loose cigarettes, yet 68% of students that reported having knowledge of such places said they were not current smokers. While 22% of students who did not smoke in the past month reported knowledge of such places, only 20% of current smokers did the same. However, a more thorough examination reveals that 23% of lifetime smokers (current or previous smokers) reported knowing of places selling loose cigarettes, but only 18% of never smokers did so. Thus, due to the binary nature of the response variable in the proposed model, the loose cigarette variable turns out to be an unsuitable predictor of current smoking behavior.

The second flaw in the model is caused by the variable concerning parents' approval. A closer look at the question and its possible answers is warranted. The question is, "Do your parents (or guardians) approve of your smoking?" The possible responses are as follows:

1. I do not smoke cigarettes.

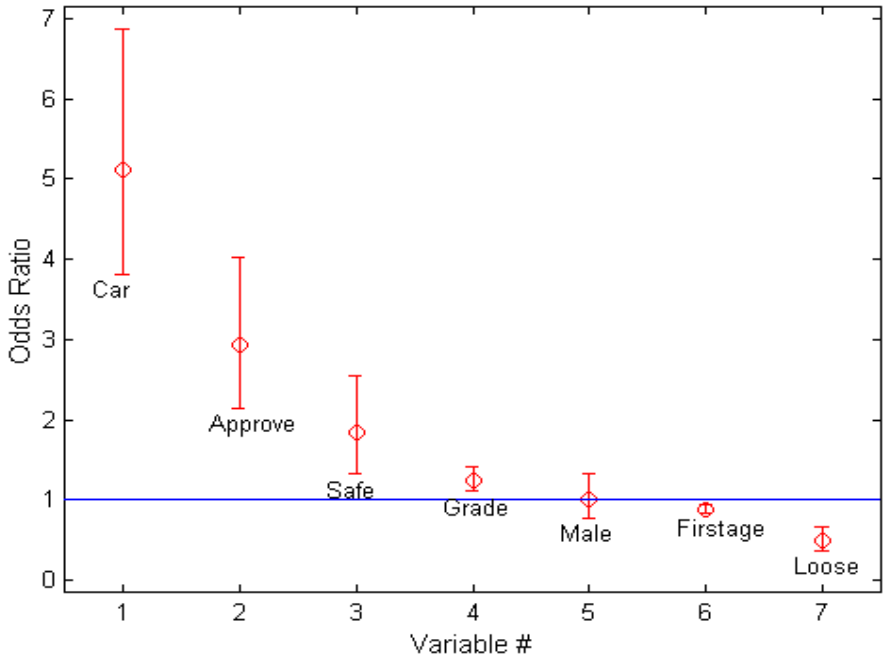

Figure 2: Odds ratio plot for current smoking per model (2)

2. Yes.

3. No.

4. My parents (or guardians) do not know that I smoke.

In the recoding that allows the use of this variable in the logistic model, all of those students who report not smoking cigarettes are treated as missing. It is worth mentioning at this point that one of the disadvantages of using logistic regression in the analysis of survey data is that introduction of more predictors into a model also tends to increase the number of missing observations. However, some explanatory variables exacerbate the problem more than others do. About two-thirds of the respondents do not currently smoke, so this single variable directly results in a large proportion of missing observations. Indeed, the proposed model is based on only 20% of the

students sampled. Unfortunately, parents' approval or disapproval of their children's smoking may actually be a worthy explanatory variable, but in this particular survey there is no way to tell how many students do not smoke because they think their parents would not approve.

There is a valuable lesson to be learned from this particular model. In logistic regression models it is not surprising to see concordance increase with the number of explanatory variables. Analysts must think carefully about the selection of variables for parsimonious models that also take advantage of, and make sense of, the wealth of information contained in huge sets of data such as those collected from statewide Youth Tobacco Surveys. Thus, although the concordance is high, model (2) is deemed unsuitable. The next goal is to construct a better model.

Once again, in an effort to provide a more plausible model that will explain the probability that a Tennessee youth reports smoking a cigarette in the past month, some variables from model (2) are eliminated and others are examined in their place. **Male**, **Grade**, **Firstage**, **Car**, and **Safe**, from the previous list of predictors, will remain in the model. The following variables are added to the newest model:

- **White**: Ethnicity. 1 = white, 0 = other.
- **Harm**: Do young people risk harming themselves if they smoke from 1 to 5 cigarettes per day? 1 = yes, 0 = no.
- **Cool**: Does the act of smoking cigarettes make young people look cool or fit in? 1 = yes, 0 = no.
- **Friends**: How many of student's four closest friends smoke cigarettes? 0 =

none, 1 = one, 2 = two, 3 = three, 4 = four.

- **Home**: Does anyone living with student currently smoke cigarettes? 1 = yes, 0 = no.

The new model, which contains ten predictor variables and estimates 11 parameters, is defined as follows:

$$\begin{aligned}\widehat{Logit}(\pi) = &-1.9768 - 0.818\,Male + 0.0854\,Grade \\ &+ 0.1364\,White - 0.0383\,Firstage + 0.9044\,Car \\ &+ 0.8356\,Safe - 0.4177\,Harm + 0.5251\,Cool \\ &+ 0.6253\,Friends + 0.1080\,Home. \end{aligned} \quad (3)$$

The percentage of pairs that are concordant is 79.2, and this model describes 43% of the total observations. More variables are now competing for a share of the explained error in the prediction of current smoking behavior. Two variables that were critical in model (1), Male and White, do not explain a sufficient amount of error when other more relevant variables are taken into consideration, and so Male (p = 0.2646) and White (p = 0.1818) are not significant predictors in model (3). Firstage is not now as significant a predictor as in model (2) discussed above (p-value has increased from less than 0.0001 to 0.0219). Home, included as a substitute for the faulty variable Approve, is not found to be significant in model (3) (p = 0.1514). However, if variable Car is removed from the model, Home becomes a significant predictor. This is a noteworthy point since the most likely sources of cigarette smoke in a car are family and friends. Still, Car has merit as an explanatory variable specifically because the smell of smoke

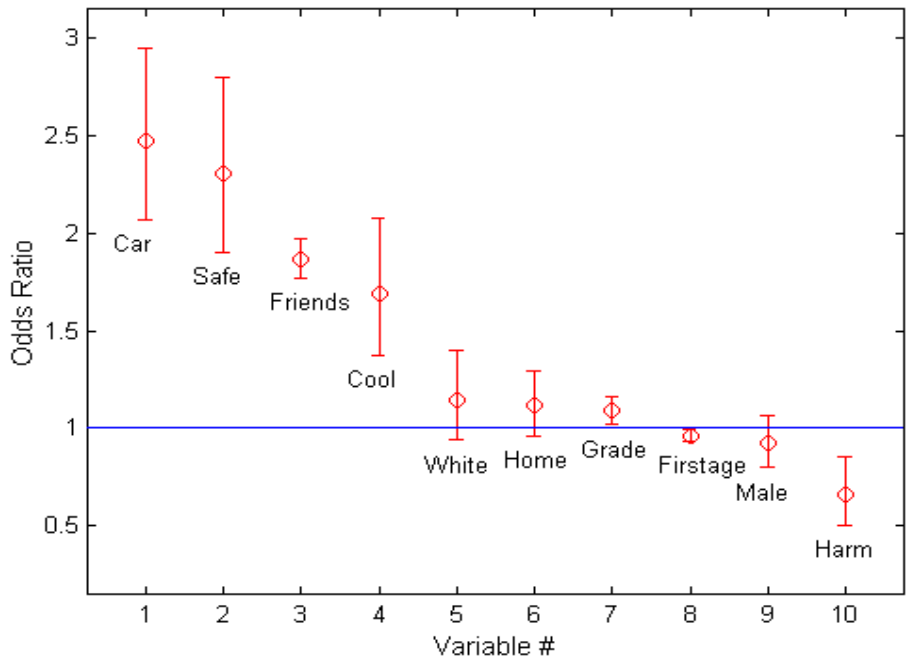

Figure 3: Odds ratio plot for current smoking per model (3)

from enclosed spaces tends to be absorbed by hair and clothing and, thus, is a good indicator of recent exposure to tobacco smoke. All associations appear to be going in the expected directions.

Most of the explanatory variables included in model (3) are positively associated with current smoking behavior. According to this model, a senior is approximately 1.3 times as likely to be a current smoker compared to a freshman. A student whose four closest friends smoke cigarettes is over twelve times as likely to be a current smoker as one whose four closest friends do not smoke. The student exposed to cigarette smoke in a car in the past week has two and one-half times the odds of reporting current smoking than one who has not been exposed. A student who believes it is safe to smoke cigarettes for just a couple of years has two and one-third times the odds of

claiming to be a current smoker than one who does not think smoking is safe. And, the student who thinks young cigarette smokers look cool is 1.7 times as likely to be a current smoker than one who does not think that smoking makes teenagers look cool. Harm is the one variable that has a negative association with regard to current smoking. Students who think that smoking one or more cigarettes daily may be a health risk are two-thirds as likely to report smoking cigarettes in the past month as compared to others. Recall that for each one of these comparisons in turn, one must assume that all other variables are held constant.

The examples of logistic models in the preceeding discussion are intended to show the informative nature of odds ratios produced by those models. It is common to display the point estimates of odds ratios and their confidence intervals in tables, but comparisons of incremental odds ratios of the response variable among the different explanatory variables might be better facilitated by graphical displays. To that end, odds ratio plots, inspired by Friendly [7] and Gelman, Pasarica, and Dodhia [8] have been devised for the three logistic regression models given above. Matlab [17] programs were developed specifically to produce these plots given the point estimates, upper margins of error, and lower margins of error (the confidence intervals are not necessarily symmetric) obtained from the SAS [30] output of Proc Logistic, and are included in appendix A.3. In general, the odds ratio plots shown here are designed to give the point estimates (in descending order) along with 95% confidence intervals, provided by SAS, for the incremental odds ratios associated with current smoking corresponding to the explanatory variables included in the model. Figure 1 displays the point estimates and 95% confidence intervals of odds ratios for current smoking

that pertain to gender, grade, and ethnicity, the explanatory variables included in model 1. Likewise, figures 2 and 3 correspond to models 2 and 3, respectively. The plots include a reference line, representing an odds ratio equal to one, to aid visual identification of odds ratios that have value significantly different from one. The 95% confidence intervals are displayed as *whiskers*, a term we borrow from the common boxplot, extending vertically up and down from the odds ratio point estimate. If the confidence interval for a particular predictor variable straddles the reference line, then the odds ratio of current smoking pertaining to incremental levels (i.e., such as an increment from category 1 to category 2) of that variable is not considered significantly different from one.

3.3 Usefulness of Logistic Regression in the Context of Youth Tobacco Surveys

The advantages of logistic modeling should be fairly clear at this point. A person wishing to design tobacco prevention and control programs stands to gain a great deal of relevant information by examining logistic regression models. The models discussed in this section allow clearer insight to identify which factors have relatively greater power to predict current cigarette use. Since logistic regression provides the capability to measure the impact of a single explanatory variable, through odds ratios, while controlling for the effects of others, policymakers may be better able to determine if individual elements of anti-tobacco programs are effective and to suggest appropriate changes or improvements. Furthermore, there are a number of binary

response variables included in the YTS that could be modeled in this manner. This list includes such variables as lifetime cigarette use, lifetime tobacco use, frequent cigarette use, attempts to quit, and desire to quit smoking, to name just a few.

4 LOGLINEAR MODELS AND INDEPENDENCE

4.1 Loglinear Models

State Youth Tobacco Surveys measure responses pertaining to more than 60 variables that describe a variety of issues related to tobacco exposure. However, as noted before, a majority of the statistics come from pairwise examination of variables. This practice probably seems reasonably safe but, as Agresti warns, *incorrect conclusions can result from studying variables two at a time* [1] (p.130). The faulty conclusions can occur as a result of, what are collectively termed as, *association reversal phenomena* [29] (p.81). The most famous of these is known as Simpson's Paradox, occurring when the association between two variables either disappears or reverses direction when covariates are taken into account. Simpson himself gives an example of each case [31]. This paradox may arise when some third variable has conditionally dependent relationships with both variables in the pair being studied [3]. The fact that the YTS contains so many questions implies that there is a realization that many factors have an impact on teen tobacco use and exposure, and that the effects of lurking variables must be controlled. When analysts investigate the association between just two variables from such a study, all of these covariates are still lurking since they are essentially ignored. Loglinear modeling allows an examination of dependencies between two or more variables and, thus, provides a means of control over lurking variables. The discussion of loglinear modeling that follows will begin with an example using two binary variables and then extend to the multivariate procedures.

Loglinear modeling of the cell frequencies in a two-way table for variables X_1 and

X_2 is similar to the chi-square test of independence between the variables. This procedure will produce regression-type models for the expected counts in the table, but, more importantly, it allows one to perform a test of significance for the interaction between the variables. If the association between two variables is statistically significant, the model for predicted counts will contain a nonzero interaction term. For this test of independence the null hypothesis states that the interaction term in the model is equal to zero, and the test is accomplished by comparing a saturated model that includes the interaction term to a reduced (or independence) model that does not. Let us denote the expected count in the $(i,j)th$ cell as m_{ij}. The saturated model is written as

$$log(m_{ij}) = \mu + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_{ij}^{X_1 X_2}$$

and the independence model is

$$log(m_{ij}) = \mu + \lambda_i^{X_1} + \lambda_j^{X_2},$$

where log is taken to be the natural logarithm function (i.e., ln). Agresti [1] outlines methods by which the parameters for either model can be determined given a two-way table of observed frequencies. For the independence model,

$$\lambda_i^{X_1} = log(p_{i+}) - (\sum_h log(p_{h+}))/I,$$

$$\lambda_j^{X_2} = log(p_{+j}) - (\sum_h log(p_{+h}))/J,$$

and

$$\mu = log(n) + (\sum_h log(p_{h+}))/I + \sum_h log(p_{+h}))/J,$$

where p_{i+} and p_{+j} are the marginal relative frequencies for the I (total number of) rows and the J columns, respectively. Now, define $\eta_{ij} = log(n_{ij})$, where n_{ij} is the observed frequency in the $(i,j)^{th}$ cell of the table, and let $\eta_{i.} = \sum_j \eta_{ij}/J$, $\eta_{.j} = \sum_i \eta_{ij}/I$, and $\mu = \eta_{..} = \sum_i \sum_j \eta_{ij}/IJ$, the grand mean obtained by averaging the natural logarithm of all cell counts. Then the other parameters in the saturated model are given by

$$\lambda_i^{X_1} = \eta_{i.} - \eta_{..},\ \lambda_j^{X_2} = \eta_{.j} - \eta_{..},$$

and

$$\lambda_{ij}^{X_1X_2} = \eta_{ij} - \eta_{i.} - \eta_{.j} + \eta_{..}.$$

In the case of a 2×2 table the saturated model has 9 parameters once all the i's and j's are taken into account. However, the following constraints on the parameters,

$$\sum_{i=1}^{2} \lambda_i^{X_1} = 0,\ \sum_{j=1}^{2} \lambda_j^{X_2} = 0, \text{ and } \sum_{i=1}^{2} \lambda_{ij}^{X_1X_2} = \sum_{j=1}^{2} \lambda_{ij}^{X_1X_2} = 0,$$

effectively reduce the number of $\lambda's$ to just three, $\lambda_1^{X_1}$, $\lambda_1^{X_2}$, and $\lambda_{11}^{X_1X_2}$, that are nonredundant. The fourth parameter in the saturated model is μ, which is fixed by the total number of observations n represented in the table. The first two constraints also hold for the independence model, giving it three parameters, μ, $\lambda_1^{X_1}$, and $\lambda_1^{X_2}$ [35]. Ramsey and Schafer [26] suggest that an extra-sum-of-squared-residuals test, which compares residuals obtained from both models, is appropriate to test the significance of the interaction term. For a 2×2 table the residuals for the saturated model are all zero, so the test statistic, which has a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom, is calculated by summing the Pearson residuals from the independence model. The test statistic should agree perfectly with a chi-square test of independence

LIFE		HOME	
		No	Yes
		(j = 1)	(j = 2)
No	(i = 1)	1771	1000
Yes	(i = 2)	2798	3592

Table 1: Lifetime cigarette use versus someone at home smokes (LIFE vs HOME)

between the two variables. At this point an example should help to clarify this discussion.

In order to illustrate loglinear modeling of the counts in a 2×2 table we will use two binary variables measured by the 2000 TnYTS. The row variable is LIFE, short for lifetime use of cigarettes. A response of *yes* represents a student who reports having ever smoked a cigarette (even one or two puffs) and *no* represents a student claiming to have never smoked a cigarette. The column variable is HOME, which indicates whether or not someone living at the student's home smokes cigarettes. The 2×2 contingency table of observed frequencies for LIFE and HOME is shown in table 1.

Matlab [17] programs were designed to perform the computations recommended by Agresti [1], and are located in appendix A.4. The first of these gives the parameters for the saturated loglinear model for expected counts in the two-way table of LIFE and HOME, which is

$$log(m_{ij}) = 7.6275 - 0.4340\, X_1 + 0.0804\, X_2 + 0.2053\, X_1\, X_2. \tag{4}$$

Indicator variable X_1 represents LIFE with response $no = 1$ and $yes = -1$, while

the similarly valued X_2 represents the column variable HOME. Notice that we are modeling the expected cell counts in the two-way table with both of the variables treated as predictors. The independence model is

$$log(m_{ij}) = 7.6516 - 0.4178\,X_1 + 0.0025\,X_2, \tag{5}$$

which yields expected counts of $m_{11} = 1382$, $m_{12} = 1389$, $m_{21} = 3187$, and $m_{22} = 3203$. The Pearson residuals from this model are 109.48, 108.93, 47.48, and 47.24, respectively, and sum to 313.13 which has exactly the same value as the χ^2 test statistic, on one degree of freedom, computed for a test of independence between LIFE and HOME. The conclusion for the χ^2 test is that LIFE and HOME are not independent variables. This parallels the conclusion when testing for lack of fit since the interaction parameter in the saturated model has value significantly different from zero. In other words, there is significant interaction between the variables LIFE and HOME. Furthermore, interpretations for the parameters of small models can be explained fairly easily in terms of odds ratios [1]. The odds ratio can be calculated easily from the saturated model for a 2×2 table. Let the odds ratio be denoted as θ. Then

$$log(\theta) = log(m_{11}m_{22}/m_{12}m_{21}) = \lambda_{11}^{X_1X_2} + \lambda_{22}^{X_1X_2} - \lambda_{12}^{X_1X_2} - \lambda_{21}^{X_1X_2} = 4\lambda_{11}^{X_1X_2}.$$

This outcome is a result of the sum-to-zero constraints on the parameters. Simply put, the odds ratio pertaining to a 2×2 table is equal to the antilog of four times the interaction parameter in the full model. For the contingency table of LIFE and HOME the odds ratio $\theta = e^{4*0.2053} = 2.3$. Thus, a student is approximately twice as likely to report having ever smoked a cigarette if someone at home smokes.

The variables LIFE and HOME provide a nice example of a significant bivariate association determined through loglinear modeling of observed counts in a 2×2 table. However, loglinear models for two-way tables may seem like overkill. After all, that familiar old friend, the chi-square test of independence does an acceptable job of indicating lack of independence between two categorical variables. The biggest advantage of loglinear modeling may be that it is also a multivariate technique capable of showing multi-way interactions between categorical variables. Another advantage is that no response variable is needed, or we can take the perspective that all variables are measured simultaneously, and so they are all response variables. This seems ideal for large surveys such as the YTS. The focus of the modeling process in the loglinear case is not on prediction, but instead on *assessing patterns of statistical dependence among subsets of variables* [35] (p.427).

Loglinear modeling seems made to order for surveys that measure many categorical variables. Yet it is not a process which is customarily done with pencil and paper. Iterative processes are often required to obtain maximum likelihood estimates for model parameters, and so one must turn to fairly sophisticated computational methods. The SAS System includes a special procedure, *Proc Catmod*, that accomplishes the analysis quickly. The output of Proc Catmod may include the parameter estimates if desired, but this can be suppressed, and usually is, by a *noparm* command. The output that is most useful is titled Maximum Likelihood Analysis of Variance, and includes a χ^2 statistic and associated p-value that indicate the significance, or lack thereof, of each parameter in the model. If the value of a parameter is not significantly different from zero, then the variable or interaction to which that

parameter belongs is not significant in the model. We are generally most interested in the presence of interactions between the variables included in the model. Certainly there are many combinations of variables measured in the YTS, and many models, that might be examined for interactions through loglinear modeling. That means that the technique may be used in an exploratory context. In order to minimize type I errors, in this case, we suggest that p-values should be very small before considering an interaction significant. The next logical step is to look at an example of loglinear modeling for a multi-way table.

4.2 Loglinear Models for Multi-way Tables from the 2000 TnYTS

It has already been noted that there may be interesting associations between variables Car, Home, and Friends. Recall that Car refers to the binary variable measuring whether or not a student has ridden in a car with someone smoking a cigarette in the past week. Home represents the binary variable corresponding to the presence or absence of a cigarette smoker in the student's home. The discrete variable Friends is the number of the student's four closest friends that smoke cigarettes. The SAS code used to obtain a loglinear model using these three variables is located in appendix A.5. Not surprisingly, all two-way interactions between the variables are significant. There is also a significant three-way interaction ($p < 0.0001$) between Car, Home, and Friends. In order to investigate the presence of some association reversal phenomenon resulting from this interaction we must look at two-way tables. The relevant

Car		Friends				
		0	1	2	3	4
No	(Car = 0)	2032	624	400	136	140
Yes	(Car = 1)	1215	1022	1080	807	1174

Table 2: Exposure to car smoke in the past week versus number of friends that smoke (Car vs Friends)

Car		Friends				
		0	1	2	3	4
No	(Car = 0)	1559	462	281	92	88
Yes	(Car = 1)	353	383	399	295	348

Table 3: Car versus Friends for students from homes where nobody smokes

tables include Car versus Friends regardless of Home (table 2), Car versus Friends for students from homes where nobody smokes (table 3), and Car versus Friends for students from homes where somebody smokes (table 4).

Let us compare the conditional distributions of Car given that Friends equals zero for the three tables; that is, divide the counts in the first column of cells by the column total and compare the resulting proportions. The measure of association known as *Gamma* is based on the number of concordant pairs and will help to compare the

Car		Friends				
		0	1	2	3	4
No	(Car = 0)	455	149	109	42	47
Yes	(Car = 1)	829	604	624	489	804

Table 4: Car versus Friends for students from homes where somebody smokes

degree of association between Car and Friends in the three tables. Table 3 shows that 81.5% of students having no close friends that smoke responded no to Car versus 18.5% that responded yes to Car, and the degree association between Car and Friends is given by $Gamma = 0.6586$. Table 2 shows that 62.6% of students having no close friends that smoke responded no to Car versus 37.4% that responded yes to Car, and $Gamma = 0.6125$. Meanwhile, table 4 shows that 35.44% of students having no close smoking friends answered no to Car versus 64.6% that said yes to Car, with $Gamma = 0.5358$. To summarize, of students with no smoking friends and nobody smoking at home, a large majority reported no exposure to car smoke in the past week. On the other hand, of students who come from homes where somebody smokes cigarettes but who have no close smoking friends, a majority claimed to have been recently exposed to car smoke. The main differences between the three tables occurs in the first column, yet the conditional distributions of Friends with respect to the given response for Car are quite similar. There is no evidence of Simpson's Paradox in this case, just a difference in the degree of association between Car and Friends when we account for Home. So, how does all of this relate to the big picture? Well, Car might be the best explanatory variable of the three in a logistic model for some response variable having to do with smoking status such as lifetime, current, or frequent cigarette use because it reflects both Friends and Home, because exposure to car smoke is noticeable, and because, according to this writer's personal experience, it can lead directly to experimentation with cigarettes. Alternatively, perhaps the second order and even third order interaction terms may be suitable in a logistic regression model.

4.3 Usefulness of Loglinear Models in the Context of Youth Tobacco Surveys

Loglinear modeling is a useful tool for analysis of the YTS not only because it is appropriate for use with categorical data but also because it can help to identify and make use of interactions that are known to exist in the myriad and complex factors that relate to adolescent tobacco use. It is a technique that assists in the search for the association reversal phenomena that may lead to mistaken conclusions based on two-way tables. Furthermore, interactions that are determined to exist between variables may also be of use in logistic regression models or in correspondence analysis, which is to be discussed in the next section.

5 CORRESPONDENCE ANALYSIS

5.1 Correspondence Analysis for Two-Way Contingency Tables

Correspondence Analysis (CA) refers to a technique for exploratory data analysis that graphically indicates the degree of association between discretely measured variables. CA is attributed to the pioneering work of H. O. Hirschfeld in 1935 and of R.A. Fisher in 1940 by Hill [11] (p.340) who describes CA as, *an analogue of principal components analysis* for the multivariate analysis of discrete data. The analysis of principal components is a data reduction technique which is popular in multivariate analysis of continuous data. The method employs all of the variables in a large collection of data to create a smaller set of new variables that can be used to approximate the larger set [14]. CA is likened to principal components analysis because singular value decomposition of residuals from a model that assumes independence between variables is employed to explain the largest proportion of the value of the chi-square test statistic in a small number of dimensions [7]. A plot is created using the two dimensions that best describe the data. Friendly [7] makes a distinction between simple correspondence analysis (SCA) and multiple correspondence analysis (MCA). SCA is applied only to bivariate data for which frequencies of responses are displayed in two-way, $r \times c$ contingency tables, while MCA is a method with similar results, suitable for plotting the relationships between the levels of two or more discrete variables. Ordinal, categorical variables, and binary response variables, such as those measured in the youth tobacco surveys, can often be coded as discrete data and,

thus, CA has great potential to enable better recognition of the relationships existing between variables in such large datasets.

Johnson and Wichern [15] present a well-designed algorithm in linear algebra, that performs SCA for a two-way table of counts. Using this algorithm as a guide, a Matlab [17] program, located in appendix A.6, that performs SCA was developed in order to illustrate the technique for two variables measured in the TnYTS. The first variable, called DAYS, represents the number of days that a student reports smoking cigarettes out of the past 30. The possible responses are as follows: zero days, one to two days, three to five days, six to nine days, ten to 19 days, 20 to 29 days, and all 30 days. Let DAYS be the row variable in a two-way table. The second variable is FRIENDS, with response categories 0, 1, 2, 3, and 4, corresponding directly to the number of the student's four closest friends that smoke cigarettes. Let FRIENDS be the column variable for the table. Since DAYS has seven categories and FRIENDS has five, cross-tabulation of the non-missing responses for the survey results in a 7×5 contingency table of frequencies and is shown in table 5.

The first step in CA of an $r \times c$ table, where r refers to the number of rows and c is the number of columns, is to place the count belonging to each cell of the table into an $I \times J$ matrix $\mathbf{N}$ of which the (i,j)th entry is the count in the corresponding (i,j)th cell of the table for row index $i = 1, 2, \ldots, r = I$ and column index $j = 1, 2, \ldots, c = J$. For DAYS and FRIENDS, $\mathbf{N}$ is a 7×5 matrix with entries n_{ij} such that $n_{11} = 2909, n_{12} = 1193, \ldots, n_{75} = 537$. The analysis of this matrix involves computation of relative frequencies, centering and scaling of the relative frequencies, and matrix singular value decomposition (SVD) which results in numerical coordinates in p dimensions

DAYS		FRIENDS				
		0	1	2	3	4
		(j = 1)	(j = 2)	(j = 3)	(j = 4)	(j = 5)
0	(i = 1)	2909	1193	802	348	314
1-2	(i = 2)	77	110	151	71	96
3-5	(i = 3)	33	64	80	61	38
6-9	(i = 4)	33	38	60	42	49
10-19	(i = 5)	27	52	71	62	108
20-29	(i = 6)	20	41	89	89	134
all 30	(i = 7)	88	102	184	238	537

Table 5: Number of days smoked versus number of best friends who smoke (DAYS vs FRIENDS)

for row points and column points representing the categories of the two variables. The number of dimensions p is the minimum of $(r-1)$ or $(c-1)$. Thus, for this example we obtain coordinates in four dimensions for the row points representing the seven levels of DAYS and for the column points representing the 5 levels of FRIENDS. Finally, a plot will display the row points and column points according to the coordinate pairs given by the first two dimensions. The procedure provides a nice exercise in matrix manipulation and algebra.

Let the sum of entries in the ith row of $\mathbf{N}$ be denoted n_{i+}, the sum of the jth column of $\mathbf{N}$ be denoted n_{+j}, and the sum of all the entries of $\mathbf{N}$ be denoted n (i.e., n_{++}). In the case of DAYS and FRIENDS, $n = 8411$ students who responded to the two questions of interest. A matrix $\mathbf{P} = [pij]_{(I \times J)}$, of relative frequencies consisting of the ratio of cell counts to the overall sum, called the *correspondence matrix* [7] [15] is computed as $\mathbf{P} = (1/n)\mathbf{N}$. Matrix $\mathbf{P}$ for the DAYS/FRIENDS data is

$$\mathbf{P} = \begin{pmatrix} .3459 & .1418 & .0954 & .0414 & .0373 \\ .0092 & .0131 & .0180 & .0084 & .0114 \\ .0039 & .0076 & .0095 & .0073 & .0045 \\ .0039 & .0045 & .0071 & .0050 & .0058 \\ .0032 & .0062 & .0084 & .0074 & .0128 \\ .0024 & .0049 & .0106 & .0106 & .0159 \\ .0105 & .0121 & .0219 & .0283 & .0638 \end{pmatrix}.$$

The next step of the analysis is to center and scale matrix $\mathbf{P}$.

Let $\mathbf{r} = \mathbf{P1}$ where $\mathbf{1}$ is a $J \times 1$ column vector of ones. Then $\mathbf{r}$ is a column vector consisting of a set of "row masses" [7] that can also be computed by dividing each row total n_{i+} by n. Also, let $\mathbf{c} = \mathbf{P}'\mathbf{1}$ where, in this case, $\mathbf{1}$ is a $I \times 1$ column vector of ones. Column vector $\mathbf{c}$ consists of a set of "column masses"; that is, the column totals (the n_{+j}'s) divided by n. The sets of row masses and column masses computed for the DAYS/FRIENDS data are

$$\mathbf{r}' = [.6618\, .0600\, .0328\, .0264\, .0380\, .0443\, .1366]$$

and

$$\mathbf{c}' = [.3789\, .1902\, .1708\, .1083\, .1517].$$

The row masses and column masses are used to center the correspondence matrix. Denote this centered matrix as $\tilde{\mathbf{P}}$, and compute $\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{rc}'$, or in other words, $\tilde{\mathbf{P}}$ is a matrix formed by subtracting the product of the (i)th row mass and the (j)th

column mass from each p_{ij}. For ease of matrix manipulation, let **Dr** be a diagonal $I \times I$ matrix with the row masses as diagonal entries and let **Dc** be a diagonal $J \times J$ matrix with the column masses as diagonal entries. The scaling of correspondence matrix $\mathbf{P}^*$ is accomplished by the following operation:

$$\mathbf{P}^* = \mathbf{Dr}^{-1/2}\,\tilde{\mathbf{P}}\,\mathbf{Dc}^{-1/2}.$$

Matrix $\mathbf{P}^* = \left[p^*_{ij}\right]_{(I\times J)}$ can also be calculated by dividing each (i,j)th entry of $\tilde{\mathbf{P}}$ by the square root of the product of the (i)th row mass and the (j)th column mass. The centered and scaled matrix for DAYS/FRIENDS is

$$\mathbf{P}^* = \begin{pmatrix} .1899 & .0450 & -.0527 & -.1132 & -.1990 \\ -.0901 & .0155 & .0760 & .0240 & .0242 \\ -.0763 & .0173 & .0522 & .0620 & -.0065 \\ -.0608 & -.0071 & .0391 & .0399 & .0288 \\ -.0933 & -.0124 & .0241 & .0506 & .0930 \\ -.1113 & -.0388 & .0345 & .0834 & .1122 \\ -.1815 & -.0860 & -.0096 & .1110 & .2995 \end{pmatrix}.$$

Continuing the CA, a generalized SVD of $\mathbf{P}^*$ ultimately produces a matrix $\mathbf{Y}$ the rows of which provide the graphing coordinates for row points pertaining to the levels of the row variable in the two-way-table. A second matrix $\mathbf{Z}$ gives the coordinates for the column points that belong to the categories of the column variable in the table. The SVD of $\mathbf{P}^*$ generates matrices of eigenvalues and eigenvectors such that

$$\mathbf{P}^* = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{V}'.$$

Diagonal matrix $\mathbf{\Lambda}_{I\times J} = diag(\lambda_1, \lambda_2, \ldots, \lambda_{J-1})$ has the reverse ranked positive constants called singular values (eigenvalues) along its main diagonal, $\mathbf{U}$ is an orthogonal $I \times I$ matrix, and $\mathbf{V}$ is a $J \times J$ orthogonal matrix. Now we specify $\tilde{\mathbf{U}} = \mathbf{D_r}^{1/2}\mathbf{U}$ and $\tilde{\mathbf{V}} = \mathbf{D_c}^{1/2}\mathbf{V}$. Then the matrix $\tilde{\mathbf{U}}$ contains normalized left singular vectors and $\tilde{\mathbf{V}}$ contains normalized right singular vectors, the unit lengths of which match the metrics defined by the row margins $\mathbf{D_r}^{-1}$ and the column margins $\mathbf{D_c}^{-1}$, respectively. Johnson and Wichern explain that, *the columns of* $\tilde{\mathbf{U}}$ *define the coordinate axes for the points representing the column conditional distributions of* $\mathbf{P}$. *Similarly, the columns of* $\tilde{\mathbf{V}}$ *define the coordinate axes for the points representing the row conditional distributions of* $\mathbf{P}$ [15] (p.774). Finally, the graphing coordinates of the row conditional distributions are computed as

$$\mathbf{Y} = \mathbf{D}_r{}^{-1}\,\tilde{\mathbf{U}}\,\mathbf{\Lambda}$$

and the coordinates for the column conditional distributions as

$$\mathbf{Z} = \mathbf{D}_c{}^{-1}\,\tilde{\mathbf{V}}\,\mathbf{\Lambda}'.$$

The first two column vectors of $\mathbf{Y}$ and $\mathbf{Z}$ provide the pairs of coordinates giving the data's "best" two-dimensional representation that will be used to graph row points and column points, respectively. The results for a two-way table are thus two sets of planar coordinates that are superimposed to produce a CA plot. Now it is time to perform SVD for the CA of DAYS and FRIENDS.

The singular value decomposition of matrix $\mathbf{P}^*$ given by the DAYS and FRIENDS

data produces the three matrices that follow:

$$\mathbf{U} = \begin{pmatrix} -.5559 & .1640 & -.0382 & -.0097 & -.8109 & .0466 & .0529 \\ .1542 & -.5387 & .7275 & -.1211 & -.2212 & .1645 & .2571 \\ .1195 & -.5614 & -.5286 & .4898 & -.1541 & -.0170 & .3567 \\ .1383 & -.2743 & -.1019 & -.3956 & -.2011 & -.8085 & -.2112 \\ .2604 & -.0999 & .1229 & .4536 & -.2551 & .1107 & -.7899 \\ .3360 & -.1260 & -.3996 & -.6030 & -.2082 & .5339 & -.1397 \\ .6731 & .5164 & .0685 & .1396 & -.3477 & -.1394 & .3402 \end{pmatrix},$$

$$\mathbf{\Lambda}_{(I\times J)} = diag(.5474, .1584, .0410, .0184, .0000),$$

and

$$\mathbf{V} = \begin{pmatrix} -.5862 & .4345 & -.1399 & -.2629 & .6156 \\ -.1747 & -.2969 & .2251 & .8003 & .4362 \\ .1170 & -.6394 & .3476 & -.5346 & .4133 \\ .3571 & -.2243 & -.8433 & .0525 & .3291 \\ .6962 & .5138 & .3127 & .0424 & .3895 \end{pmatrix}.$$

Furthermore, calculation of the two sets of coordinates gives (only the first two

dimensions are shown),

$$\mathbf{Y} = \begin{pmatrix} -.3741 & .0319 \\ .3445 & -.3482 \\ .3612 & -.4908 \\ .4662 & -.2674 \\ .7308 & -.0811 \\ .8734 & -.0948 \\ .9969 & .2213 \end{pmatrix}$$

and

$$\mathbf{Z} = \begin{pmatrix} -.5213 & .1118 \\ -.2193 & -.1078 \\ .1550 & -.2450 \\ .5939 & -.1079 \\ .9785 & .2089 \end{pmatrix}.$$

The spatial interpretation of points in the two-dimensional CA plot is important to note. The Euclidean distances between pairs of points representing different levels of a single variable are related directly to the statistical (chi-square) distance between the corresponding pairs in that variable's profiles. Essentially, this means that row points (or column points) that are close together in the plot indicate rows (columns) having similar conditional probability distributions across the columns (down the rows). On the other hand, the Euclidean distance between a row point and a column point has no such clear distance relation [15]. However, an intuitive interpretation

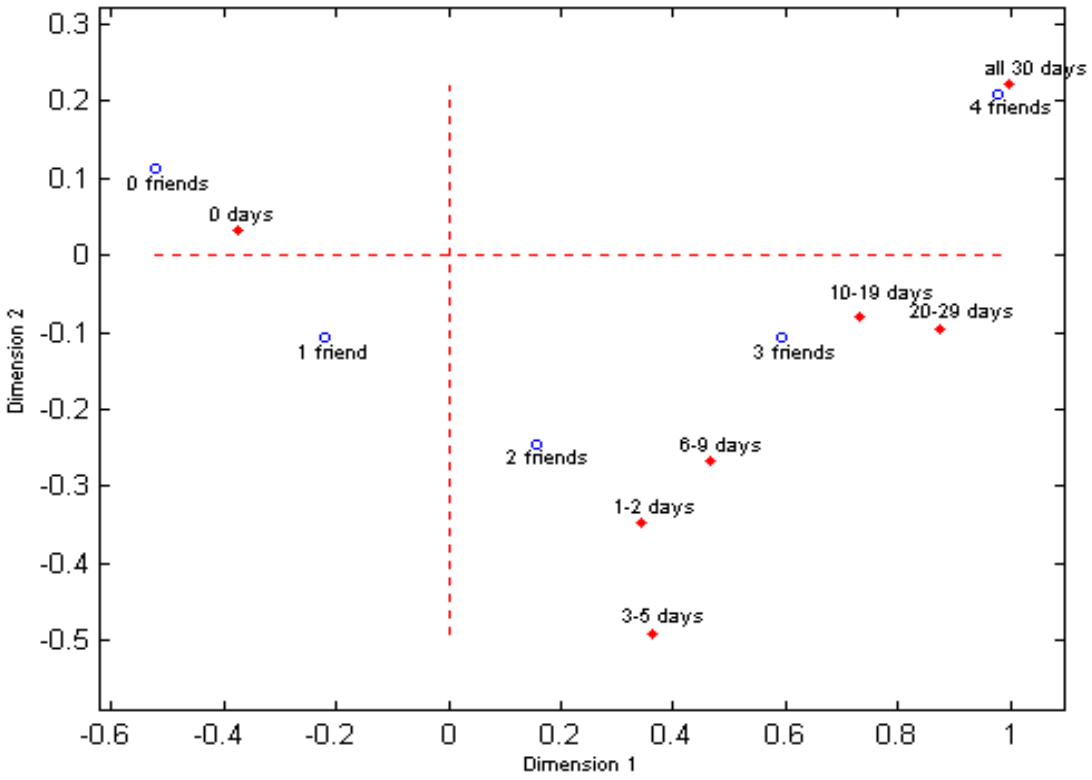

Figure 4: Correspondence Analysis Plot of DAYS and FRIENDS

exists for the distance between two points belonging to different variables as a visual expression of the residual between the observed and expected counts $n_{ij} - e_{ij}$. Thus, a given row point appears nearer to some column point when the residual is positive and farther away if the residual is negative [7]. It should be helpful, at this point, to view and interpret an actual CA plot.

The CA plot for DAYS and FRIENDS, displayed in figure 4, gives interesting insight into the relationship between these variables. However, examination of the row profiles and column profiles will also make the discussion of the plot clearer.

Row profiles consist of each cell frequency divided by the corresponding row total. Likewise the column profiles give each cell frequency divided by the matching column total. In the case of the DAYS and FRIENDS table, the row profiles show the conditional distributions of FRIENDS given a particular category of DAYS, and the column profiles are the conditional distributions of DAYS with respect to number of FRIENDS. This distinction is important since SCA essentially gives each DAYS point as a weighted average of the scores in the conditional distribution of FRIENDS. Likewise, each FRIENDS point is a weighted average of the scores for the DAYS categories [7]. The row profiles $\mathbf{R}$, and the column profiles $\mathbf{C}$ are shown below in matrix form as

$$\mathbf{R} = \begin{pmatrix} .5226 & .2143 & .1441 & .0625 & .0564 \\ .1525 & .2178 & .2990 & .1406 & .1901 \\ .1196 & .2319 & .2899 & .2210 & .1377 \\ .1486 & .1712 & .2703 & .1892 & .2207 \\ .0844 & .1625 & .2219 & .1937 & .3375 \\ .0536 & .1099 & .2386 & .2386 & .3592 \\ .0766 & .0888 & .1601 & .2071 & .4674 \end{pmatrix}$$

and

$$\mathbf{C} = \begin{pmatrix} .9128 & .7456 & .5581 & .3820 & .2461 \\ .0242 & .0687 & .1051 & .0779 & .0752 \\ .0104 & .0400 & .0557 & .0670 & .0298 \\ .0104 & .0237 & .0418 & .0461 & .0384 \\ .0085 & .0325 & .0494 & .0681 & .0846 \\ .0063 & .0256 & .0619 & .0977 & .1050 \\ .0276 & .0638 & .1280 & .2613 & .4208 \end{pmatrix}.$$

First, note the pattern of row points representing the seven categories of DAYS. The CA plot shows that the row point representing *1 to 2 days* is near to the row point representing *6 to 9 days*. The proximity in the two points reflects the similarity in proportions found in the second and fourth rows of **R**. This means that the FRIENDS profiles for students who claim smoking cigarettes on one to two days is similar to those for students reporting six to nine days. Also, the row points representing *10 to 19 days* and *20 to 29 days* are close to each other since the entries in the fifth and sixth rows of **R** are comparable. In other words, the conditional distributions of FRIENDS are nearly the same for *10 to 19* and *20 to 29* days of smoking cigarettes.

Next, look at the positions of the column points in the plot. No column point for the FRIENDS variable is very near any other column point. Looking at the columns of **C** gives evidence of a fairly large difference when comparing the first proportion in each column. Thus, the conditional distributions for DAYS tend to be different for every level of FRIENDS.

Finally, the most dramatic element of the plot is the graphic depiction of the

association between the variables DAYS and FRIENDS. Note the proximity of point *0 days* to point *0 friends*, and that of point *all 30 days* to the *4 friends* point. The observed frequency of students responding *0 days* and *0 friends*; that is, the $(i = 1, j = 1)$th cell, is 2909, while the expected count is 2109.01. Therefore, the residual for this cell is 799.99, which is large and positive. Likewise, the residual for the $(7, 5)$th cell is $537.00 - 174.31 = 362.69$. However, the residual for the $(7, 1)$th cell is $88.00 - 435.37 = -347.37$, and is large and negative. Thus, in the CA plot, the row point representing *all 30 days* is far from the column point representing *0 friends*. Additionally, the column point for *2 friends* is relatively close to the row points representing *1 to 2 days*, *3 to 5 days*, and *6 to 9 days*, while the *3 friends* column point is close to row points for *10 to 19 days* and *20 to 29 days*. Clearly, from the plot, the number of days a student claims to have smoked in the past month is directly associated with the number of the student's four closest friends that smoke. Although this information can be readily obtained from the chi-square analysis of the two-way table by a knowledgeable researcher, the value of the CA plot is that it displays the strength and direction of association between two discrete variables in a forthright and powerful way that can be recognized even by those uninitiated in the chi-square techniques. In the case of CA it seems that a picture really is worth a thousand words. However, the picture does not provide the entire story of CA.

Besides the plot there is another result of CA worth mentioning; a mathematical result called *inertia*. Inertia is a numerical measure that describes the amount of information from the chi-square analysis of a two-way table remaining in each dimension [15]. The total inertia for the CA of a two-way table is defined as χ^2/n. Total

inertia can also be calculated as the sum of the elements on the main diagonal of the matrix formed by the operation $\mathbf{D_r}^{-1}\tilde{\mathbf{P}}\mathbf{D_c}^{-1}\tilde{\mathbf{P}}'$ and, as such, represents the magnitude of difference from an independence model that requires explanation, according to Jobson [14]. The total inertia involved in the CA of the DAYS and FRIENDS table is 0.3268 ($\chi^2 = 2748.56$). Since the CA plot is the outcome of the first two dimensions resulting from the analysis, we are most interested in the inertia pertaining to dimensions one and two. The components of the total inertia that pertain to each dimension consist of the individual squared singular values on the main diagonal of matrix $\mathbf{\Lambda}$. Therefore, the inertia components for the CA of DAYS and FRIENDS are 0.2997, 0.0251, 0.0017, and 0.0003, listed per ascending dimension. In order to compute the proportion of the information from the chi-square statistic retained in each dimension, simply divide each component by the total inertia. Thus, the inertia associated with the first dimension of the CA plot for DAYS and FRIENDS is 0.2997, accounting for 91.7% of the total inertia, and the inertia related to the second dimension is 0.0251, which accounts for an additional 7.7%. Together the two dimensions explain 99.4% of the total inertia, suggesting that the two-dimensional representation approximates the data very well.

5.2 Correspondence Analysis for Multi-Way Tables

Multiple correspondence analysis (MCA) is not a simple extension of SCA, yet the analysis is easily done using SAS software [30] and the *Proc Corresp* command. Suppose we have K discrete variables and that the number of levels for variable k is J_k. The K-way contingency table, having size $J = J_1 \times J_2 \times \cdots \times J_k$, represents $n = n_{++\ldots}$

total observations. The MCA of such a K-way table is accomplished by SVD of a *Burt* matrix [7]. The symmetric partitioned Burt matrix is described generically as

$$\mathbf{B} = \begin{pmatrix} \mathbf{N}_{[1]} & \mathbf{N}_{[12]} & \cdots & \mathbf{N}_{[1K]} \\ \mathbf{N}_{[21]} & \mathbf{N}_{[2]} & \cdots & \mathbf{N}_{[2K]} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{N}_{[k1]} & \mathbf{N}_{[k2]} & \cdots & \mathbf{N}_{[K]} \end{pmatrix}$$

where the $\mathbf{N}_{[i]}$ blocks located on the main diagonal give the marginal frequencies for the (i)th variable. According to Friendly, SVD of the Burt matrix *produces scores for the categories of all variables so that the greatest proportion of the bivariate, pairwise associations in all off-diagonal blocks is accounted for in a small number of dimensions* [7] (p.169). Therefore, the MCA plot allows exploration of bivariate associations among larger groups of variables. This property is ideal for surveys containing large numbers of categorical variables such as the TnYTS. Here are some pointers, also provided by Friendly [7], that will aid interpretation of the MCA plots.

- The centroid (i.e., weighted average) of the full set of category points for a variable is located at the origin of the plot. This holds for each variable.
- A single variable's contribution to inertia increases with the number of response levels.
- The amount of inertia contributed by a particular category of a given variable is inversely related to its marginal frequency.
- The two points representing a binary variable will be located on a line pass-

ing through the origin, and the distance from the origin to each point is also inversely related to its marginal frequency.

So, interpretation of MCA plots is fairly similar to that of SCA plots. There are, however, critical differences between SCA and MCA. The main difference is in the way inertia is computed. In the case of MCA for a two-way table of size $(J_1 \times J_2)$, total inertia does not depend on the chi-square statistic at all but, instead, is calculated as $(J_1+J_2-2)/2$. In addition, MCA produces J_1+J_2-2 dimensions but half of these are deemed artifacts. The artifacts are easily disregarded, though, since the components of the total inertia pertaining to these dimensions have value less than one half. For the analysis of K-way tables total inertia is computed as $J/K-1$ and the nontrivial dimensions are considered to be those with inertia components greater in value than $1/K$. Consequently, the proportions of χ^2 explained by each dimension tend to be undervalued, leading to rather conservative estimates of the association accounted for in the dimensions that are plotted [7]. Despite these limitations, MCA still provides a useful and easily grasped technique in the exploratory analysis of categorical data. The next section includes some examples of MCA plots for various sets of variables from the TnYTS.

5.3 Correspondence Analysis Graphs for the 2000 TnYTS

The MCA plots shown here have a slightly different appearance compared to figure 4, which was created as a Matlab plot [17]. The multiple correspondence analysis that created the following plots was performed in SAS [30], the coding for which is displayed in appendix A.7. The two-dimensional coordinates were then graphed

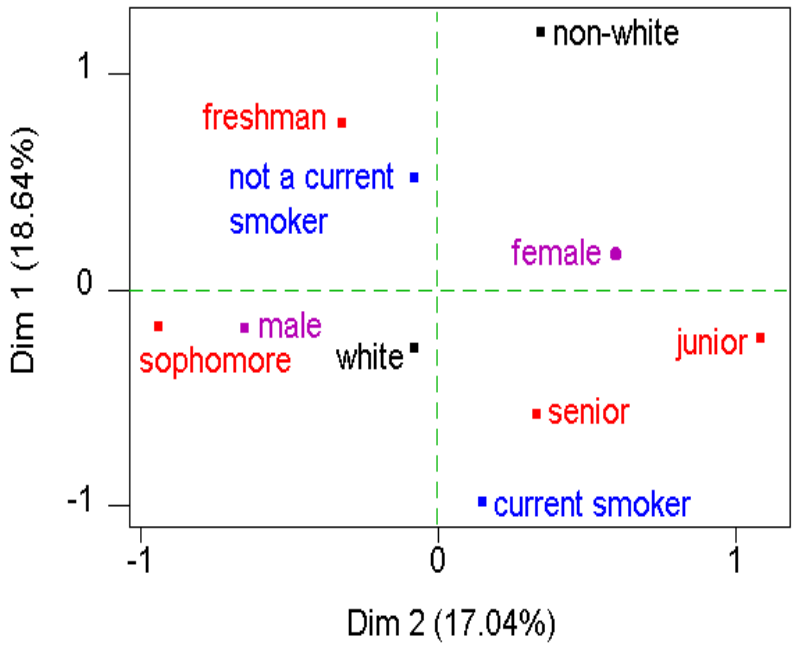

Figure 5: MCA Plot of Smoking Status, Gender, Grade, and Ethnicity

in Minitab [20], a statistical software preferred in this application for the ease and flexibility of labeling and coloring points within a plot.

The first MCA plot considered is shown in figure 5. Notice that the *conservative* values given by Proc Corresp in SAS [30] for the proportion of total inertia contributed per dimension are specified with the axes labels. If this set of variables seems familiar it is because it was discussed previously as an example of logistic regression in model (1). Recall that the binary response for this model is whether or not the student has smoked a cigarette in the past 30 days. The predictor variables include gender, grade, and ethnicity. The most striking association shown here is that freshmen do

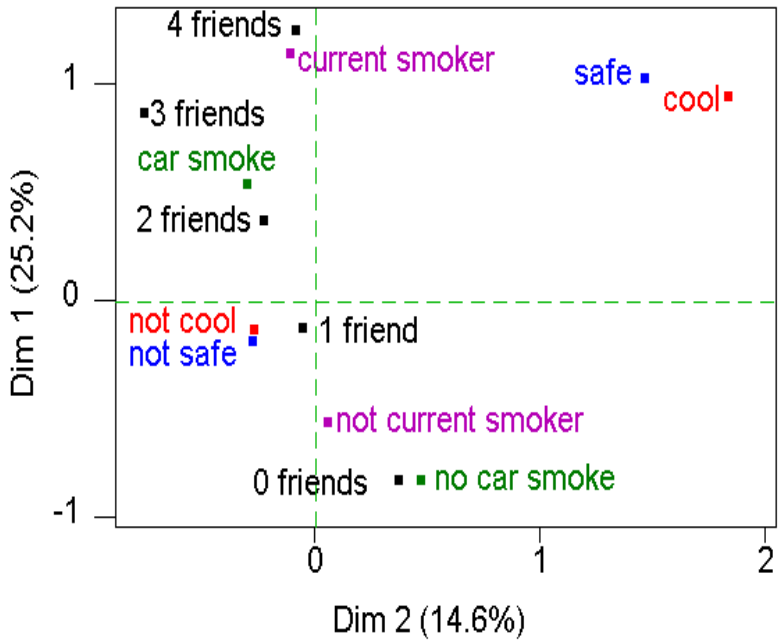

Figure 6: MCA Plot of Smoking Status, Car, Safe, Cool, and Friends

not tend to be current smokers while seniors are more likely to be current smokers. Also, being white is associated with current smoking. There does not appear to be a clear association between gender and smoking. This helps to emphasize the point that significance of associations in a χ^2 tests of independence is sometimes exaggerated by very large sample sizes.

The next MCA plot, in figure 6, shows relationships between a student's current smoking behavior and some of the more significant predictor variables included in logit model (3) including Car, Safe, Cool, and Friends. It is not surprising that a student's recent exposure to cigarette smoke in cars increases with the number of

close friends that smoke. This association is shown clearly in the plot, along with the fact that a student who does not currently smoke cigarettes is most likely to have no close friends that smoke, but may think one cigarette-smoking friend is tolerable. Exposure to car smoke and the number of close friends that smoke are key indicators for a couple of reasons. Firstly, as mentioned previously, some people start smoking cigarettes as a direct response to being in a car with the smoke from other peoples' smoke. Secondly, since smoke from enclosed spaces such as cars often leaves a distinct odor in a person's clothing and hair, it is a marker of sorts. It is also important for those interested in designing prevention or cessation programs to note that students claiming not to be current smokers probably think that smoking is not safe and that smoking cigarettes does not necessarily make young people look cool.

The final example of MCA, figure 7, shows associations between several variables within the realm of tobacco use that were measured by TnYTS, including smokable and smokeless types, non-tobacco cigarettes such as kreteks, a blend of tobacco and clove extract, and Indian bidis (beedies) consisting of brown leaves packed with tobacco and tied up with a thread [40], as well as the number of cigarette smoking friends. The main ideas conveyed by this plot are that students willing to smoke cigarettes are more likely to engage in use of other tobacco products compared to those who have never smoked cigarettes, and furthermore, use of various types of tobacco is positively associated with the number of friends that smoke. However, experimentation with the cigarette alternatives, kreteks and bidis, appears to be a fringe behavior.

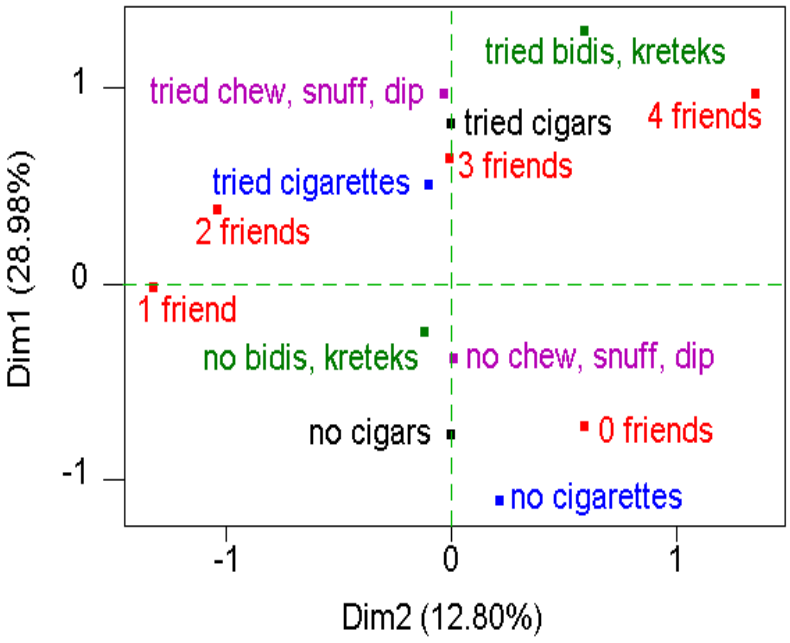

Figure 7: MCA Plot of Student's Experimentation with Different Types of Tobacco and Number of Closest Friends that Smoke Cigarettes

5.4 Usefulness of Correspondence Analysis in the Context of Youth Tobacco Surveys

Perhaps the most useful aspect of correspondence analysis is that it offers an exciting alternative for the exploratory data analysis of categorical variables. The math is somewhat more involved than that needed to produce a bar chart or pie chart, but the graphs produced are capable of conveying much useful information without requiring a great deal of viewer education, and computer software such as SAS [30], Matlab [17], or Minitab [20] make the analysis fairly easy. Moreover, correspondence analysis can be extended in order to use significant interaction terms that are identified by loglinear modeling [7].

6 CONCLUSION

Categorical data analysis is a rapidly expanding specialty field in the study of statistics, and we have mentioned only a few of the tools available. The state Youth Tobacco Survey is a fairly sophisticated instrument that attempts to measure many complex variables related to an important public health issue. However, the methods of analysis involving YTS data that have been reported to this point do not appear to make proper use of contemporary techniques. Logistic modeling, loglinear modeling, and correspondence analysis are three procedures recommended here because they offer powerful means for gaining new and relevant insight into the issues surrounding teen tobacco use as measured by the YTS, and their use is supported by popular statistical software packages such as SAS. The next step in this study is to publish the results and recommendations discussed here in a journal dedicated to the study and control of tobacco use.

BIBLIOGRAPHY

[1] A. Agresti, *Categorical Data Analysis*, Wiley and Sons, New York (1990).

[2] U. E. Bauer, T. M. Johnson, R. S. Hopkins, and R. G. Brooks, Changes in youth cigarette use and intentions following implementation of a tobacco control program. *Journal of the American Medical Association.* 284(6) (2000) 723–728.

[3] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge (1975).

[4] S. C. Carvajal, D. E. Wiatrek, R. I. Evans, C. R. Knee, and S. G. Nash, Psychosocial determinants of the onset and escalation of smoking: Cross-sectional and prospective findings in multiethnic middle school samples. *Journal of Adolescent Health.* 27(4) (2000) 255–265.

[5] E. Conlisk and S. H. Malek, Results from the 1999 North Carolina youth tobacco survey. *North Carolina Medical Journal.* 62(5) (2001) 256–259.

[6] Florida Department of Health, 2000 Florida Youth Tobacco Survey results. *FYTS.* 3(1) (Revised version dated June 21, 2000), `www.doh.state.fl.us/Disease_ctrl/epi/FYTS/vol3rep_1.pdf`.

[7] M. Friendly, *Visualizing Categorical Data*, SAS Publishing, Cary, NC (2000).

[8] A. Gelman, C. Pasarica, and R. Dodhia, Statistical computing and graphics: Let's practice what we preach: Turning tables into graphs. *The American Statistician.* 56(2) (2002) 121–130.

[9] Georgia Department of Human Resources and the Coalition for a Healthy and Responsible Georgia, *The Burden of Tobacco in Georgia* (2000), www.ph.dhr.state.ga.us/publications/reports.shtml.

[10] Georgia Department of Human Resources Division of Public Health, *Georgia Youth Tobacco Survey, Summary Report* (1999), www.ph.dhr.state.ga.us/programs/tobacco/pdfs/summaryreport99.pdf.

[11] M. O. Hill, Correspondence analysis: A neglected multivariate method. *Applied Statistics.* 23(3) (1974) 340– 354.

[12] K. A. Horn, X. Gao, G. A. Dino, and S. Kamal– Bahl, Determinants of youth tobacco use in West Virginia: A comparison of smoking and smokeless tobacco use. *American Journal of Drug and Alcohol Abuse.* 26(1) (2000) 125–138.

[13] T. T. K. Huang, B. A. Unger, and L. A. Rohrbach, Exposure to, and perceived usefulness of, school-based tobacco prevention programs: Associations with susceptibility to smoking among adolescents. *Journal of Adolescent Health.* 27(4) (2000) 248–254.

[14] J. D. Jobson, *Applied Multivariate Data Analysis Volume II: Categorical and multivariate methods*, Springer–Verlag, New York (1992).

[15] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River, NJ (1998).

[16] Kansas Department of Health and Environment, *1999 Kansas Youth Tobacco Survey Report no. 1* (1999), `www.kdhe.state.ks.us/tobacco/resources/kyts_99.pdf`.

[17] *Matlab: the language of technical computing*, The MathWorks, Inc. Version 5.3.0.620a (R11) Student Edition (1999), and Version 6.0.0.88 Release 12, (2000).

[18] P. McCullagh and J. A. Nelder, *Generalized Linear Models (2nd ed)*, Chapman and Hall/CRC, Boca Raton (1989).

[19] MidSOUTH Prevention Institute at ULAR, Youth tobacco use in Arkansas remains above national level. *Prevention Outlook.* 2(4) (2001).

[20] *Minitab Statistical Software*, Minitab Inc. Release 13.1 (2000).

[21] Minnesota Department of Health Center for Health Statistics, *Teens and Tobacco in Minnesota, Executive summary: Results from the Minnesota Youth Tobacco Survey* (2000), www.health.state.mn.us/divs/opa/ytssumm.pdf.

[22] Mississippi State Department of Health and the Social Science Research Center at MSSTATE, *1998 Mississippi Youth Tobacco Survey Report 1* (1998), www.msdh.state.ms.us/documents/tobacco.1998yts.pdf.

[23] Missouri Department of Health Division of Chronic Disease Prevention and Health Promotion, *Tobacco Use Among Missouri Middle School Students*(1999), www.health.state.mo.us/SmokingAndTobacco/TotalMSReport.pdf.

[24] New Jersey Department of Health and Senior Services and the University of Medicine and Dentistry of New Jersey-School of Public Health, *1999 New Jersey Youth Tobacco Survey: A statewide report* (2000), www.state.nj.us/health/as/yts/yts.pdf.

[25] Oklahoma State Department of Health, *Oklahoma Youth Tobacco Survey Report no. 1* (1999), www.health.state.ok.us/program/tobac/oyts/.

[26] F. L. Ramsey and D. W. Schafer, *The Statistical Sleuth: A course in methods of data analysis*, Duxbury Press, Belmont, CA (1997).

[27] N. A. Rigotti, J. E. Lee, and H. Weschler, US college students' use of tobacco products. *Journal of the American Medical Association.* 284(6) (2000) 699–705.

[28] P. Rode and J. Oswald, New findings from the Minnesota youth tobacco survey. *Minnesota Medicine.* 84:7 (2001) 42–46.

[29] M. L. Samuels, Simpson's paradox and related phenomena. *Journal of the American Statistical Association.* 88(421) (1993) 81–88.

[30] *SAS System*, SAS Institute Inc. Cary, NC Release 8.00.00 MOP110199 (1999).

[31] E. H. Simpson, The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological).* 13(2) (1951) 238–241.

[32] S. Soldz, P. Kreiner, T. W. Clark, and M. Krakow, Tobacco use among Massachusetts youth: Is tobacco control working? *Preventive Medicine.* 31 (2000) 287–295.

[33] South Dakota Department of Health Tobacco Control Program, *The 2000 South Dakota Youth Tobacco Survey Executive Summary* (2000), www.state.sd.us/doh/News/ytssummary.htm.

[34] South Dakota KIDS COUNT Project, Teens and smoking. *Facts on Kids in South Dakota.* 2(00) (2000), `www.usd.edu/brbinfo/brb/kc/pdf_files/Teens%20and%20smoking%20single%20041900.pdf`.

[35] M. E. Stokes, C. S. Davis, and G. G. Koch, *Categorical Data Analysis Using the SAS System*, SAS Institute, Inc., Cary, NC (1995).

[36] Tennessee Department of Health and Tennessee Tobacco Surveillance Program, *1999 Tennessee Youth Tobacco Survey Report 2: Environmental tobacco smoke* (2000), hitspot.utk.edu/ chrg/hit/main/reports/tnyts/1999/tnytsrpt2.pdf.

[37] Tennessee Department of Health and Tennessee Tobacco Surveillance Program, *1999 Tennessee Youth Tobacco Survey Report 3: Social influences: Kids and tobacco* (2000), hitspot.utk.edu/ chrg/hit/main/reports/tnyts/1999/tnytsrpt3.pdf.

[38] Tennessee Department of Health and Tennessee Tobacco Surveillance Program, *2000 Tennessee Youth Tobacco Survey Report 1: Tobacco use and prevalence* (2000), www.state.tn.us/health/Downloads/2000TnYTS.pdf.

[39] Texas Department of Health Bureau of Disease and Injury Prevention, *1998 Texas Youth Tobacco Survey Report 1: Current tobacco use* (1998), www.ssrc.msstate.edu/socialclimate/texas/rep1.pdf.

[40] U. S. Department of Health and Human Services, CDC Surveillance Summaries: Youth tobacco surveillance United States, 1998–1999. *Morbidity and Mortality Weekly Report.* 49:SS–10 (2000).

[41] U. S. Department of Health and Human Services, CDC Surveillance Summaries: Youth tobacco surveillance United States, 2000. *Morbidity and Mortality Weekly Report.* 50:SS–4 (2001).

APPENDICES

.1 Recoding of TnYTS Variables for SAS

Base SAS Code Showing Recoding and New Variable Names

```
options ls=80 ps=60;

data  high ;
set tobacco.tnhstate ;

age=cr1;
if age<=3 then newage=14;
if age=4  then newage=15;
if age=5  then newage=16;
if age=6  then newage=17;
if age=>7 then newage=18;

male=cr2-1;

if cr3=. then grade =.;
else grade = cr3 + 5;

if cr5=.  then black=.;
if cr5<=2 then black=0;
if cr5=3  then black=1;
if cr5>=4 then black=0;
if cr5=.  then white=.;
if cr5<=5 then white=0;
if cr5=6  then white=1;

evrsmoke=2-esmoke;

if cr7=. then firstage=.;
if cr7=1 then firstage=.;
if cr7=2 then firstage=8;
if cr7=3 then firstage=9;
if cr7=4 then firstage=11;
if cr7=5 then firstage=13;
if cr7=6 then firstage=15;
if cr7=7 then firstage=17;
```

```
if cr7=. then preteen=.;
else if cr7=1 then preteen=.;
else if cr7>=5 then preteen=0;
else preteen=1;

if cr8=. then nbrpacks=.;
else if cr8<=1 then nbrpacks=0;
else if cr8=8  then nbrpacks=2;
else nbrpacks = 1;

evdaily=2-cr9;

if cr10=1 then la30smok=0;
if cr10=. then la30smok=.;
if cr10>1 then la30smok=1;
bbnotl30=evrsmoke-la30smok;

if cr10=1  then la30inte=.;
if cr10=.  then la30inte=.;
if cr10>=2 then la30inte=cr10;

if cr13=. then menthol=.;
if cr13=1 then menthol=.;
if cr13=2 then menthol=1;
if cr13=3 then menthol=0;

if tnr16=. then loose=.;
else loose=2-tnr16;

if cr16=. then cigproof=.;
if cr16=1 then cigproof=.;
if cr16=3 then cigproof=0;
if cr16=2 then cigproof=1;

if cr17=. then refused=.;
if cr17=1 then refused=.;
if cr17=3 then refused=0;
if cr17=2 then refused=1;

if cr21=1 then tryquit=.;
```

```
if cr21=. then tryquit=.;
if cr21=2 then tryquit=1;
if cr21=3 then tryquit=0;

if cr22=1 then wantquit=.;
if cr22=. then wantquit=.;
if cr22=2 then wantquit=1;
if cr22=3 then wantquit=0;

if cr23=1 then manyquit=.;
if cr23=. then manyquit=.;
if cr23=2 then manyquit=0;
if cr23=3 then manyquit=1;
if cr23>3 then manyquit=2;

if cr25=. then evchsndp=.;
else evchsndp = 2-cr25;

if tnr26=1 then approve=.;
if tnr26=. then approve=.;
if tnr26=2 then approve=1;
if tnr26=3 then approve=0;
if tnr26=4 then approve=.;

if cr30=. then evrcigar=.;
else evrcigar = 2-cr30;

if cr35=. then evrbidis=.;
else if cr35=4 then evrbidis=0;
else evrbidis=1;

if cr41=1 then discudan=1;
if cr41=2 then discudan=1;
if cr41=3 then discudan=1;
if cr41=4 then discudan=0;

if cr42=. then tobadict=.;
else if cr42>=3 then tobadict=0;
else tobadict=1;
```

```
if cr43=1 then morefrie=1;
if cr43=2 then morefrie=1;
if cr43=3 then morefrie=0;
if cr43=4 then morefrie=0;
if cr43=. then morefrie=.;

if cr44=1 then cool=1;
if cr44=2 then cool=1;
if cr44=3 then cool=0;
if cr44=4 then cool=0;
if cr44=. then cool=.;

if cr45=1 then harm=1;
if cr45=2 then harm=1;
if cr45=3 then harm=0;
if cr45=4 then harm=0;
if cr45=. then harm=.;

if cr46=. then safe=.;
else if cr46>2 then safe=0;
else safe=1;

if cr47=. then ablequit=.;
if cr47=1 then ablequit=.;
if cr47=2 then ablequit=1;
if cr47=3 then ablequit=0;

if tnr48=1 then liveshor=1;
if tnr48=2 then liveshor=1;
if tnr48=3 then liveshor=0;
if tnr48=4 then liveshor=0;
if tnr48=. then liveshor=.;

if cr48=1 then progquit=.;
if cr48=. then progquit=.;
if cr48=2 then progquit=1;
if cr48=3 then progquit=0;

if cr49=1 then pracscho=1;
if cr49=. then pracscho=.;
```

```
if cr49>1 then pracscho=0;

if cr50=. then commuact=.;
else if cr50=1 then commuact=1;
else commuact = 0;

if cr51=. then anticomm=.;
else if cr51>2 then anticomm=1;
else anticomm=0;

if cr52=. then actors=.;
else if cr52=1 then actors=0;
else if cr52=5 then actors=0;
else actors = 1;

if cr53=. then athletes=.;
else if cr53=1 then athletes=0;
else if cr53=5 then athletes=0;
else athletes = 1;

if cr54=. then ntrnetad=.;
else if cr54=1 then ntrnetad=0;
else if cr54=5 then ntrnetad=0;
else ntrnetad = 1;

if tnr53=1 then doctor=.;
if tnr53=. then doctor=.;
if tnr53=2 then doctor=1;
if tnr53=3 then doctor=0;

if tnr54=1 then dentist=.;
if tnr54=. then dentist=.;
if tnr54=2 then dentist=1;
if tnr54=3 then dentist=0;

if cr57=. then roomsmok=.;
else if cr57=1 then roomsmok=0;
else roomsmok = 1;

if cr58=. then carsmoke=.;
```

```
else if cr58=1 then carsmoke=0;
else carsmoke = 1;

if cr59=. then scndharm=.;
else if cr59>=3 then scndharm=0;
else scndharm = 1;

somehome=2-cr60;

hmfriend=cr62-1;
if cr62=6 then hmfriend=.;

if hmfriend>=1 then somefrie=1;
else if hmfriend=. then  somefrie=.;
else somefrie=0;

if tnr63=. then poster=.;
else if tnr63>2 then poster=0;
else tnr63=1;

hmfrichw=cr63-1;
if cr63=6 then hmfrichw=.;

if tnr65=. then bllboard=.;
else if tnr65>2 then bllboard=0;
else tnr65=1;

if tnr69=. then mislead=.;
else if tnr69>=3 then mislead=0;
else mislead = 1;

label
esmoke =   'ever smoked a cigarette'
morefrie = 'think that y. people who smoke have more friends'
liveshor = 'think that smokers have shorter lives'
somehome = 'somebody at home smokes yes=1'
discudan = 'both or one parent has discussed danger yes=1'
approve =  'kids smoke parent knows and =1 approves '
tryquit =  'smoked during the past 12 and tried to quit 1=yes'
evrsmoke = 'person has ever smoked, 1=yes'
```

```
male =     'gender, 1=male 0=female'
evdaily =  'ever smoked cigarettes daily, 1=yes'
la30smok = 'smoke in last 30 days, 1=yes'
la30inte = 'frequency of smoking last 30 days'
somefrie = 'at least one of the closest 4 friends smokes'
bbnotl30 = 'ever smoked but not last 30 days'
hmfriend = 'how many of 4 closest friends smoke'
pracscho = 'practice to say no in school '
progquit = 'participated program to quit'
doctor =   'doctor has talked about danger (if visited)'
dentist =  'dentist has talked about danger (if visited)'
harm =     'think that young people risk harming yes=1'
cool =     'think that smoking makes look cool'
manyquit = 'times (0,1 or more) that have tried to quit'
wantquit = 'Want to stop smoking'
newage =   'Age in years 14(or under),15,16,17, 18+ '
/* new ones (2-5-02)*/
black =    'student is black, 1=yes'
white =    'student is white, 1=yes'
firstage = 'age when 1st smoked whole cig, . if . or nonsmoker,
8 if 8yrs, 9 if 9 or 10yrs, 11 if 11 or 12yrs,
13 if 13 or 14yrs, 15 if 15 or 16yrs, 17 if >= 17yrs'
preteen =  'student was under 13 yrs old when 1st smoked
whole cig, 1=yes'
nbrpacks = 'lifetime nbr of cigarettes smoked, missing = .,
zero to a few puffs = 0, between 1 cig and 4 pcks=1,
5 pcks or more = 2'
menthol =  'usually smoke menthols, 1=yes'
evchsndp = 'ever chew, snuff, or dip, 1=yes'
evrcigar = 'ever tried cigar products (even a puff or 2), 1=yes'
tobadict = 'can people get addicted to tobacco, 1=yes'
roomsmok = 'been in a room with someone smoking ipw, 1=yes'
carsmoke = 'been in a car with someone smoking ipw, 1=yes'
actors =   'ever see actors in TV, movies use tobacco, 1=yes'
athletes = 'ever see athletes use tobacco on TV, 1=yes'
ntrnetad = 'ever see ads for tobacco on internet, 1=yes'
mislead =  'think tobaco co.s mislead youngsters, 1=yes'
scndharm = 'think second hand cig smoke is harmful to you, 1=yes'
cigproof = 'proof of age demanded upon selling cigs ipm, 1=yes'
refused =  'sell of cigs to you refused due to age ipm, 1=yes'
```

```
commuact = 'participated in anti-tobacco community activity ipy,
1=yes'
grade =    'grade in school'
evrbidis = 'ever tried bidis or kreteks, 1=yes'
safe =     'think it is safe to smoke for a year or two, 1=yes'
ablequit = 'cig smoker could quit if he/she wanted to, 1=yes'
loose =    'are loose cigarettes sold in area of home, 1=yes'
anticomm = 'seen or heard anti-smoking commercials ipm, 1=yes'
poster =   'seen anti-smoking poster(s)ipm, 1=yes'
bllboard = 'seen anti-smoking billboards ipm, 1=yes'
; % NOTE: SAS code using these variables will run below this semi-colon.
```

.2 Logistic Regression Programs

SAS Code: Logistic Regression and Two-Way Tables of Smoking Status versus Gender, Grade, and Ethnicity for Model 1

```
proc logistic descending;
model la30smok=male grade white;
proc freq;
tables la30smok*male/chisq cmh measures;
tables la30smok*grade/chisq cmh measures;
tables la30smok*white/chisq cmh measures;
run;
```

SAS Code: Logistic Regression of Smoking Status versus Grade, Firstage, Safe, Car, Approve and Loose for Model 2

```
proc logistic descending;
model la30smok=male grade firstage safe carsmoke approve loose;
run;
```

SAS Code: Logistic Regression of Smoking Status versus Gender, Grade, Ethnicity, Firstage, Car, Safe, Harm, Cool, Friends, and Home for Model 3

```
proc logistic descending;
model la30smok=male grade white firstage carsmoke safe harm cool
hmfriend somehome;
run;
```

.3 Programs for Odds Ratio Plots

Matlab Program for Odds Ratio Plot from Model 1

```
% Title: logitgram1.m
% This program charts the point estimates of incremental
% odds ratios and 95% confidence intervals corresponding to
% predictor variables in logistic regression model (3).
% INPUT NEEDED ============================================
% Enter variables names in order of descending odds ratio.
% Names of variables should go in quotes and separated by ;
VarName={'White';'Grade';'Male'};
% Enter odds ratio and margins of error
A=[1   1.577  0.173 0.194  % variable, pt est, lo marg, up marg
   2   1.149  0.045 0.046
   3   1.095  0.090 0.097];
% Enter axes limits as [xmin xmax ymin ymax]
V=[0.1 3.9 0.85 1.85];
% NO MORE INPUT IS NEEDED ==================================
% Label locations may need adjustment
varlabel=A(:,1); % variables
oddsrato=A(:,2);  % pt est for odds ratio from sas output
lomarger=A(:,3);  % margin of error to be subtracted from pt est
upmarger=A(:,4);  % margin of error to be added to pt est
nvar=length(varlabel);
oneliney=[1;1];
onelinex=[0;V(2)];
labLocy=0.975*(oddsrato-lomarger);
labLocx=varlabel-0.075;
figure (1)
errorbar(varlabel,oddsrato,lomarger,upmarger,'bo')
 hold on;
plot(onelinex,oneliney,'r'),axis(V);
hold off;
ylabel('Odds Ratio','Fontsize',11);
xlabel('Variable #', 'Fontsize',11);
set(gca, 'xtick', varlabel')
% annotation
for i=1:nvar;
text(labLocx(i,1),labLocy(i,1),VarName(i,1))
end;
```

```
end;
```

Matlab Program for Odds Ratio Plot from Model 2

```
% Title: logitgram2.m
% This program charts the point estimates of incremental
% odds ratios and 95% confidence intervals corresponding to
% predictor variables in logistic regression model (3).
% INPUT NEEDED ============================================
% Enter variables names in order of descending odds ratio.
% Names of variables should go in quotes and separated by ;
VarName={'Car';'Approve';'Safe';'Grade';'Male';'Firstage';'Loose'};
% Enter odds ratio and margins of error
A=[1   5.118  1.308 1.756  % variable, pt est, lo marg, up marg
   2   2.938  0.796 1.091
   3   1.832  0.513 0.714
   4   1.252  0.147 0.166
   5   1.013  0.237 0.308
   6   0.884  0.052 0.055
   7   0.489  0.123 0.165];
% Enter axes limits as [xmin xmax ymin ymax]
V=[0.5 7.5 -0.15 7.15];
% NO MORE INPUT IS NEEDED ==================================
% Location of labels may need adjustment.
varlabel=A(:,1); % variables
oddsrato=A(:,2);  % pt est for odds ratio from sas output
lomarger=A(:,3);  % margin of error to be subtracted from pt est
upmarger=A(:,4);  % margin of error to be added to pt est
nvar=length(varlabel);
oneliney=[1;1];
onelinex=[0;V(2)];
labLocy=(oddsrato-lomarger)-0.2;
labLocx=varlabel-0.25;
figure (1)
errorbar(varlabel,oddsrato,lomarger,upmarger,'bo')
 hold on;
plot(onelinex,oneliney,'r'),axis(V);
hold off;
ylabel('Odds Ratio','Fontsize',11);
xlabel('Variable #','Fontsize', 11);
```

```
set(gca, 'xtick', varlabel')
% annotation
for i=1:nvar;
text(labLocx(i,1),labLocy(i,1),VarName(i,1))
end;
end;
```

Matlab Program for Odds Ratio Plot from Model 3

```
% Title: logitgram3.m
% This program charts the point estimates of incremental
% odds ratios and 95% confidence intervals corresponding to
% predictor variables in logistic regression model (3).
% INPUT NEEDED ============================================
% Enter variables names in order of descending odds ratio.
% Names of variables should go in quotes and separated by ;
VarName={'Car';'Safe';'Friends';'Cool';'White';'Home';'Grade';
 'Firstage';'Male';'Harm'};
% Enter odds ratio and margins of error
A=[1   2.470  0.398 0.476  % variable, pt est, lo marg, up marg
   2   2.306  0.403 0.489
   3   1.869  0.098 0.103
   4   1.691  0.317 0.389
   5   1.146  0.208 0.254
   6   1.114  0.153 0.177
   7   1.089  0.071 0.076
   8   0.962  0.031 0.032
   9   0.921  0.123 0.143
   10  0.659  0.153 0.198];
% Enter axes limits as [xmin xmax ymin ymax]
V=[0.25 10.75 0.15 3.15];
% NO MORE INPUT IS NEEDED ==================================
% Location of labels may need adjustment.
varlabel=A(:,1); % variables
oddsrato=A(:,2);  % pt est for odds ratio from sas output
lomarger=A(:,3);  % margin of error to be subtracted from pt est
upmarger=A(:,4);  % margin of error to be added to pt est
nvar=length(varlabel);
oneliney=[1;1];
onelinex=[0;V(2)];
```

```
labLocy=(oddsrato-lomarger)-0.125;
labLocx=varlabel-0.4;
figure (1)
errorbar(varlabel,oddsrato,lomarger,upmarger,'bo')
 hold on;
plot(onelinex,oneliney,'r'),axis(V);
hold off;
ylabel('Odds Ratio','Fontsize',11);
xlabel('Variable #','Fontsize',11);
set(gca, 'xtick', varlabel')
% annotation
for i=1:nvar;
text(labLocx(i,1),labLocy(i,1),VarName(i,1))
end;
end;
```

.4 Programs for Loglinear Models of Counts from 2×2 Tables

Matlab Program for Saturated Model of Counts

```
% Title:  SatLifeHome.m
% Performs Loglinear Modeling of 2 x 2 Table
% Saturated model (Agresti p.132)
% Binary variables: Ever smoked (Life) (rows: 1 = no, 2 = yes)
% versus Someone at home smokes (Home) (cols: 1 = no, 2 = yes)

% Labels
a = '2 x 2 contingency table of observed counts:';
c = '2 x 2 table of log(counts):';
d = 'Parameters for saturated loglinear model:';
e = 'Expected counts derived from the model:';

% Description of two-way table
h = 'Row variable is Lifetime cigarette use';
i = 'row 1 = no, row 2 = yes';
j = 'Column variable is Someone at home smokes';
k = 'col 1 = no, col 2 = yes';
l = 'Bivariate Loglinear Saturated Model';
m = '****************************************';
disp(l)
disp(h)
disp(i)
disp(j)
disp(k)
disp(m)

% Data and computations
I = 2;  % number of rows (Life)
J = 2;  % number of columns (Home)
% Enter Obs = observed counts
disp(a)
Obs = [1771 1000; 2798 3592]   % matrix of counts
X = [1  1  1  1
     1  1 -1 -1
     1 -1  1 -1
     1 -1 -1  1]; % indicator matrix
disp(c)
```

```
eta = log(Obs) % ln(observations)
reta = eta*ones(J,1);   % row margin ln(obs)
ceta = eta'*ones(I,1); % col margin ln(obs)
etaidot = reta/J; % row mean ln(obs)
etadotj = ceta/I; % col mean ln(obs)
etaddot = sum(sum(eta))/(I*J); % grand mean
Llife = etaidot(1)-etaddot;  % lambda^X1
Lhome = etadotj(1)-etaddot; % lambda^X2
Llxh = eta(1)-etaidot(1)-etadotj(1)+etaddot; % lambda^X1X2
disp(d)
Pars = [etaddot; Llife; Lhome; Llxh]% parameters
disp(e)
Expected = exp(X*Pars)
```

Matlab Program for Independence Model of Counts

```
% Title: IndLifeHome.m
% Performs Loglinear Modeling of 2 x 2 Table
% Independence model (Agresti p.131)
% Binary variables: Ever smoked (Life) (rows: 1 = no, 2 = yes)
% versus Someone at home smokes (Home) (cols: 1 = no, 2 = yes)

% Labels
a = '2 x 2 contingency table of observed counts:';
b = 'The sum of cell frequencies:';
c = '2 x 2 table of relative frequencies:';
d = 'Parameters for loglinear model:';
e = 'Expected counts derived from the model:';
f = 'Pearson residuals:';
g = 'Chi-square statistic:';

% Description of two-way table
h = 'Row variable is Lifetime cigarette use';
i = 'row 1 = no, row 2 = yes';
j = 'Column variable is Someone at home smokes';
k = 'col 1 = no, col 2 = yes';
l = 'Bivariate Loglinear Independence Model';
m = '****************************************';
disp(l)
disp(h)
disp(i)
disp(j)
disp(k)
disp(m)


% Data and computations
I = 2;  % number of rows (Life)
J = 2;  % number of columns (Home)
% Enter Obs = Obc = observed counts
disp(a)
Obs = [1771 1000; 2798 3592]  % matrix of counts
Obc = [1771; 1000; 2798; 3592];% column vector of counts
disp(b)
```

```
n = sum(sum(Obs))  % sum of counts
disp(c)
P = Obs/n   % matrix of cell proportions
lP = log(P);  % matrix of ln(proportions)
Ri = P*ones(J,1);  % row masses
Cj = P'*ones(I,1);  % col masses
lRi = log(Ri);      % ln(row masses)
lCj = log(Cj); % ln(col masses)
lRbar = sum(lRi)/2;  % mean ln(row masses)
lCbar = sum(lCj)/2; % mean ln(col masses)
Llife = lRi(1)-lRbar; % lambda^X1
Lhome = lCj(1)-lCbar; % lambda^X2
mu = log(n)+lRbar+lCbar; % constant
disp(d)
Pars = [mu; Llife; Lhome] % parameters
X = [1 1 1
   1  1 -1
   1 -1  1
   1 -1 -1]; % indicator matrix
disp(e)
Ex = exp(X*Pars) % expected counts
SqDiff = (Obc-Ex).*(Obc-Ex);
disp(f)
Res = SqDiff./Ex % residuals
disp(g)
Chisq = sum(Res) % chi square
```

.5 Programs for Loglinear Models of Multi-way Tables

SAS Programs for Loglinear Models

```
proc catmod;
model carsmoke*hmfrien*somehome=_response_/ noresponse noiter noparm;
loglin carsmoke|hmfrien|somehome;
run;
```

.6 Program for Simple Correspondence Analysis

Matlab Program for Simple Correspondence Analysis

```
% Title:  DaysFriends.m
% Performs Correspondence Analysis for r x c Table
a = '7 x 5 contingency table:';
b = 'The sum of cell frequencies:';
c = 'Correspondence matrix of relative frequencies:';
d = 'Row masses:';
e = 'Column masses:';
f = 'Row Profiles:';
g = 'Column Profiles:';
h = 'The centered and scaled correspondence matrix:';
i = 'Singular Value Decomposition of Ps:';
j = 'S is the diagonal matrix of singular values.';
k = 'Coordinates of the row profiles:';
l = 'Coordinates of the column profiles:';
m = 'Total inertia (chi_square/n):';
q = 'Inertia Component by Ascending Dimension:';
r = 'Proportion of Chi-Square Statistic by Dimension:';
t = '_____________________________________________________________________';
v = 'Correspondence Analysis for A Two-Way Table:';

% Description of two way table
w = 'Row variable is STUDENT SMOKED HOW MANY DAYS IN PAST 30';
x = 'index i:  1=0, 2=1-2, 3=3-5, 4=6-9, 5=10-19, 6=20-29, 7=ALL 30 DAYS';
y = 'Column variable is HOW MANY OF 4 CLOSEST FRIENDS SMOKE';
z = 'index j:  1 = NONE, 2 = ONE, 3 = TWO, 4 = THREE, 5 = FOUR';

% Labels for row points and column points
str1(1)={'0 days'};
str1(2)={'1-2 days'};
str1(3)={'3-5 days'};
str1(4)={'6-9 days'};
str1(5)={'10-19 days'};
str1(6)={'20-29 days'};
str1(7)={'all 30 days'};
str2(1)={'0 friends'};
str2(2)={'1 friend'};
str2(3)={'2 friends'};
```

```
str2(4)={'3 friends'};
str2(5)={'4 friends'};

%Display Heading
disp(v)
disp(t)
disp(w)
disp(x)
disp(y)
disp(z)
disp(t)

%===============================================================
% NOTE: Enter correct cell frequencies for N,
% the number of rows for I, and number of columns for J below.
disp(a)
N  = [2909 1193 802 348 314
    77 110 151 71 96
    33 64 80 61 38
    33 38 60 42 49
    27 52 71 62 108
    20 41 89 89 134
    88 102 184 238 537]       % I x J contingency table of counts
I  = 7; % nbr of rows
J  = 5; % nbr of columns
%===============================================================

disp(b)
n  = sum(sum(N)) % sum of cell counts
disp(c)
P  = (1/n)*N % correspondence matrix
disp(d)
ri = P*ones(J,1)        % row masses
disp(e)
cj = P'*ones(I,1) % column masses
Dr = diag(ri); % r_weight matrix
Dc = diag(cj); % c_weight matrix
disp(f)
R  = inv(Dr)*P % row profiles
Ct = inv(Dc)*P';
```

```
disp(g)
C  = Ct'% column profiles
Pc = P-(ri*cj'); % centered corr matrix
disp(h)
Ps = Dr^(-1/2)*Pc*Dc^(-1/2) % scaled corr matrix
disp(i)
disp(j)

% S is the diagonal matrix of singular values.
[U, S, V] = svd(Ps) % singular value decomposition
Ut = Dr^(1/2)*U % Set U~
Vt = Dc^(1/2)*V  % Set V~
disp(k)
Y  = inv(Dr)*Ut*S % row profiles coordinates
disp(l)
Z  = inv(Dc)*Vt*S'% column profiles coordinates
disp(m)
TI  = trace(inv(Dr)*Pc*inv(Dc)*Pc')
disp(q)
S_sq = S.*S; % inertia by ascending axes
Inertia = [S_sq(1:I+1:J*(I+1))] % pick diagonal elements of A
disp(r)
Pchi = Inertia./TI % Proportion of chi_sq by asc. axes

% Plotting the graph
rprcoor1=Y(:,1);
rprcoor2=Y(:,2);
cprcoor1=Z(:,1);
cprcoor2=Z(:,2);
rowsrows=length(rprcoor1);
rowscols=length(cprcoor1);
coor1ma1=max(rprcoor1);
coor1ma2=max(cprcoor1);
coor1max=max(coor1ma1,coor1ma2);
coor1mi1=min(rprcoor1);
coor1mi2=min(cprcoor1);
coor1min=min(coor1mi1,coor1mi2);
coor2mi1=min(rprcoor2);
coor2mi2=min(cprcoor2);
coor2min=min(coor2mi1,coor2mi2);
```

```
coor2ma1=max(rprcoor2);
coor2ma2=max(cprcoor2);
coor2max=max(coor2ma1,coor2ma2);
horlin=[coor1min;coor1max];
zhorlin=[0;0];
verlin=[coor2min;coor2max];
zverlin=[0,0];
plot(cprcoor1,cprcoor2,'ro','MarkerSize', 3);
axis([coor1min-0.1 coor1max+0.1 coor2min-0.1 coor2max+0.1]);
hold on
plot(rprcoor1,rprcoor2,'b*','MarkerSize', 3);
plot(horlin,zhorlin,':');
plot(zverlin,verlin,':');

% Annotation
xlabel('Dimension 1','Fontsize',10);
ylabel('Dimension 2','Fontsize',10);
text(rprcoor1(1)-.05,rprcoor2(1)+0.02, [str1(1)],'Fontsize',8);
text(rprcoor1(2)-.05,rprcoor2(2)+0.02, [str1(2)],'Fontsize',8);
text(rprcoor1(3)-.05,rprcoor2(3)+0.02, [str1(3)],'Fontsize',8);
text(rprcoor1(4)-.05,rprcoor2(4)+0.02, [str1(4)],'Fontsize',8);
text(rprcoor1(5)-.05,rprcoor2(5)+0.035,[str1(5)],'Fontsize',8);
text(rprcoor1(6)-.05,rprcoor2(6)+0.02, [str1(6)],'Fontsize',8);
text(rprcoor1(7)-.05,rprcoor2(7)+0.02, [str1(7)],'Fontsize',8);
text(cprcoor1(1)-.05,cprcoor2(1)-0.025,[str2(1)],'Fontsize',8);
text(cprcoor1(2)-.05,cprcoor2(2)-0.025,[str2(2)],'Fontsize',8);
text(cprcoor1(3)-.05,cprcoor2(3)-0.025,[str2(3)],'Fontsize',8);
text(cprcoor1(4)-.05,cprcoor2(4)-0.025,[str2(4)],'Fontsize',8);
text(cprcoor1(5)-.05,cprcoor2(5)-0.025,[str2(5)],'Fontsize',8);
hold off
```

.7 Multiple Correspondence Analysis Programs

SAS Code for MCA of Smoking Status, Gender, Grade, and Ethnicity, figure 5

```
proc corresp short mca outc=coords;
tables la30smok male grade white;
run;
```

SAS Code for MCA of Smoking Status, Car, Safe, Cool, and Friends, figure 6

```
proc corresp short mca outc=coords;
tables la30smok carsmoke safe cool hmfriend;
run;
```

SAS Code for MCA of Student's Experimentation with Different Types of Tobacco and Number of Closest Friends that Smoke Cigarettes, figure 7

```
proc corresp short mca outc=coords;
tables esmoke evrcigar evchsndp evrbidis hmfriend;
run;
```

VITA

DEBORAH SUSAN HOSLER

Address:

793 Harbor Springs Rd.
Kingsport, TN 37664

Education:

Miami University, Oxford, OH (B.S., Home Economics, 1979)
East Tennessee State University (ETSU), Johnson City, TN (B.S., Math, 2000)
ETSU (M.S., Mathematical Sciences Mathematical Statistics, 2002)

Experience:

Graduate Assistant, ETSU, Department of Mathematics, 2000–2002
Mathematics Tutor, ETSU, Office of Special Programs, 1995–2000

Organizations:

American Statistical Association
Phi Kappa Phi
Kappa Mu Epsilon
Mathematical Association of America

Honors:

Outstanding Graduate Assistant, 2002
Mathematics Faculty Award, 2000