Answer to Sample Question

This dataset comes from a CBS survey of 1246 respondents in the High School graduating class of 2000, conducted in December 1998. Thus, the respondents are high school juniors, not seniors. The sampling frame is not documented. However, from the codebook, articulating the state in which the respondent resides, it can be deduced that this is a national sample. Moreover, despite the attempt to codify region, state and urban/rural settings, it is using a weighting technique to correct for under/over sampling of the units. Therefore, this can best be described as a convenience sample, attempting to use the techniques of a disproportionate stratified sample of students, in an attempt to achieve external validation.

The questions were of three types: 1) those that attempted to assess feelings concerning social problems confronting students, such as racial and gender prejudice, as well as questions about education and aspirations, sex, pregnancy and abortion, drugs, alcohol and smoking, cheating, stealing and lying, happiness, self-worth and suicide; 2) those that that attempted to assess demographic factors, such as age, sex, # of siblings, religion, year in school, etc.; and finally 3) those that attempted to assess information about the respondents and their context, such as state and region of residence, living arrangement, parental relationships, use of technology, and sex/drug/cigarette/diet/alcohol related behavior.

All of these questions appear to be discrete and largely non-quantitative. That is, there were no apparent composite variables, either scale or index, and no apparent distinction made between independent and dependent variables. However, the defining characteristic of the survey, that is, the thread among the questions, appears to be the degree to which social problems confronting American teenagers manifest themselves in certain behavioral patterns.

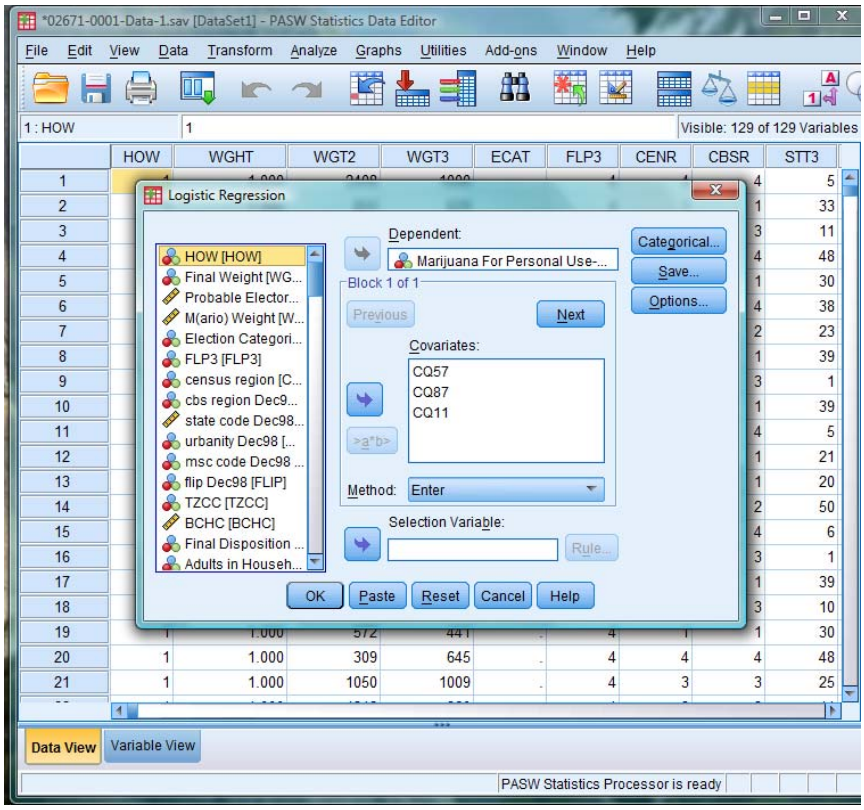
Therefore, the most likely findings one would look for in these data, would be the relationship between demographic factors as antecedent variables (e.g. age, sex and race), certain behavior patterns

as intervening variables (e.g. parental relations, occupational aspirations, GPA, etc) thoughts, feelings and other, more anti-social, behavior as dependent variables (discrimination, sex, drugs and alcohol use, contemplating suicide). For example, are females who have high GPA, college aspirations and live at home, less likely to become pregnant or have friends who are?

Combining the apparent lack of stratified random sampling and the measurement being largely non-quantitative, the analysis of these data appear to be restricted to multivariate models dealing with frequency analysis. That is, these data seemingly do not lend themselves to the models of latent structure analysis based on correlations among quantitative variables (e.g. factor and cluster analysis). Rather it would appear they are candidates for log linear and particularly logistic regression models.

Accordingly, and as an example for this exercise, I will attempt to analyze the following independent variables. I want to know if how often you smoke marijuana (question 57), whether you have a job (question 87) and how you rate your relationship with your parents (question 11) predict the dependent variable, whether you favor or oppose legalizing the possession of marijuana for personal use (question 58). All variables were transformed to exclude non-responses.

To accomplish this analysis, a logistic binary regression was used (i.e. the DV is dichotomous). In this technique, the ability of a unit movement in the independent variables is assessed to see their effect on the change in the ratio of the probabilities of dichotomous outcomes of the dependent variable. More particularly, whether being a marijuana smoker, having no job and having a bad relationship with your parents is likely to affect the probability of being in favor of legalizing marijuana. This type of analysis is the most flexible in this context as it: 1) doesn't require quantitative independent variables; 2) does not preclude correlated terms; 3) does not require an underlying normal distribution on the variables. Therefore, using SPSS, I have selected "analyze", "regression" and "binary logistic" as the procedure (see screen below).



The results, predicated on the two important output tables, indicate that these variables are significant predictors of attitudes toward legalizing marijuana. Yet while they reproduce results of who would be expected to oppose the legalization, overall they do not reproduce as significantly who is likely to favor it. On average, 77% of the cases were correctly classified. However, this is because 97% of the cases opposing were correctly classified, while only 29% of those favoring were correctly classified.

Classification Table[a]

Observed			Predicted		
			Marijuana For Personal Use-Dec98		Percentage Correct
			Favor	Oppose	
Step 1	Marijuana For Personal Use-Dec98	Favor	101	252	28.6
		Oppose	28	818	96.7
	Overall Percentage				76.6

a. The cut value is .500

More particularly, the effects table indicates that smoking marijuana (q57) strongly predicts the probability of favoring, while having a strong relationship with your parents predicts the tendency to oppose. Both are significant (p<.001). Having a job seems to have little effect and is not significant.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CQ57	.838	.082	104.255	1	.000	2.313
	CQ87	.011	.145	.006	1	.940	1.011
	CQ11	-.451	.098	21.221	1	.000	.637
	Constant	-2.167	.471	21.163	1	.000	.115

a. Variable(s) entered on step 1: CQ57, CQ87, CQ11.

In short, this dataset is largely a product of the desire to have tablular data on high school student feelings and behaviors toward social problems and the demographics associated therewith. This is typical of media studies, where pivot tables are all that are presented. However, as can be seen from this cursory analysis, much richer data mining can be utilized. Given the time, it would be useful to develop scales and indexes, and to transform these into dummy variables such that more latent structure can be realized from these questions.