Biostatistics 200B – Project Fall 2013 – CTS Dataset

Matthew Theisen

Introduction

The data for analysis is a survey of youth aged 12-17 in California regarding smoking habits and other life data. The data for this analysis is a subset which only includes results from subjects which reported having smoked. The outcome of interest is a binary variable which represents whether or not the subject has smoked every day for a month. Since the outcome is binary, it is natural to use a logistic regression framework.

The dataset contains 686 units with 17 variable entries. One variable is for ID# and one variable is for the outcome ('cigmonth'). An additional variable representing whether or not the subject had ever smoked was 'Yes' for all units, leaving 14 covariates. Analysis was carried out in R. Six entries out of the total (17*686 = 11,662) were reported as "don't know". For purposes of analysis, these were all changed to 'No'.

Analyses

The first portion of my analysis is to look at the mean of the outcome variable based on cross-categorization by each of the covariates. This section of the analysis doesn't provide information about statistical significance but can be instructive for further analysis. The results of this first analysis are presented in Table 2.

Based on this analysis, it is logical to collapse the five levels of rdaganst into 2. For rdaganst, code 5 and 3 collapse into 1 based on the similarity of their means and the small number of samples. 4 can collapse into 2 based on the similar description ("a few" vs. "no" commercials). Similar processes were carried out for bdaganst and tvaganst. Some additional categories were collapsed as follows based on small size of some categories and/or exchangibility based on probability of cigmonth being true. Ethnic 3 & 5 were collapsed into 1 and 4 was collapsed into 2. Tvaganst 5 & 4 were collapsed into category 3. Bdaganst 3 was collapsed into 2. Religsvc 2 & 3 were collapsed into 1. Additionally, cigmonth 2 ("no") was recategorized as 0 so the results of the logistic regression would be more intuitive to understand.

The next step of the analysis was to carry out marginal logistic regressions. Using this approach, an estimate of statistical significance can be obtained for the effect of each predictor variable. For this step, it is necessary to decide whether to code non-binary predictors as a 'factor' or numerical value. For example, it is appropriate to treat age as a numerical value since it is reasonable to expect that the effect of age varies in a continuous manner. However, for many of the other predictor variables ('rdaganst' for example) it isn't reasonable to expect the 1-5 coding of the surveys to vary in some continuous manner with the effect on the outcome variable. Thus, this and other similar covariates are coded as 'factors' – where the effect of each level is evaluated as a separate predictor. For binary predictors, it isn't necessary to make this distinction since the analysis would be identical in either case.

Using the results provided by the analysis of the marginal logistic regressions, it can be used to guide to determine where main effects and interactions may exist. It is a general heuristic of statistical analysis that higher order interactions are accompanied by lower order ones – so it is logical to seek out interactions between factors which have significant main effects. However, many models were tested but no significant interactions were found. Using that principle, and employing likelihood ratio tests to accept or reject candidate models, the following model was developed.

Constructing a Model

The model was selected after finding the model whose parameters were significant.

```
logit <- glm(cigmonth ~ age + ethnic + parntsmk + tmsport + rdaganst +
bdaganst, data = mydata, family = "binomial")
```

The predictors had their levels collapsed as previously described. The deviance parameter for the logistic regression without collapsing the predictors was 636.26 with 14 parameters. After collapsing the categories, the deviance parameter was 630.46 with 7 parameters. A lower deviance indicates a higher likelihood for the model, implying that the model with collapsed categories had a higher likelihood even though it had a lower number of parameters.

Hosmer – Lemeshow Goodness of Fit Test

The Hosmer-Lemeshow goodness of fit test is available as a function in R in the ResourceSelection package. The test requires the observed outcome and expected value as arguments. To calculate the expected value in r, the predict function can be used – for example:

```
mydata$predictP <- predict(logit, newdata = mydata, type = "response")
```

The code for the corresponding HL-GOF test is then the following:

```
hoslem.test(mydata$cigmonth, mydata$predictP, g = 10)
```

The result of the HL-GOF test is a chi-squared value of 10.1 on 8 degrees of freedom for p=0.2581. This result indicates that the test doesn't reject the model.

Results of Model

The model yields the following table:

Table 1 – Odds ratios and 95% confidence intervals associated with selected covariates for the probability of having smoked a cigarette every day for a month

Factor	OR	95% CI
Age (1 year increase)	1.24	(1.04, 1.48)
Black, Hispanic or other ethnicity	1.00	--
White or Asian/Pacific Islander	2.09	(1.36, 3.23)
Parents who smoke	0.42	(0.28, 0.62)

No participation in a team sport	2.00	(1.34, 2.98)
Heard a lot of radio commercials against smoking	1.00	--
Heard a few or no radio commercials against smoking	0.41	(0.26, 0.67)
Saw a lot of billboards against smoking	1.00	--
Heard a few or no billboards against smoking	2.38	(0.95, 5.98)

Table 2 – Table of means of cigmonth value – Data reported as follows – variable level (count): fraction with cigmonth = 1

age	12 (8): 0.2500	13 (28): 0.0714	14 (66): 0.2121	15 (113): 0.1327	16 (208): 0.2115	17 (263): 0.2509
sex	1 (370): 0.1811	2 (316): 0.2405				
ethnic	1 (198): 0.1262	2 (367): 0.2588	3 (24): 0.1667	4 (50): 0.2400	5 (47): 0.1489	
parntsmk	1 (281): 0.2989	2 (405): 0.1456				
Tvaganst	1 (288): 0.2188	2 (316): 0.1867	3 (10): 0.3000	4 (53): 0.2453	5 (19): 0.2632	
Rdaganst	1 (57): 0.3508	2 (288): 0.1632	3 (28): 0.3214	4 (288): 0.2048	5 (25): 0.3200	
bdaganst	1 (57): 0.1053	2 (415): 0.2120	3 (214): 0.2289			
famargue	1 (292): 0.2157	2 (394): 0.2030				
unhappy	1 (174): 0.2528	2 (239) 0.1966	3 (171): 0.1812	4 (102): 0.2059		
nofuture	1 (136): 0.2279	2 (135): 0.200	3 (154): 0.2208	4 (261): 0.1954		
steadyfd	1 (558): 0.2258	2 (128): 0.1328				
tmsport	1 (362): 0.1464	2 (324): 0.2778				
talkhelp	1 (652): 0.2116	2 (34): 0.1471				
religsvc	1 (160): 0.1750	2 (138): 0.1811	3 (181): 0.1989	4 (207): 0.2609		