Targeting Non-users of Contraception Among Sexually-active Female Teens

Haylin Belay, Michael Discenza, and Andrew Pinelli

March 29, 2013

Introduction and Purpose

For this project, we analyzed secondary data from the Data Archive on Adolescent Pregnancy and Pregnancy Prevention (DAAPPP) in the Social Science Electronic Data Library. The study was conducted in the mid-1980s. Researchers surveyed adolescents visiting a San Francisco health clinic regarding their decision-making about contraceptive use. The study measured general attitudes towards contraceptive methods, social expectations about contraceptive use (or disuse), intentions to use (or not use) contraception, and sexual behavior and decision-making. Our goal was to investigate whether the data could be used to identify what behaviors, attitudes, and demographic characteristics are predictors for unprotected sex. Specifically, we wanted to focus on employing predictors in our model that could be used to target public health interventions that could spur higher contraceptive usage among populations that are at a high risk of teenage pregnancy. These interventions could take the form of ad campaigns, revised health curricula, and increased access to contraception at local clinics.

Data Preparation

In our analysis, we used only a subset of the available data. We narrowed our focus to female, sexually active teens. This seemed to be the ideal population given that more females than males were present in the data set, certain questions in the survey pertained more to women than men, and survey directions were ambiguous for male subjects. Moreover, because the survey instrument included many detailed questions about the nature and duration of contraception use, it seems more likely that female subjects would have a more complete view of their personal history.

The raw data set contained 1016 variables. A significant portion of these variables related to attitudes toward specific methods of contraception and extremely detailed use history. Given our purpose of targeting intervention for at-risk females, the aforementioned minutia would be of little use. Thus we chose to examine only 16 variables pertaining to 234 sexually active female subjects that are observable outside of clinical context.

```
##  [1] "sex_wo_contraception"                         "age"
##  [3] "ethnicity"                                    "religion"
##  [5] "highest_yr_school_completed"                  "know_preg_peera"
##  [7] "fathers_education"                            "mothers_education"
##  [9] "educational_aspiration"                       "dont_participate_in_school_activites_1"
## [11] "smoking_freq"                                 "drinking_freq"
## [13] "drug_freq"                                    "age_frist_drinking"
```

Preparing the dataset for analysis required significant data transformation and cleaning. We found duplicates of the selected 16 variables of interest. These duplicate variables were present because the dataset was assembled from a number of different survey instruments. Some subjects had responses for items in only one of the duplicate variables and others for both. In order to assemble our set of complete and atomic features for our analysis, we needed to combine these variables and did so on a case by case basis.

The documentation concerning the coding of the data was rather incomplete, so we needed to piece together an understanding of the meaning of the raw data using an incomplete data dictionary and examining the survey instruments themselves. In particular, ″NAs,″ which were encoded as ″-9″ in the raw data, posed a problem in the following predictors: father′s education, mother′s education, educational aspiration, drinking frequency, drug use frequency, and smoking frequency. We had to make assumptions regarding the nature of these ″NAs.″ For father′s education and mother′s education, the survey did not include an option for subjects to indicate that they did not have or know either of their parents or their parents′ level of educational attainment. As a result, we included the ″NAs″ as a category of that variable, as it may have been indicative of their underlying family circumstance.

The frequency of substance use variables represented a different problem in that, each of these questionsallowed a response category of ″never.″ For the vast majority of subjects, ″NAs″ were present in each of these three variables indicating that their survey instruments did not contain these questions. We did not want to throw out these observations given the small size of our dataset, so we coded them as a distinct category. We were careful not to use any of the ″NA″ categories as reference categories because we did not know their true composition.

Upon examining the 16 variables′ distributions of responses, we noticed that many categories had a very low frequency. Additionally, for many of the questions, there were 6-10 possible response categories, making regression coefficients highly numerous and difficult to interpret. Thus, we decided to recode some of the predictors so that they had categories with larger response rates that were more easily interpretable. For fathers education, mothers education, and educational aspiration, we regrouped ten categories into five: did not graduate high school, graduated highschool or some college, graduated college or schooling beyond college, other schooling after high school (i.e. trade school), and no response. Regarding drinking, drug, and cigarette frequency, we regrouped seven categories into three: abstain, partake, and no response.

All source code pertaining to the construction of the dataframe that we used for analysis can be found in the files ″constructing_dataset.r″ and ″recoding_ distributions.r.″

Model Fitting

[Note that in this section, we do not follow what would be the ideal procedure, but retrace our actual data analysis procedure to recreate the path to our own insights about the data]

Our practical goal was, through examining various logistic models, not necessarily to pick the model with the best statistical properties, but to gain an understanding of how to target high-risk populations by only selecting demographic features that could predict whether someone had or had not engaged in sex without any conraception. Because the purpose of the models we fit was to predict a binary response variable, we used a logistic regression model. We were not particularly interested in predictive accuracy, but rather in the significance and size of coefficients in the regression, which would be able to guide targeting efforts to at-risk populations. This consideration led us to focus exclusively on parametric models. In light of this, we did not run a generalized additive model, as it is semi-parametric in nature nor use any of the non-parametric tree-based methods. While these models are useful for predictive purposes in determining the risk level for sex without contraception on an individual level, they are not useful in helping us understand the aggregate the dynamics of contraception use in young females.

LASSO and Full Model Fitting

The first approach to model fitting and variable selection that we took was using LASSO regression to fit a logistic model. LASSO is a shrinkage and selection method for generalized linear regression that minimizes the sum of squared errors subject to bound on the absolute values of the coefficients. This method allows for a variable to be partly included in the model. We chose LASSO over a Ridge regression as the values of coefficients are driven to zero more quickly in Lasso, building the model selection process into the analysis.

```
library(glmnet)
# to use glmnet, we need to recode the factor variables as binary
full_model <- sex_wo_contraception ~ age + ethnicity + religion + highest_yr_school_completed +
    know_preg_peera + fathers_education + mothers_education + educational_aspiration + dont_participate_
    smoking_freq + drinking_freq + drug_freq + age_frist_drinking
yc <- as.matrix(study_data_converted[, 1])
x_matc <- model.matrix(full_model, data = study_data_converted)
# running
glmnetModelc_lasso <- glmnet(x_matc, yc, family = "binomial", alpha = 1)
# plot(glmnetModelc_lasso)
my.cvc <- cv.glmnet(x_matc, yc, family = "binomial")
predict(glmnetModelc_lasso, type = "coef", s = my.cvc$lambda.min)
```

```
## 36 x 1 sparse Matrix of class "dgCMatrix"
##                                                  1
## (Intercept)                                0.5613
## (Intercept)                                .     
## age                                        .     
## ethnicity2                                 .     
## ethnicity3                                 .     
## ethnicity4                                 .     
## ethnicity5                                 .     
## ethnicity6                                 .     
## ethnicity7                                 .     
## religion1                                  .     
## religion2                                  .     
## religion3                                  .     
## religion4                                  .     
## religion5                                  .     
## highest_yr_school_completed                .     
## know_preg_peera2                           .     
## fathers_educationgrad.college.beyond       .     
## fathers_educationgrad.hs.some.college      .     
## fathers_educationno.response               .     
## fathers_educationother.schooling           .     
## mothers_educationgrad.college.beyond       .     
## mothers_educationgrad.hs.some.college      .     
## mothers_educationno.response               .     
## mothers_educationother.schooling           .     
## educational_aspirationgrad.college.beyond  .     
## educational_aspirationgrad.hs.some.college .     
## educational_aspirationno.response          .     
## educational_aspirationother.schooling      .     
## dont_participate_in_school_activites_12    .     
## smoking_freqno_response                    .     
## smoking_freqpartake                        .     
## drinking_freqno_response                   .     
## drinking_freqpartake                       .     
## drug_freqno_response                       .     
## drug_freqpartake                           .     
## age_frist_drinking                         .     
```

```
glmnetModelc_lasso$nulldev
```

```
## [1] 306.7
```

Though the LASSO algorithm would have been an ideal solution for automated variable selection, it did not select any variables so we needed to take another approach to modeling the data. We proceeded to fit a full logistic regression model that included all of our predictors.

```
lr1c <- glm(full_model, data = study_data_converted, family = binomial(link = logit))
summary(lr1c)
```

```
##
## Call:
## glm(formula = full_model, family = binomial(link = logit), data = study_data_converted)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.052  -1.125   0.691   0.917   1.669
##
## Coefficients: (1 not defined because of singularities)
##                                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                -2.01e+01   1.46e+03   -0.01    0.989
## age                                         1.01e-01   1.14e-01    0.89    0.376
## ethnicity2                                  3.19e-01   4.50e-01    0.71    0.478
## ethnicity3                                  9.36e-01   6.13e-01    1.53    0.127
## ethnicity4                                 -1.59e+01   1.46e+03   -0.01    0.991
## ethnicity5                                  8.19e-01   1.09e+00    0.75    0.454
## ethnicity6                                  1.09e+00   9.13e-01    1.20    0.231
## ethnicity7                                  6.97e-01   5.83e-01    1.19    0.232
## religion1                                   1.61e+01   1.46e+03    0.01    0.991
## religion2                                   1.55e+01   1.46e+03    0.01    0.992
## religion3                                   1.63e+01   1.46e+03    0.01    0.991
## religion4                                   1.65e+01   1.46e+03    0.01    0.991
## religion5                                   1.64e+01   1.46e+03    0.01    0.991
## highest_yr_school_completed                 1.79e-02   2.55e-02    0.70    0.484
## know_preg_peera2                            6.51e-01   9.23e-01    0.71    0.481
## fathers_educationgrad.college.beyond        5.33e-01   5.35e-01    1.00    0.319
## fathers_educationgrad.hs.some.college       5.33e-01   5.09e-01    1.05    0.295
## fathers_educationno.response                2.12e-01   7.51e-01    0.28    0.778
## fathers_educationother.schooling            8.89e-01   8.44e-01    1.05    0.292
## mothers_educationgrad.college.beyond       -2.44e-01   5.69e-01   -0.43    0.667
## mothers_educationgrad.hs.some.college       2.62e-01   5.05e-01    0.52    0.604
## mothers_educationno.response               -1.63e+01   1.46e+03   -0.01    0.991
## mothers_educationother.schooling            2.48e-02   7.61e-01    0.03    0.974
## educational_aspirationgrad.college.beyond   7.75e-01   1.11e+00    0.70    0.486
## educational_aspirationgrad.hs.some.college  1.15e+00   1.14e+00    1.01    0.311
## educational_aspirationno.response           1.57e+01   1.46e+03    0.01    0.991
## educational_aspirationother.schooling       9.84e-01   1.30e+00    0.76    0.450
## dont_participate_in_school_activites_12     1.45e-01   3.46e-01    0.42    0.674
## smoking_freqno_response                    -5.68e-02   1.43e+00   -0.04    0.968
## smoking_freqpartake                         9.54e-01   4.26e-01    2.24    0.025 *
## drinking_freqno_response                    9.62e-01   1.45e+00    0.66    0.509
## drinking_freqpartake                        2.29e-01   4.49e-01    0.51    0.611
## drug_freqno_response                              NA         NA      NA       NA
## drug_freqpartake                            5.51e-03   4.56e-01    0.01    0.990
## age_frist_drinking                          8.02e-03   1.36e-02    0.59    0.554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 306.66  on 233  degrees of freedom
## Residual deviance: 276.92  on 200  degrees of freedom
## AIC: 344.9
##
## Number of Fisher Scoring iterations: 14

# explained deviance:
(lr1c$null.deviance - lr1c$deviance)/lr1c$null.deviance

## [1] 0.097
```

Exhaustive Search 1: All Levels of Factor Variables

The only statistically significant coefficient that this model fit was for partaking in smoking. The explained deviance of this model is 0.097, indicating that about 10 percent of the variation in the dependent variable is explained by the model. A model such as this, which had a large number of insignificant predictors is far from ideal so we needed to use other methods to select a smaller number of variables for inclusion.

One such method we explored was exhaustive search. We fit 2048 (2 to the 11) models of predictors that seemed to, based on the results of the full model, have some capacity to explain the variation in our response variable. We then sorted these models by their AIC. The top six resulting models are shown below:

	Model	AIC	Explained.Deviance
1	y ˜ smoking_freq	305.77	0.02
2	y ˜ age+smoking_freq	306.54	0.03
3	y ˜ highest_yr_school_completed+smoking_freq	307.30	0.02
4	y ˜ know_preg_peera+smoking_freq	307.43	0.02
5	y ˜ dont_participate_in_school_activites_1+smoking_freq	307.50	0.02
6	y ˜ age+highest_yr_school_completed+smoking_freq	307.80	0.03

The best model according to AIC that the above search found was using the only the smoking variable (including all of its levels), which had an AIC of 305.77. It is important to note that these models include as predictors, all indicator variables for a given categorical variable and do not allow us to construct the best model from all combinations of individual indicator variables independent of the categorical group from which they originated. For instance, in the above method, we could not select for the final model that included indicator variables for a significant factor level of religion and mothers education, but that dropped all other indicators for religion and mothers education.

Exhaustive Search 2: Individual Indicator Variables

Another approach would be to examine the full set of individual indicator variables. Had we attempted to do exhaustive model search for the best subset of all indicator variables, we would have needed to fit 34359738368 (2 to the 35) different models, which would not have been feasible. Therefore we used Rs leaps package as part of another approach to model fitting.

In perhaps a novel way, we used leaps to select indicator variables that contained strong signals. We selected the top 5 five models of size 1 using both forward and backward model search, then used the variables included in models of up to a size of 4 variables. This yielded a top group of 7 recurring indicator variables. We then did exhaustive model search with those variables and found that the following 10 models had the lowest AICs:

	Model	AIC	Explained.Deviance
1	y ~ ethnicity4+smoking_freqpartake+mothers_educationgrad.hs.some.college	307.45	0.02
2	y ~ smoking_freqpartake+mothers_educationgrad.hs.some.college	307.57	0.02
3	y ~ ethnicity4+mothers_educationgrad.hs.some.college	307.96	0.02
4	y ~ ethnicity4+smoking_freqpartake+mothers_educationgrad.college.beyond	308.09	0.02
5	y ~ smoking_freqpartake+mothers_educationgrad.college.beyond	308.10	0.01
6	y ~ smoking_freqpartake	308.22	0.01

We get a rather interesting result. The best model, as given by AIC contains a predictor, ethnicity4, the binary variable representing whether a subject is Japanese. When we examine that model however, it becomes clear that it is not significant (see both the summary and F-tests below).

```
##
## Call:
## glm(formula = y ~ ethnicity4 + smoking_freqpartake + mothers_educationgrad.hs.some.college,
##     family = binomial(link = logit), data = best_subset_leaps)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.673  -1.254   0.916   0.919   1.103
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              0.179      0.220    0.81    0.415
## ethnicity4                             -15.216    882.743   -0.02    0.986
## smoking_freqpartake                      0.465      0.297    1.57    0.118
## mothers_educationgrad.hs.some.college    0.471      0.277    1.70    0.088 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 306.66  on 233  degrees of freedom
## Residual deviance: 299.45  on 230  degrees of freedom
## AIC: 307.5
##
## Number of Fisher Scoring iterations: 13
## Single term deletions
##
## Model:
## y ~ ethnicity4 + smoking_freqpartake + mothers_educationgrad.hs.some.college
##                                       Df Deviance AIC F value Pr(>F)
## <none>                                        300 308
## ethnicity4                             1      302 308    1.63   0.20
## smoking_freqpartake                    1      302 308    1.93   0.17
## mothers_educationgrad.hs.some.college  1      302 308    2.25   0.14
```

We also look at the model without ethnicity4, which has a slightly higher AIC value:

```
##
## Call:
## glm(formula = y ~ smoking_freqpartake + mothers_educationgrad.hs.some.college,
##     family = binomial(link = logit), data = best_subset_leaps)
```

```
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.667  -1.252   0.915   0.928   1.105
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              0.173      0.219    0.79     0.43
## smoking_freqpartake                      0.481      0.296    1.62     0.10
## mothers_educationgrad.hs.some.college    0.447      0.275    1.62     0.10
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 306.66  on 233  degrees of freedom
## Residual deviance: 301.57  on 231  degrees of freedom
## AIC: 307.6
##
## Number of Fisher Scoring iterations: 4
## Single term deletions
##
## Model:
## y ~ smoking_freqpartake + mothers_educationgrad.hs.some.college
##                                       Df Deviance AIC F value Pr(>F)
## <none>                                        302 308
## smoking_freqpartake                    1      304 308    2.07   0.15
## mothers_educationgrad.hs.some.college  1      304 308    2.03   0.16
```

Results

In the full model, the model with the overall lowest AIC for both complete factor variables and individual factor levels, smoking behavior had a consistently positive coefficient and the most statistically significant predictor in the regression.

To choose between the three models, each of which point to smoking as the most important variable in the analysis, we looked to AIC. The best model from of the exhaustive search of entire full categorical variables has the best AIC of any model that we fit. The only predictors in this model are the levels of the smoking variable. Thus we conclude that using smoking behavior to target at-risk populations is the best result that we can generate. No other variables through any of our analysis seem to hold consistent and significant predictive value.

Having at this point established a final model, we would normally look at residual plots and check assumptions that underlie all linear regression models. In this case, however, that does not make sense. Residuals will not be normally distributed, because they can only take one of six values, two for each of the three possibilities with regard to smoking - smoking, non-smoking, and no response, where one of the values is for records with a 1 response and one for a 0 response. A better way to understand our result is to look at a two-way table because we have a binary response variable and a categorical predictor that can be represented by two binary indicator variables:

```
##    yc
##      0  1
##   0 55 88
##   1 30 61
```

Among non-smokers and those who we did not have information about smoking, 88\143 (0.615) have had sex without contraception and among smokers, 61\91 (0.670) have had sex without contraception. This means smokers are about 9 percent more likely have had sex without contraception (1-(61\91)\(88\143)). To get a better picture, we can construct a two-way table, dropping records that have no smoking data associated with them and we get an even stronger effect of smoking:

```
##           y_noNA
## smoke_noNA  0  1
##          0 31 31
##          1 24 57
```

Only 50 perecent (31\62) of non smokers have had sex without contraception and whereas almost 65 perecent (57\81) of smokers have had sex without contraception. Therefore we can say that smokers have sex without contraception at a rate that is 130 percent higher than non-smokers

```
simple.demo <- glm(y_noNA ~ smoke_noNA, family = binomial(link = logit))
summary(simple.demo)

##
## Call:
## glm(formula = y_noNA ~ smoke_noNA, family = binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.560  -1.177   0.838   0.838   1.177
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.09e-15   2.54e-01    0.00    1.000
## smoke_noNA  8.65e-01   3.52e-01    2.46    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 190.56  on 142  degrees of freedom
## Residual deviance: 184.40  on 141  degrees of freedom
## AIC: 188.4
##
## Number of Fisher Scoring iterations: 4

exp(simple.demo$coefficients[2])

## smoke_noNA
##      2.375
```

In the final model, shown above, the odds of having had sex without contraception for an individual that smokes over the odds for an individual that does not smoke, the odds ratio, is exp(.8650)=2.375, holding all other regressors constant. In terms of a percent change, the odds for someone that partakes in smoking to have had sex without contraception are 137.5 percent higher than for someone who does not smoke, holding all else constant. Therefore, we see that that the logistic regression model shows similar effects of smoking on having sex without contraception. The results from the two-way table and the logistic regression differ by 5.5 percent. This difference is likely attributable to the signal of the data being forced though the logit function.

Discussion

This dataset was tedious to work with for a number of reasons. First, constructing a dataset of relevant predictors before the model fitting process was a challenge. The data itself, as we discovered when fitting models, did not contain a very strong signal, which is why we struggled to find statistically significant predictors. We can concretize this assumption about signal by looking at the pairwise correlation between the response variable and each of the predictors:

Variable	Correlation
smoking_freqpartake	0.10
mothers_educationgrad.hs.some.college	0.10
ethnicity6	0.07
educational_aspirationgrad.hs.some.college	0.06
fathers_educationgrad.hs.some.college	0.06
ethnicity3	0.06
know_preg_peera2	0.06
drinking_freqno_response	0.06
smoking_freqno_response	0.06
drug_freqno_response	0.06
fathers_educationother.schooling	0.05
ethnicity7	0.05
educational_aspirationno.response	0.05
age	0.05
religion4	0.04
educational_aspirationother.schooling	0.04
highest_yr_school_completed	0.04
drug_freqpartake	0.04
religion5	0.03
religion1	0.03
dont_participate_in_school_activites_12	0.02
ethnicity5	0.01
mothers_educationother.schooling	0.00
ethnicity2	-0.00
drinking_freqpartake	-0.01
fathers_educationno.response	-0.01
religion3	-0.02
fathers_educationgrad.college.beyond	-0.03
educational_aspirationgrad.college.beyond	-0.07
ethnicity4	-0.09
mothers_educationno.response	-0.09
mothers_educationgrad.college.beyond	-0.09
religion2	-0.10

Beyond these challenges, there are a number of idiosyncrasies that this data presented. These features are both interesting and relevant for assessing the validity of our conclusions.

In our final model, the no response category for smoking was statistically significant at a 5 percent level and the coefficient was positive. In examining these responses, it seems clear that those subjects with no response were given a survey missing the question because they also had no response for drug and alcohol use. Because the coefficient that we fit for the smoking no response group is positive, we might hypothesize that this groups unobserved smoking behavior is skewed toward the behavior of the smoking group.

Our analysis was a statistical exercise to identify at risk populations in a given geography at a specific time. Given that the data was collected between 1984-86 in San Francisco, it is unclear as to whether or not the results of our analysis are applicable to populations of teenage girls today or in different geographical regions. Thus, the external validity of our analysis is questionable. Between the mid-1980s and today, attitudes and behaviors relating to contraception and to smoking have changed dramatically. Additionally, without knowing the specific locations where the data were collected, we cannot know if ethnic factors from the sampled populations apply to the same ethnic groups outside of that area. In sum, temporal and situational considerations may greatly limit the generalizability of our conclusions.

The last issue with the data, and one that we were under-equipped to either characterize or compensate for, was a partial temporal dependence. We had teens of many different ages reporting on their behavior. This presented a challenge because the response variable that we chose is not regarding behavior in the present moment, it is about a behavior that may or may not have occurred in the past and still occur in the future. It stands to reason that teens who have are older and have been sexually active for a longer period of time may have engaged in risky behaviors more often than younger teens who have not yet had the opportunity to do so.