Smoking Data Analysis

STAT 420 - Dalpiaz

By: Charlie Diamond (csdiamo2)

Devlin McKay (dcmckay3)

Malcolm Taylor (mtaylr11)

Erica Williams (elwilli2)

**Introduction**

For this project we analyzed a set of data that has to do with smoking. This data set has observations on 10,000 indoor workers, which is a subset of a 18,090 observation data set collected as part of the National Health Interview Survey in 1991 and then again (with different respondents) in 1993. The data set contains information on whether individuals were, or were not, subject to a workplace smoking ban, whether or not the individuals smoked and other individual characteristics. This data was provided by Professor William Evans of the University of Maryland and were used in his paper with Matthew Farrelly and Edward Montgomery: “Do Workplace Smoking Bans Reduce Smoking?” *American Economic Review,* September 1999, Vol. 89, No. 4, 728-747. Charlie had also gotten the data from his Econometrics 471 class.

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| *smoker* | =1 if current smoker, 0 if otherwise |
| *smkban* | =1 if there is a work area smoking ban, =0 otherwise |
| *age* | age in years |
| *hsdrop* | =1 if high school dropout, =0 otherwise |
| *hsgrad* | =1 if high school graduate, =0 otherwise |
| *colsome* | =1 if some college, =0 otherwise |
| *colgrad* | =1 if college graduate, =0 otherwise |
| *black* | =1 if black, =0 otherwise |
| *hispanic* | =1 if Hispanic, =0 otherwise |
| *female* | =1 if female, =0 otherwise |

In terms of these variables, the educational binary indicators refer to the highest level of education attained and thus are mutually exclusive. Out of the 10 variables, 9 of them are binary, while age is in integers.The binary variables take a value of 1 if the person has that characteristic and a 0 if they don’t. The response variable is if the person is a current smoker or not. The predictor variables are the smoking ban in a work area, their age, their education level, and other individual characteristics. Our research question is:

* Is education level a significant predictor of smoking habits?

**Materials and Methods**

Logistic models were fit with a person’s smoking habits as a response to three combinations of predictor variables. The Bayesian Information Criterion was applied to each model to eliminate unnecessary variables. Cross validation was used to compare the models. The model with the lowest cross validation estimate would be the best model.

**Initial Models**

To analyze the data, several large models were chosen to include important aspects of the data. The log odds ratio as a function of a linear combination of all of the predictor variables (basic), a linear combination of the predictor variables except the terms relevant to education (basic - edu) and a combination of all of the predictor variables and every possible interaction term (huge). These models were chosen because they encompass a wide range of possible models, which could predict the likelihood of a person being a smoker. We were interested in seeing if the simple model without education would be a better predictor than a model with education. It would also be interesting to see whether education terms would remain in the basic model and huge model after variable selection. The individual effects of the education related predictors could be compared with the other predictor variables.

**Logistic Regression**

When the response variable for a regression is binary, a logistic regression is a good method to fit a model. Linear models would violate the categorical interpretation of the response. A logistic regression uses the odds of one outcome over another. By using the odds, the response is bound between 0 and 1. The logistic equation is

=

where pi is the probability, pi /(1-pi) is the odds ratio, the *B’s* are the coefficients, the x’s are the predictor variables, and the e is the error term. It is important to notice that a change in beta coefficients and predictor variables does not correspond directly to a change in the odds ratio but rather a change in the log odds ratio.

**Model Selection**

After fitting the first three models, three more models were fit using P-Value Model selection. Three typical ways to select a model are backward, forward, and stepwise selection. Backwards selection was chosen because it is possible for a set of variables to have considerable predictive capability even though any subset of them does not. The process of backward selection is to start with all of the predictors in the model, and then step by step eliminate the least significant variable, which is the variable with the largest P-value. After the removal of each variable, the variables are ranked according to the P-value until removing no variables has the lowest P-value.

**Bayesian Information Criterion**

The two main methods used to determine the validity of a model are Akaike Information Criterion and Bayesian Information Criterion. AIC and BIC are typically found in the form where L is the likelihood function, p is the number of parameters in the model, and k equals 2 for AIC and k equals log(n) for BIC. We chose to perform the BIC. By definition, BIC is an estimate of a function of the probability of the model being true, under a certain Bayesian setup, so that the lower the BIC of the model, the more likely it is to be considered the true model. We chose to perform the BIC because of the large sample size (n = 10,000). One thing to keep in mind is that for a reasonable sample size, the BIC generally will pick a smaller model than AIC.

After the Bayesian Information Criterion was applied to each of the initial three models, cross validation was performed on each model to determine the test error coefficient.

**Cross Validation**

Cross validation was performed on all six of the models to determine the error rate of each model. Cross validation is the process of randomly dividing the data into a set number of groups, then fitting models without one of the groups. The excluded data from the group is then *checked* against the new model. The new model will not perfectly predict the data from the group and the errors can be compiled and compared against the total number of observations. The response is qualitative, which means we can calculate a ratio of the misclassifications to the number of observations in the group. (Misclassifications/observations in group) This ratio is called the test error ratio.

The best model will have a lower test error ratio which indicates it makes the least number of mistakes when data is re-sampled. The predictive power can be found by finding 1- the test error estimate. The best model similarly will have the highest predictive power.

**Results**

The statistical program, R, and the programming environment, RStudio, were used to analyze the data set. The packages ‘haven’ and ‘boot’ were also imported to help analyze the data, while ‘knit’ was used to output the data into the appendix. ‘Haven’ was used to interpret the data file and ‘boot’ was used to assist in cross validating the models. After setting several variables, such as the number of observations, and creating a vector for the cross validation results, the analysis began.

The Appendix contains a table for each of the models which contains the variables and the probability for the Beta values. The probability is the increase if the predictor is set to 1. The results from the Bayesian variable selection can be found in the table.

After the Bayesian information criterion was performed, there were six models that predicted the likelihood of a person being a smoker. K-fold cross validation was used to determine which model predicted the most accurately. The data set is large and the computing power was not a concern, so K was set to 10. The K value being set to 10 means the data was separated into 10 groups before the new models were fit and the test errors found from the excluded data. The results from the “modified huge” model turned out to be 0.44 for smkban, 0.50 for age, 0.95 for hsdrop, 0.90 for hsgrad, 0.77 for colsome, 0.61 for colgrad, 0.40 for hispanic, 0.45 for female, 0.49 for age:hsdrop, 0.50 for age:hsgrad, and 0.36 for hsdrop:hispanic.

One reason why the modified huge model had the lowest cross validated result is because it had the most variables to select from. The cross validated results differed by at most 0.0086, with the modified huge model having the lowest estimate, and the basic minus education model having the largest. The cross validated results can be seen as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Basic | Mod Basic | Huge | Mod Huge | Basic-edu | Mod Basic-edu |
| 0.1736 | 0.1737 | 0.1736 | 0.1734 | 0.1820 | 0.1735 |

**Conclusion**

Using the “modified huge” regression, we chose to test specific data points to help visualize our research question of whether or not education level is an adequate predictor of smoking habits. The best model turned out to be the “modified huge” one. In terms of the research question, all of the education terms remain in the model, and the interaction terms were related to education. They have a strong influence when predicting the model. That being said, we do think that education level is a significant predictor of smoking habits.

**Appendix**

Given 30 year old, white male, p(smoking)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Smkbn | Age | HSDrop | HSGrad | ColSome | ColGrad | Black | Hispanic | Female | (P) |
| 0 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .1139 |
| 0 | 30 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | .1697 |
| 0 | 30 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | .3022 |
| 0 | 30 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | .4085 |
| 0 | 30 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | .5638 |

Given 30 year old, white female, p(smoking)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Smkbn | Age | HSDrop | HSGrad | ColSome | ColGrad | Black | Hispanic | Female | (P) |
| 0 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | .0967 |
| 0 | 30 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | .1454 |
| 0 | 30 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | .2650 |
| 0 | 30 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | .3651 |
| 0 | 30 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | .5183 |

Given 30 year old, hispanic male, p(smoking)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Smkbn | Age | HSDrop | HSGrad | ColSome | ColGrad | Black | Hispanic | Female | (P) |
| 0 | 30 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | .0775 |
| 0 | 30 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | .1178 |
| 0 | 30 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | .2206 |
| 0 | 30 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | .3111 |
| 0 | 30 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | .3263 |

Given 30 year old, hispanic female, p(smoking)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Smkbn | Age | HSDrop | HSGrad | ColSome | ColGrad | Black | Hispanic | Female | (P) |
| 0 | 30 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | .0654 |
| 0 | 30 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | .1001 |
| 0 | 30 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | .1908 |
| 0 | 30 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | .2732 |
| 0 | 30 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | .2874 |

Basic - probabilities

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| smkban | age | hsdrop | hsgrad | colsome | colgrad | black | hispanic | female |
| 0.4376 | 0.4981 | 0.8734 | 0.8210 | 0.7650 | 0.6046 | 0.4627 | 0.3578 | 0.4530 |

Mod Basic probabilities (p)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| smkban | age | hsdrop | hsgrad | colsome | colgrad | hispanic | female |
| 0.4373 | 0.4981 | 0.8720 | 0.8199 | 0.7635 | 0.6042 | 0.3605 | 0.4524 |

Mod huge probabilities

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| smkban | age | hsdrop | hsgrad | colsome | colgrad | hispanic | female | age: hsdrop | age: hsgrad | hsdrop: hispanic |
| 0.4360 | 0.5007 | 0.9511 | 0.8984 | 0.7711 | 0.6138 | 0.3953 | 0.4544 | 0.4945 | 0.4959 | 0.3644 |

Basic- Edu Fit

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| smkban | age | black | hispanic | female |
| 0.4008 | 0.4984 | 0.5002 | 0.4477 | 0.4743 |

Mod Edu Fit

|  |  |
| --- | --- |
| smkban | age |
| 0.3990 | 0.4985 |

Huge Probabilities

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| smkban | age | hsdrop | hsgrad | colsome | colgrad | black | hispanic | female |
| 0.3009 | 0.4993 | 0.9356 | 0.8681 | 0.6970 | 0.5438 | 0.2633 | 0.4935 | 0.5494 |
|  | smkban:age | smkban: hsdrop | smkban: hsgrad | smkban: colsome | smkban: colgrad | smkban: black | smkban: hispanic | smkban: female |
|  | 0.5004 | 0.6599 | 0.6269 | 0.6091 | 0.5754 | 0.5275 | 0.4884 | 0.5298 |
|  |  | age: hsdrop | age:hsgrad | age:colsome | age:colgrad | age:black | age:hispanic | age:female |
|  |  | 0.4944 | 0.4968 | 0.5011 | 0.5015 | 0.5042 | 0.5047 | 0.4992 |
|  |  |  | hsdrop: hsgrad | hsdrop :colsome | hsdrop: colgrad | hsdrop: black | hsdrop: hispanic | hsdrop: female |
|  |  |  | NA | NA | NA | 0.5351 | 0.1815 | 0.4747 |
|  |  |  |  | hsgrad: colsome | hsgrad: colgrad | hsgrad: black | hsgrad: hispanic | hsgrad: female |
|  |  |  |  | NA | NA | 0.5228 | 0.2647 | 0.4235 |
|  |  |  |  |  | colsome: colgrad | colsome: black | colsome: hispanic | colsome: female |
|  |  |  |  |  | NA | 0.5930 | 0.3167 | 0.4446 |
|  |  |  |  |  |  | colgrad: black | colgrad: hispanic | colgrad: female |
|  |  |  |  |  |  | 0.7131 | 0.3818 | 0.3959 |
|  |  |  |  |  |  |  | black: hispanic | black: female |
|  |  |  |  |  |  |  | 0.5861 | 0.4398 |
|  |  |  |  |  |  |  |  | hispanic: female |
|  |  |  |  |  |  |  |  | 0.4037 |