Generalized Linear Modeling of Health Risk

By

Rebecca Anne Black

Thesis Project

Submitted in partial fulfillment of the

Requirements for the degree of

MASTER OF SCIENCE IN PREDICTIVE ANALYTICS

June, 2013

Dr. Vivek Ajmani, First Reader

Dr. Michael Speed, Second Reader

ABSTRACT

Generalized Linear Modeling of Health Risk

Rebecca A. Black

Health Care and health care costs have been frequent topics in the popular media for quite some time. The recession of 2009 and consequent high rate of unemployment in America has severely taxed the ability of the average citizen to pay for health-related costs, and it is not clear whether the Affordable Care Act has helped or worsened this issue. At the same time, health care costs themselves have increased at the federal level due to an aging population and consequent strain on the Medicare system. Unarguably, one of the most important policy challenges of the future is to find a way to control the cost of health care at all levels and at the same time preserve high quality and readily available medical care for everyone. As such, an understanding of the health status of the American populace as well as a knowledge of the risk factors associated with the health status of those individuals can be a useful tool as we move forward.

This study provided a descriptive and predictive analysis of the health risk behavior of a sample of the American populace. Using SAS, several regression models were fit to two response variables. A backward variable selection procedure was performed, giving us a set of significant factors associated with health status. The coefficient estimates for these factors give us a measure to gauge effect magnitudes for future health initiatives aimed at risk factor mitigation and health care cost management.

Table of Contents

List of Tables

List of Graphs

Introduction

Health Care and health care costs have been frequent topics in the popular media for quite some time. The recession of 2009 and consequent high rate of unemployment in America has severely taxed the ability of the average citizen to pay for health-related costs, and it is not clear whether the Affordable Care Act has helped or worsened this issue. At the same time, health care costs themselves have increased at the federal level due to an aging population and consequent strain on the Medicare system. Unarguably, one of the most important policy challenges of the future is to find a way to control the cost of health care at all levels and at the same time preserve high quality and readily available medical care for everyone. As such, an understanding of the health status of the American populace as well as a knowledge of the risk factors associated with the health status of those individuals can be a useful tool as we move forward.

This study will provide both descriptive and predictive analysis of the health risk behavior of a sample of the American populace. The descriptive analysis will give us a snapshot of the general medical profile of the population with respect to a variety of factors such as access to health care, whether or not the individual has a "personal" doctor, and whether they have trouble paying for medical care. Additionally a profile of a host of health issues will be formed, telling us of the prevalence of systemic and cardiac problems, disability status, tobacco use, and alcohol consumption. The predictive analysis will focus on two response variables: general health status and number of healthy days in the previous month. The analysis will control for several demographic factors as well as incorporate the factors from the descriptive analysis as explanatory variables.

The dataset we will use is sourced from the Centers for Disease Control (CDC) Behavioral Risk Factor Surveillance System. As a part of this surveillance program, the CDC gathers yearly data on a large random sample of adult residents of the United States. The 2012 dataset comprises an ideal body of information to facilitate an analysis to address the issues described above. As the dataset includes demographic information as well as risk behaviors and health status, it makes it possible to conduct a thorough Exploratory Data Analysis (EDA) as well as several types of regression analyses to investigate associations between the selected response variables and a large number of explanatory variables.

The CDC dataset possesses two desirable properties that make it a good source of data for our analysis. Firstly, it is a random sample which allows us to make inferences about the American population in general, and secondly, it is quite large (just under 500,000 records,) which enables us to have confidence in our estimates of factor-level effects in the predictive analysis. Thus the results we derive can provide an accurate reference for policy makers as they seek to further understand our nation's health challenges and design programs that address those challenges while optimizing the cost outlay associated with program implementation.

This study will provide insightful analysis that can aid policy makers in understanding the health-related challenges of individuals in the U.S., as well as describe significant associations between behavioral factors and health status. This information will prove to be a useful tool to assist in further development and refinement of recent health policy decisions and programs as decision makers seek to improve services and trim unnecessary health care expenditures.

Additionally, this study may provide a cogent body of knowledge with which to modify the current legislation surrounding the Affordable Care Act. Much of the results contained herein

suggest that there exist individuals who will be adversely affected by the new restrictions - individuals who were previously cared for quite well by the existing system and may experience a reduction in health care affordability under the new system.

Review of the Literature

There are three thematic areas in which I sought references: Health Policy, Model Specification, and Goodness of Fit. The search for Health Policy Literature focused on issues faced by policy makers in designing new legislation, policy, and initiatives. The search for Model Specification and Goodness of Fit Literature focused on uncovering the current expert opinions on model choice and fit measures appropriate for Big Data settings.

Health Policy Literature

Although the Affordable Care Act, Medicare Reform, and general health care policy has been discussed at length in the popular media, those discussions tend to overlook the three critical aspects of any health care issue at the policy level, namely equitable allocation of medical care, quality of care, and health care costs, both at the individual and governmental level. These factors are variable and can theoretically be "tuned" to optimal levels to conform to resource constraints such as regional physician density, local population disability levels, proximity to large medical centers, and even demographic factors such as median age, income, and family size. Perhaps understanding the factors that drive these resource constraints can aid policymakers in constructing and implementing more appropriate and cost effective health care measures.

In this vein, American College of Physicians 2009 discusses the challenges involved in controlling health care costs while at the same time ensuring high quality and adequate care for all. Similarly, American College of Physicians 2011 discusses the notion of health care resource allocation and the implications on the cost and quality of that care. The underlying notion here is that health care must meet the needs of patients or it is essentially useless--that high quality care

may depend on the needs of the patient, and may not simply be some predefined “one size fits all” standard.

Finally, as prevention is a key factor in cost management, American College of Physicians 2012 examines several Medicare reform proposals and assesses the implications of each on the quality of care for Medicare benefit recipients with regard to prevention, wellness, and management of chronic conditions.

Model Specification Literature

My initial exploratory analysis will include a range of potential models for each of the two response variables. The discrete response variable is a count, and so I will fit an Ordinary Least Squares Regression model, a Poisson Regression model, a Quasi-Poisson Regression model, and a Negative Binomial Regression model. Sheather 2009 and Chatterjee and Hadi 2012 provides excellent coverage of the issues surrounding OLS regression models. Myers, Montgomery, Vining, and Robinson 2010 discuss Generalized Linear Models and the choices involved in fitting each of the three models in my analysis. Finally, Hilbe 2011 covers Negative Binomial Regression models quite extensively, and offers much useful guidance about the modeling process in general.

The categorical response variable in this analysis is a 5-category ordered measure. In this case I will fit both an Ordinal Logistic Regression model as well as a Binary Logistic Regression model to a recoded dichotomized version of the variable. In both cases, Hosmer, Lemeshow, and Sturdivant 2013 and Allison 2012 provide excellent discussion of these models. Allison in particular discusses the Cumulative Logit Model, the Adjacent Categories model, and the

Continuation Ratio model--all options when fitting a Logistic Regression model to an ordered categorical response variable.

Goodness of Fit Literature

After a model has been fit to data, it is essential to conduct a goodness of fit assessment. There are a wide range of goodness of fit measures, and these can differ depending on the model used. Additionally the suitability of a given measure can depend on sample size and number of independent variables retained in the analysis. Chatterjee and Hadi 2012 provide a standard discussion of goodness of fit measures for Ordinary Least Squares regression models, including R^2, Adjusted R^2, the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The latter two measures incorporate a penalty for a large number of independent variables, making them desirable for models in a Big Data setting.

Hilbe 2011 discusses goodness of fit for Negative Binomial Regression models, explaining that in this case the fit is essentially a comparison of the suitability of the Negative Binomial model and the suitability of the Poisson model. Fit statistics in this case include the Score test or Lagrange Multiplier test.

Applicable to Generalized Linear Models in general, Hilbe 2011; Hosmer, Lemeshow, and Sturdivant 2013; and Allison 2012 discuss the deviance criterion, which allows us to compare models from different Generalized Linear Model families and favors models with smaller deviance.

Methods

I undertook eight broad steps in the modeling process. The first step involved data acquisition from the CDC, followed by data cleaning and reduction of the dataset to include only the variable subset under consideration. This was followed by exploratory data analysis and initial modeling runs with the data partitioned into training and test sets. After this, a variable selection process was employed followed by a final candidate model selection. After final candidate models were selected they were run using the test set as an insurance against overfitting. After the test set models were run, the final step involved merging the training and test data and completing a final run on the merged dataset.

The dataset was acquired from the CDC Behavioral Risk Factor Surveillance System website (CDC 2013.) The dataset contained 359 variables and 475,687 observations and was a 95 megabyte SAS Transport file. After downloading the file I imported the data into JMP for further processing.

The survey included question subsets targeted at specific states, and so after the data were loaded into JMP, I reduced the variable set from 359 to 108 to limit the analysis to variables common to all respondents in the survey. I then addressed the issue of missing data.

I chose to delete records for which the two response variables GENHLTH (reflecting general health status) and PHYSHLTH (reflecting number of unhealthy days in the previous month) were missing, refused, or not sure, as I felt it inappropriate to consider imputation for response variables. Additionally, the explanatory variables SMOKDAY3 (reflecting the number of cigarettes smoked per day) and HADSIGM3 (reflecting a previous sigmoidoscopy procedure) had nearly 50% missing values, and so I chose to delete those variables entirely.

The next issue concerned variable recoding. There were three cases under consideration here--mixed units, category aggregation, and dummy variable creation. The weight variable had values in both pounds and kilograms, so I standardized all values to pounds. I created a new response variable GENHLTH_LOG as a dichotomized version of GENHLTH to use in the binary logistic regression analysis. Next, I recoded all variables responses of "don't know" or "refused" to missing. Additionally, I chose to delete the variables HISPANC and MRACE because the mixed race combinations created far too many categories to add meaningful information to the analysis.

Finally I created dummy variables for all of the categorical variable choices. Ordinarily this is accounted for automatically in SAS analyses, however as I was using a variable selection procedure later in the analysis, dummy variable creation was required.

At this point in the modeling process I took the prepared JMP dataset and imported it into SAS for further analysis. Once the data were in SAS, I conducted an exploratory data analysis. For the six continuous variables, I used the PROC MEANS procedure to calculate the mean, median, standard deviation, and 95% CI for the mean. I also used PROC UNIVARIATE to display histograms and probability plots for these variables, including Goodness-of-Fit Tests for the Normal Distribution. For the remaining 37 categorical variables, I produced histograms and summary statistics, along with contingency tables for selected variables.

After the EDA I turned to the initial model runs. For all analyses I partitioned the dataset into a training set / test set split of 70% and 30% respectively. For the set of categorical response models I used the GENHLTH variable, and performed two types of analysis: Binary Logistic Regression and Ordinal Logistic Regression. The Binary Logistic Regression was run on a

dichotomized version of the GENHLTH variable with the response categories 1-4 (corresponding to adequate health status) mapped to "1" and the response category 5 (corresponding to poor health) mapped to "0". The Ordinal Logistic Regression was run on the original GENHLTH variable. In both cases this initial run included every possible explanatory variable in the model statement.

For the set of count response models, I used the PHYSHLTH variable and performed Ordinary Least Squares Regression, Poisson Regression, Quasi-Poisson Regression, and Negative Binomial Regression.

After these initial runs were complete, I ran a second set of models using the SELECTION=BACKWARD option in the model statements for the Logistic, Ordinal Logistic, and Ordinary Least Squares Regression. As SAS 9.3 PROC GENMOD does not offer a variable selection option, I used the PROC GLMSELECT procedure on the data to select a set of significant variables, with the intention of using those resulting variables for the final Poisson and Negative Binomial model runs.

After the variable selection procedures were implemented, I performed final runs for each model using the variable set identified by the selection procedure. The result was a set of coefficient estimates for each of the selected explanatory variables as well as a set of goodness of fit criteria for those models.

After obtaining the coefficient estimates for each of the five models, I then ran the same model statements on the test set, and compared the two sets of coefficients. The estimates were nearly identical, so taking the advice of Hilbe 2011, I merged the test and training sets and

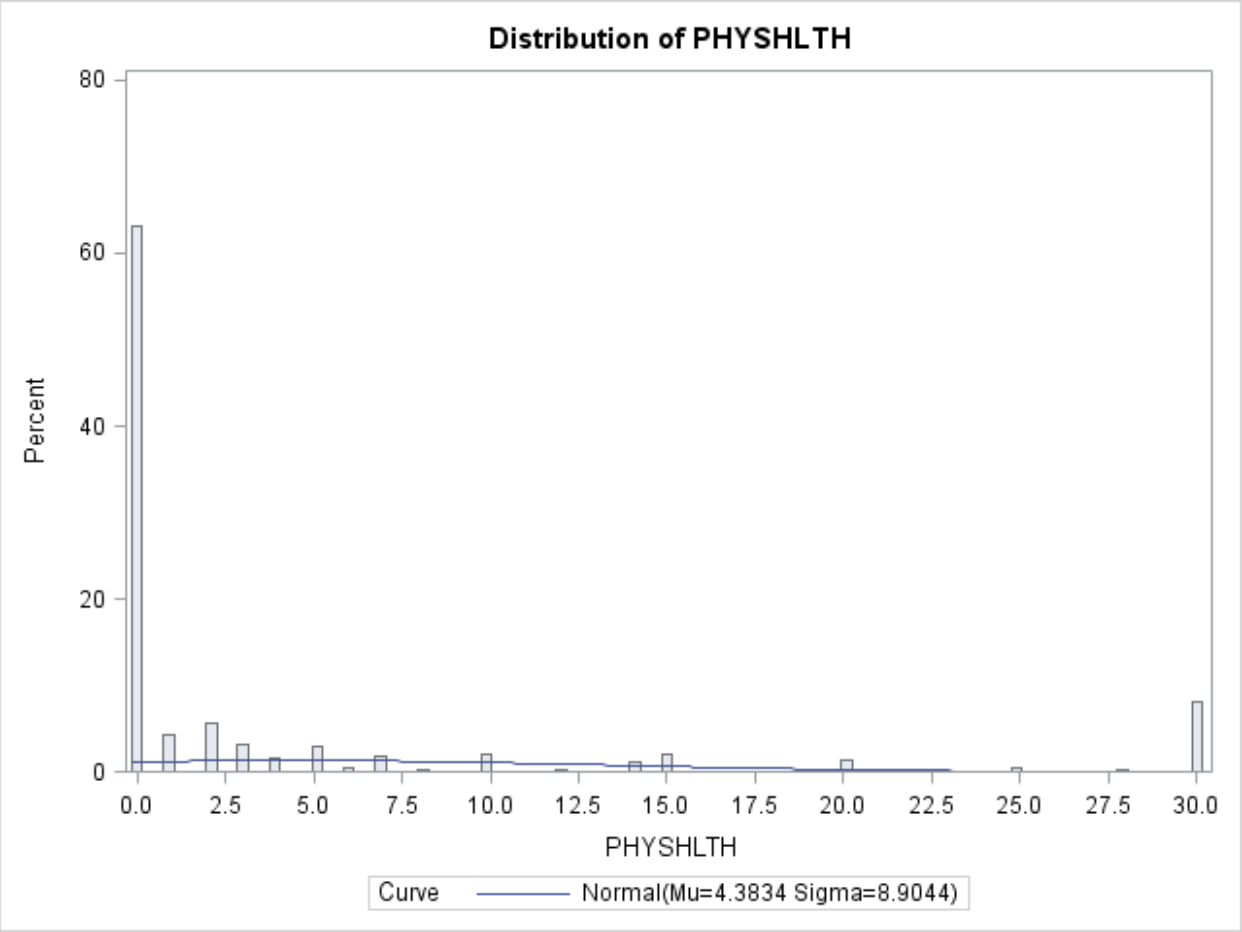
performed one last set of model runs on the combined dataset, taking the resulting coefficient estimates as the final set to be used in the results interpretation.

Results

Exploratory Data Analysis

This dataset has 108 variables, of which 43 relate directly to health behaviors. These 43 variables consist of five continuous or count variables and 38 categorical variables.

The count variable PYSHLTH (corresponding to the question " …for how many days during the past 30 days was your physical health not good?") has a mean of 4.4, a median of 0, and standard deviation of 9. PHYSHLTH is very right skewed, and the Kolmogorov-Smirnov test for normality has a p-value less than 0.01, confirming that the responses are non-normal. The empirical distribution below provides a graphical confirmation of these test statistics.



Graph 1

We know that as the mean of a Poisson random variable increases, the Poisson distribution converges to the Normal distribution. The low mean for PHYSHLTH indicates that a straight-forward OLS model is not the preferred one and that the Poisson and Negative Binomial Regression models are more appropriate.

The categorical variable GENHLTH (corresponding to the question "Would you say that in general your health is...(1) Excellent, (2) Very Good, (3) Good, (4) Fair, or (5) Poor") shows 18% of respondents with Excellent health, 32% with Very Good health, 31% with Good health, 13% with Fair health, and just under 6% with Poor health. Thus 94% of respondents report a health status of fair or better.

Among the explanatory variables, it is instructive to consider health care status. Just over 88% of respondents have health care coverage, 85% of respondents have one or more individuals they consider their personal doctor, and 12% report that cost was a barrier to seeking health care in the 12 months prior to the survey. Additionally, 84% of respondents report seeing a doctor for a routine checkup in the previous 2 years.

Under the category of chronic illness, we see 6% of respondents with a previous heart attack, 6% with angina, 4% with a previous stroke, 13% with asthma, 8% with skin cancer, 2.5% with COPD, 25% with some form of arthritis, and 10% with diabetes.

Taken individually, these summary statistics tell us that the great majority of individuals are in reasonable health, have few sick days per month, and have health coverage as well as a personal physician. Furthermore most have had a recent physical exam. The prevalence of chronic illness is somewhat less promising, with a fairly high percentage of asthma, diabetes, and arthritis. This suggests a greater focus on health education and preventative care may be of

benefit. We next turn to the aggregate affect of these factors on the two response variables GENHLTH and PHYSHLTH.

GENHLTH Model

The GENHLTH response variable is an indicator of general health status. The original variable format is a five-category ordered response. After fitting an ordinal logistic regression model to the original variable, I transformed the original variable to a binary response by aggregating the first four choices and leaving the last choice unchanged. I then fit a binary logistic regression model to the transformed variable.

Ordinal logistic regression.

For my initial analysis I fit a cumulative logit model with all possible explanatory variables. SAS uses a complete case analysis, so just over 18,000 records were deleted due to missing values, giving us ~307,000 records to use for the analysis itself. The initial model had a likelihood ratio test p-value of less than 0.0001, showing good overall model utility. The AIC was 742401.45, and the area under the ROC Curve was 0.782.

I then ran the same analysis with a backward elimination variable selection procedure. The procedure discarded 34 variables, leaving us with 115 significant variables. However removing the less significant explanatory variables did not reduce the AIC at all. As the area under the ROC curve for this model is so low, I turn to a binary logistic regression to seek a better model for these data.

Binary logistic regression.

The initial binary logistic regression model fit was quite good, with 94.8% correctly classified instances, and an area under the ROC curve of 0.926. Many explanatory variables had

p-values greater than 0.10, so I next fitted the full model with a variable selection procedure, using the backward elimination method as recommended by Chatterjee and Hadi 2012. I used p=0.05 as the cutoff point for inclusion.

The backward elimination procedure discarded 68 variables, leaving 81 variables in the reduced model. It should be noted that this count includes dummy variables, which explains the large number of variables left in the model.

The reduced model has an AIC of 82359.018, which is moderately lower than the AIC of the full model. The Likelihood Ratio test of the global utility of the reduced model has a p-value of <0.0001, with a c-statistic of 0.926 and 94.8 % correctly classified instances.

As as final step before accepting this model, I used the coefficient estimates from the training set run to predict the response variable values in the test set. The results were very favorable, with 124,067 correct predictions out of a total of 138,823, which means the training set model correctly classified 89.4% of the test set response values.

The reduced model appears to be a good fit for the data, and is considerably better than the ordinal logistic regression model. Thus I use the binary logistic regression reduced model coefficient estimates for interpretation. The procedure is modeling the probability that the respondent reports a health status of fair or better, and the odds ratio estimates are reproduced below:

Table 1

Binary Logistic Regression
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
No personal doctor	1.098	1.024	1.177
No difficulty paying medical costs	1.466	1.394	1.542
Never had clinical breast exam	0.742	0.690	0.798
Has had PAP test	1.162	1.084	1.247
Has had PSA test	1.129	1.056	1.207
Female	1.289	1.187	1.400
Has been tested for HIV	1.089	1.004	1.182
Never tested for HIV	1.129	1.048	1.218
Alabama Resident	0.777	0.695	0.868
Alaska Resident	0.727	0.590	0.897
Arizona Resident	0.814	0.707	0.937
Arkansas Resident	0.755	0.653	0.873
California Resident	0.881	0.783	0.992
Hawaii Resident	0.621	0.530	0.729
Indiana Resident	0.848	0.742	0.969
Kentucky Resident	0.886	0.798	0.984
Louisiana Resident	0.651	0.583	0.727
Mississippi Resident	0.630	0.562	0.707
Nevada Resident	0.814	0.676	0.980
NewJersey Resident	0.837	0.746	0.938
NewMexico Resident	0.742	0.652	0.845
NorthCarolina Resident	0.818	0.730	0.916
Oklahoma Resident	0.844	0.745	0.955
SouthCarolina Resident	0.821	0.739	0.913
Tennessee Resident	0.634	0.560	0.719
Texas Resident	0.730	0.641	0.832
Utah Resident	0.810	0.711	0.922
Virginia Resident	0.846	0.721	0.993
WestVirginia Resident	0.703	0.612	0.807
Wyoming Resident	0.813	0.695	0.951

Table 1

Binary Logistic Regression
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Married	0.873	0.834	0.915
Divorced	0.863	0.819	0.909
Separated	0.852	0.771	0.942
Never had a stroke	1.323	1.250	1.401
Has asthma	0.861	0.821	0.903
Has had skin cancer	0.652	0.483	0.880
Never had skin cancer	0.703	0.523	0.944
Never had any type of cancer	1.861	1.773	1.953
Has COPD	0.696	0.582	0.832
Does not have COPD	1.230	1.031	1.466
Had checkup in past 12 months	1.506	1.353	1.675
Had checkup in past 12-24 months	1.667	1.485	1.872
Had checkup more than 5 years ago	1.511	1.328	1.720
Exercised in past month	1.958	1.884	2.034
Has had heart attack	0.663	0.627	0.702
Has angina or heart disease	0.805	0.705	0.920
Does not have angina or heart disease	1.283	1.128	1.458
Never attended school	0.480	0.346	0.665
Attended grade 1-8	0.745	0.683	0.812
Attended grade 9-11	1.364	1.285	1.449
Attended 1-3 years of college	1.624	1.522	1.732
Graduated college	1.767	1.643	1.902
Employed	1.728	1.587	1.883
Homemaker	1.226	1.100	1.367
Student	1.635	1.302	2.054
Retired	1.149	1.055	1.252
Unable to work	0.508	0.469	0.550
Yearly income less than 10k per year	0.807	0.761	0.855
Yearly income greater than 25k but less than 50k	1.164	1.106	1.225

Table 1

Binary Logistic Regression
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Yearly income greater than 50k	1.530	1.431	1.635
Pregnant	2.207	1.288	3.782
Activities limited due to physical, mental, or emotional problems	0.668	0.574	0.777
No limitations due to physical, mental, or emotional problems	2.496	2.141	2.910
Has disability that requires special equipment (e.g. cane, wheelchair)	0.533	0.511	0.556
Does not use chewing tobacco or snuff	1.136	1.034	1.249
Has had pneumonia vaccine	0.896	0.859	0.935
Always uses seatbelt	1.223	1.132	1.321
Sometimes uses seatbelt	1.304	1.189	1.430
Does not have arthritis	1.241	1.189	1.295
Never diagnosed with depression	1.444	1.384	1.505
Has kidney disease	0.792	0.642	0.977
Never had kidney disease	1.439	1.176	1.763
Never had vision difficulties	1.460	1.403	1.518
Has diabetes	0.638	0.611	0.666
Last dental checkup less than 1 year ago	1.235	1.176	1.296
Last dental checkup 2-5 years ago	1.131	1.073	1.192
Age	0.996	0.994	0.998
Number of children	1.051	1.023	1.079
Weight (in lb.)	1.001	1.001	1.002
Number of falls in previous year	0.981	0.978	0.985

These results reveal a clearer picture about the effects of the factors on reported health status. For example, while 88% of respondents have health coverage, someone without health coverage is 0.816 times as likely to report fair to excellent health as someone with health

insurance. Respondents who have no difficulty paying for health care are 1.466 times as likely to report fair to excellent health as someone who has difficulty with health care costs.

Turning to the effect of chronic illness on health status, we see that respondents who have never had a previous stroke are 1.323 times as likely to report fair to excellent health as someone with a history of stroke, someone with asthma is 0.861 times as likely to report fair to excellent health as someone without asthma, and someone without COPD is 1.230 times as likely to report fair to excellent health as someone with COPD.

As for the effect of preventive care, we see that someone who has had a checkup at any time in the past is between 1.5 and 1.67 times as likely to report fair to excellent health as someone who has never had a checkup, and someone who exercises regularly is 1.958 times as likely to report fair to excellent health as someone who does not exercise regularly.

Examining the effect of demographic variables, we see that someone who graduates high school is 1.364 times as likely to report fair to excellent health as someone who did not graduate high school, and that the attending college offers an even greater positive effect. On the other hand, those who are unable to work are 0.508 times as likely to report fair to excellent health as someone who is able to work, and those who make less than $10,000 per year are 0.807 times as likely to report fair to excellent health as someone who makes more than $10,000 per year. Finally, a woman who is pregnant is 2.207 times as likely to report fair to excellent health as someone who is not pregnant.

PHYSHLTH Model

The PHYSHLTH response variable is an indicator of the number of days in the previous 30 days for which the respondent had ill health. This is a count variable taking on integer values in

the range [0,30]. As such, it is probable that the counts can be modeled by the Poisson or Negative Binomial distribution. Now, the Poisson Random variable assumes counts from 0 to infinity, but in reality the probability of those high counts converge to zero rapidly. As evidenced by the empirical distribution on page 15, this is exactly what we see happening. So we can certainly attempt to model the variable by the Poisson regression model.

Now, a key assumption of the Poisson distribution is that the mean of the data is equal to the variance. We already know that in this case this assumption is violated as evidenced by the results of the exploratory data analysis on page 15. There are two alternative methods to explore in this case. I will look at the Quasi-Poisson models as well as the Negative Binomial, and compare the efficacy and fit of all four alternatives. However as a first pass in the overall modeling process for this response variable I will take a look at the OLS model in addition to the alternatives described above.

Ordinary Least Squares Regression

As before, for my initial analysis I fit a model with all possible explanatory variables. SAS uses a complete case analysis, so just over 18,000 records were deleted due to missing values, giving us ~307,000 records to use for the analysis itself. The initial model had an F-test p-value of less than 0.0001, showing good overall model utility. However the Adjusted R-Square was very low at 0.3439. Additionally, many explanatory variables had variance inflation factors greater than 10, with one as high as 530. Finally, the model residuals were highly heteroscedastic, showing violation of the ordinary least squares regression assumptions. It is clear that this model is inappropriate for these data, and so I turn next to the Poisson and negative binomial models.

Poisson Regression

The initial Poisson regression model with all possible explanatory variables had an AIC of 2,906,222. This model is based on the assumption that the mean of the response variable (PHYSHLTH) given all the explanatory variables is equal to its variance given all the explanatory variables. That is

$$E(PHYSHLTH|X)=Var(PHYSHLTH|X).$$

As indicated by the Poisson Regression Model output, the Deviance/DF value equals 5.2 which indicates that this model suffers from over-dispersion. As explained by Allison (2012), under the null hypothesis of equidispersion, the distribution of the deviance statistic is a Chi-Square with degrees of freedom equal to the difference between the sample size and the number of explanatory variables in the model. Also note that the expected value of a Chi-Square random variable is equal to its degrees of freedom. Therefore, under the assumption of equidispersion, the ratio Deviance/DF should be equal to 1. This is clearly violated when applying the Poisson model to this data set. Therefore, the Poisson regression model is an inappropriate model for this response variable. Overdispersion commonly occurs due to omitted variables. The Poisson regression model does not have a scale parameter to account for the omitted variables. A consequence of overdispersion is that the standard errors of the regression coefficients are underestimated resulting in false rejection of the null hypothesis that the regression coefficients are no different from zero.

Two commonly suggested alternatives to the Poisson regression model in the presence of overdispersion are the Quasi Poisson regression model which adjusts the standard errors of the Poisson regression model coefficients using the square-root of the Deviance/DF statistic and the

negative binomial regression model which adds a dispersion parameter to the regression model to account for the omitted variables that commonly cause overdispersion in Poisson regression models.

Quasi Poisson regression

As a first step toward creating a parsimonious model using Quasi Poisson regression, I used a variable selection procedure to effectively reduce the number of explanatory variables. As I mentioned before, in SAS 9.3 PROC GENMOD does not allow for a variable selection option (however this has been implemented in SAS 9.4.) To circumvent this issue I used PROC GLMSELECT, which employs a backward selection procedure in the setting of GLM (without regard to the link function.) This procedure discarded 105 variables, leaving 44 variables in the reduced model. Once again, it should be noted that this count includes dummy variables, which explains the large number of variables left in the model.

After obtaining the reduced set of explanatory variables, I then fit the resulting model using PROC GENMOD with the Poisson link function and the DSCALE option to adjust the Poisson standard errors to correct for the overdispersion. The DSCALE option takes the standard errors from the original Poisson regression model and multiplies it by the square-root of the Deviance/ DF statistic. Note that the deviance is simply the difference in the log-likelihood of the fitted model versus the saturated model. The parameter values remain unchanged while their standard errors increase. The increase in standard errors therefore lower the values of the test statistics associated with the parameters when compared to what they were in the Poisson regression model. The AIC for this model was 2,915,900 - identical to that of the Poisson, however two variables that were significant in the Poisson ceased to be significant in this adjusted model.

My next step was to fit the same model but using the PSCALE option to adjust the standard errors. The PSCALE option takes the standard errors from the original Poisson regression model and multiplies it by the square-root of the Pearson-Deviance/DF statistic. Note that the Pearson deviance is simply the sum of the squares of the ratio of the difference between the observed count and the expected count and the expected count. As with the DSCALE option, the AIC for this model was 2,915,900 - identical to that of the Poisson, however this time four variables that were significant in the Poisson ceased to be significant in this adjusted model.

Now clearly the Quasi Poisson model offered some rudimentary improvement over the Poisson, in that the adjusted standard errors allowed for a modest reduction in the explanatory variable set. However the lack of reduction in AIC over the Poisson regression model suggests that the Quasi-Poisson adjustments to the standard errors were not significant enough to make up for the Poisson over dispersion and that the Negative Binomial regression model may provide for a better fit.

Negative binomial regression

The initial negative binomial regression model with all explanatory variables had an AIC of 1,131,581--significantly better than the analogous Poisson model. As there were a great many explanatory variables with very high p-values, I used a variable selection procedure to formulate a more parsimonious model, just as I did in the Quasi Poisson case. I used PROC GLMSELECT, which employs a backward selection procedure in the setting of GLM (without regard to the link function.) This procedure discarded 105 variables, leaving 44 variables in the reduced model. Once again, it should be noted that this count includes dummy variables, which explains the large number of variables left in the model.

After obtaining the reduced set of explanatory variables, I then fit the resulting model using PROC GENMOD with the negative binomial link function. The AIC for this model was 1,133,415 - similar to that of the full model, however there still remained a handful of non-significant explanatory variables.

As a step toward building a final model, I fit the reduced training set model to the test set. Once again, I used a complete case analysis, which omitted 7706 observations with missing values and leaving ~131,000 records for the analysis itself.

The model fit to the test set resulted in coefficient estimates that were generally very similar to that of the training set estimates but with slightly higher standard errors (as one might expect given the lower sample size.) There were however 11 instances of mismatches in variable significance between the training set and test set model fits.

As suggested in Hilbe 2011, since the two sets of coefficient estimates are similar, I then combined the training and test sets and fit a final model using the entire dataset. Before proceeding however, I deleted the 11 variables that showed mismatches in significance, as it seemed reasonable to assume that the predictive or explanatory ability of those factors were not general enough to be useful outside of the training dataset.

The final model fit to the entire dataset once again was a complete case analysis, which omitted 23097 observations with missing values and left 441,461 records for the analysis itself. The AIC was 1,625,378 - somewhat greater than for the training set only model, but still significantly better than that of the Poisson and Quasi Poisson models.

The parameter estimates for the pooled model can now be used to interpret the effects of selected explanatory variables on the number of days in the previous 30 days for which the

respondent had ill health. The coefficient estimates for all statistically significant explanatory variables are reproduced below:

Table 2

Negative Binomial Regression Coefficient Estimates

Effect	Estimate	Standard Error	$e^{(Estimate)}$
Has health plan	0.0427	0.0132	1.0436
No personal doctor	-0.1285	0.0117	0.8794
No difficulty paying medical costs	-0.4391	0.0121	0.6446
Has had PSA Test	-0.0952	0.0101	0.9092
Never tested for HIV	-0.0619	0.0079	0.9400
Widowed	-0.082	0.0118	0.9213
Never married	0.0396	0.0109	1.0404
Never had a stroke	-0.1444	0.0182	0.8655
Never had asthma	-0.2064	0.011	0.8135
Never had any type of cancer	-0.3502	0.0125	0.7045
Has COPD	0.1051	0.0547	1.1108
Does not have COPD	-0.1676	0.0535	0.8457
Has checkup in past 12 months	-0.1579	0.0254	0.8539
Had checkup in past 12-24 months	-0.2469	0.0262	0.7812
Had checkup more than 5 years ago	-0.22	0.0285	0.8025
Exercised in past month	-0.3702	0.0087	0.6906
Has had heart attack	-0.1794	0.0164	0.8358
Has angina or heart disease	0.1754	0.0166	1.1917
Graduated high school	-0.1556	0.0139	0.8559
Attended 1-3 years of college	-0.1991	0.0144	0.8195

Graduated college	-0.3298	0.0147	0.7191
Employed	-0.2003	0.0118	0.8185
Student	-0.0708	0.0254	0.9316
Retired	-0.0428	0.0132	0.9581
Unable to work	0.2217	0.0176	1.2482
Yearly income less than 10k per year	0.1793	0.0171	1.1964
Yearly income greater than 10k but less than 25k	0.142	0.0096	1.1526
Activities limited due to physical, mental, or emotional problems	0.5353	0.033	1.7080
No limitations due to physical, mental, or emotional problems	-0.4765	0.0324	0.6210
Has disability that requires special equipment (e.g. cane, wheelchair)	0.3986	0.0127	1.4897
Always uses seatbelt	-0.0407	0.0166	0.9601
Sometimes uses seatbelt	-0.0419	0.0192	0.958965672345(
Does not have arthritis	-0.3941	0.0083	0.674286624206
Never diagnosed with depression	-0.2663	0.0096	0.766209230249(
Never had kidney disease	-0.2812	0.0195	0.754877344912(
Has vision difficulties (even with glasses)	0.2159	0.0098	1.240978274968(
Has diabetes	0.2452	0.011	1.277876863021(
Last dental checkup between 3-5 years ago	0.0374	0.0118	1.038108181073
Weight (in lb.)	0.0002	0.0001	1.000200020001
Number of falls in previous year	0.0132	0.0015	1.013287504596

Examining the interpretations for these effects, we see that the predicted ratio of the number of days of ill health in the previous 30 days for someone with health insurance to

someone without health insurance is 1.04. Since this is very close to one, we can conclude that having health insurance is not associated with greater number of sick days per month, assuming all other explanatory variables are held constant. On the other hand, the predicted ratio of sick days per month for someone with no difficulty paying for medical costs to someone who does struggle with medical costs is 0.6446. In this instance, the number of sick days per month is predicted to be significantly lower for those with no difficulty paying medical costs, assuming all other explanatory variables are held constant.

Some significant demographic effects include inability to work and a yearly income less than $10,000 per year. Those who are unable to work are predicted to have 1.25 sick days per month for every one sick day experienced by someone who is able to work, and those making less than $10,000 per year is predicted to have 1.2 sick days per month for every one sick day experienced by someone making more than $10,000 per year, both situations assuming all other explanatory variables are held constant.

Chronic illness and disability are very strongly associated with greater numbers of sick days per month. For example, those who have activity limitations due to physical, mental, or emotional problems are predicted to have 1.7 sick days per month for every one sick day experienced by someone without activity limitations due to these problems (assuming all other explanatory variables are held constant.) Similarly someone with a disability requiring them to use special equipment are predicted to have 1.5 sick days per month for every one sick day experienced by those without a disability requiring the use of special equipment (assuming all other explanatory variables are held constant.)

Finally, while someone with diabetes is predicted to have 1.26 sick days per month for every one sick day experienced by someone without diabetes (holding all other explanatory variables constant,) it is interesting to note that increased weight alone is not significantly associated with a greater number of sick days per month.

Conclusion

The results of this analysis can be used and implemented in a variety of ways. The primary motivation behind this study was to assess the general state of health issues affecting residents of the United States, as well as to understand the significant factors associated with those health problems in order to facilitate optimal design of health policy and programs. The optimality criterion in this case focused on minimization of unnecessary expenditures and maximization of service availability to the populations that are in greatest need of care.

The results of the exploratory analysis indicate that most U.S. residents have health coverage (please not that this **predates** that implementation of the Affordable Care Act,) and do not experience cost as a barrier to seeking health care. This suggests that a widespread federal health insurance initiative should focus on the minority of residents without coverage in order to best utilize public health care funds. On the other hand, many residents experience a variety of chronic health challenges, like asthma, diabetes, and arthritis, which support efforts toward prevention of these problems whenever possible.

The results of the regression analyses indicate that although the majority of residents have health coverage, those without health coverage as well as those with difficulty paying for health care are less likely to report fair to excellent health. Additionally, those with difficulty paying for health care are predicted to have a greater number of sick days per month than people with no difficulty paying for health care. This too suggests that federal initiatives should focus on the residents without health coverage and with difficulty paying for health related costs.

The regression results also support our intuition that chronic illness is associated with a higher incidence of ill health as well as greater numbers of sick days per month. Additionally, a

routine physical exam as well as regular exercise are significantly associated with better health as well as fewer sick days. This all suggests that a focus on wellness and prevention could greatly improve the general quality of life for a large segment of the population as well as result in a generally healthier populace.

The regression results for demographic effects show that increased levels of education are associated with generally better health as well as fewer sick days per month. Additionally, being unable to work or making less than $10,000 per year is strongly associated with poor health as well as a greater number of sick days. As such, a focus on increasing health care access for the uneducated as well as those living in poverty would be of benefit.

This study focused on the effect of a large set of explanatory variables on two response variables. This was an appropriate focus for the purpose of large scale or federal policy health care issues, however there are a number of ways this work could be extended to provide a finer or more local focus. In particular, a more refined analysis could be conducted involving selected interactions in order to find more complex effects between explanatory variables. This would be appropriate for exploring targeted interventions in a setting of constrained financial resources.

In summary, it is clear that health status is significantly affected by myriad factors, not all health related. In an environment of limited funds and significant need, it is necessary to target interventions wisely and appropriately. This results of this study suggested several ways in which to implement those interventions.

References

Ajmani, V.B. (2009). *Applied econometrics using the SAS system*, Hoboken, NJ: Wiley-Interscience.

American College of Physicians. (2009). *Controlling Health Care Costs While Promoting The Best Possible Health Outcomes*. Philadelphia: American College of Physicians.

American College of Physicians. (2011). *How Can Our Nation Conserve and Distribute Health Care Resources Effectively and Efficiently?* Philadelphia: American College of Physicians.

American College of Physicians. (2012). *Reforming Medicare in the Age of Deficit Reduction*. Philadelphia: American College of Physicians.

CDC. (2012). *Behavioral risk factor surveillance system, 2012 data.* Retrieved from http://www.cdc.gov/brfss/annual_data/annual_2012.html

Greene, W.H. (2011). *Econometric Analysis (7th Edition)*, Upper Saddle River, NJ.

Hilbe, J.M. (2011). *Negative Binomial Regression*, Cambridge, UK: Cambridge University Press.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression (2nd ed.)*. New York: Wiley.

Kuhn, M. (2013). *Applied Predictive Modeling.* New York, NY: Springer.

Myers, R.H., Montgomery, D.C., Vining, G.G. & Robinson, T.J. (2010). *Generalized Linear Models: with Applications in Engineering and the Sciences*, New York: Wiley.

Appendix

SAS Code for Binary Logistic Regression Analysis

```
data LogModel;
set THESIS.BECKYDATA;

*Create a role variable to split the dataset into train/
test sets;
u=uniform(123);
if u < 0.7 then role="TRAINING";
else role = "TESTING";

*Create dummy variables for the Logistic version of the
categorical response variable and the 35 categorical
explanatory variables;

if GENHLTH=1 or GENHLTH=2 or GENHLTH=3 or GENHLTH=4 then
GENHLTH_LogD=1;
else if GENHLTH=5 then GENHLTH_LogD=0;

*HLTHPLN1 Base category is HLTHPLN1=7 and 9;
if HLTHPLN1=1 then HLTHPLN1_Y=1; else HLTHPLN1_Y=0;
if HLTHPLN1=2 then HLTHPLN1_N=1; else HLTHPLN1_N=0;

*PERSDOC2 Base category is PERSDOC2=7 and 9;
if PERSDOC2=1 or PERSDOC2=2 then PERSDOC2_Y=1; else
PERSDOC2_Y=0;
if PERSDOC2=3 then PERSDOC2_N=1; else PERSDOC2_N=0;

*MEDCOST Base category is MEDCOST=7 and 9;
if MEDCOST=1 then MEDCOST_Y=1; else MEDCOST_Y=0;
if MEDCOST=2 then MEDCOST_N=1; else MEDCOST_N=0;

*HADMAM Base category is HADMAM=7 and 9;
if HADMAM=1 then HADMAM_Y=1; else HADMAM_Y=0;
if HADMAM=2 then HADMAM_N=1; else HADMAM_N=0;
if HADMAM=99 then HADMAM_NA=1; else HADMAM_NA=0;

*PROFEXAM Base category is PROFEXAM=7 and 9;
```

```
if PROFEXAM=1 then PROFEXAM_Y=1; else PROFEXAM_Y=0;
if PROFEXAM=2 then PROFEXAM_N=1; else PROFEXAM_N=0;
if PROFEXAM=99 then PROFEXAM_NA=1; else PROFEXAM_NA=0;

*HADPAP2 Base category is HADPAP2=7 and 9;
if HADPAP2=1 then HADPAP2_Y=1; else HADPAP2_Y=0;
if HADPAP2=2 then HADPAP2_N=1; else HADPAP2_N=0;
if HADPAP2=99 then HADPAP2_NA=1; else HADPAP2_NA=0;

*PSATEST1 Base category is PSATEST1=7 and 9;
if PSATEST1=1 then PSATEST1_Y=1; else PSATEST1_Y=0;
if PSATEST1=2 then PSATEST1_N=1; else PSATEST1_N=0;
if PSATEST1=99 then PSATEST1_NA=1; else PSATEST1_NA=0;

*HIVTST6 Base category is HIVTST6=7 and 9;
if HIVTST6=1 then HIVTST6_Y=1; else HIVTST6_Y=0;
if HIVTST6=2 then HIVTST6_N=1; else HIVTST6_N=0;

*_STATE Base category is _STATE=66 and 72;
if _STATE=1 then _STATE_Alabama=1; else _STATE_Alabama=0;
if _STATE=2 then _STATE_Alaska=1; else _STATE_Alaska=0;
if _STATE=4 then _STATE_Arizona=1; else _STATE_Arizona=0;
if _STATE=5 then _STATE_Arkansas=1; else _STATE_Arkansas=0;
if _STATE=6 then _STATE_California=1; else
_STATE_California=0;
if _STATE=8 then _STATE_Colorado=1; else _STATE_Colorado=0;
if _STATE=9 then _STATE_Connecticut=1; else
_STATE_Connecticut=0;
if _STATE=10 then _STATE_Delaware=1; else
_STATE_Delaware=0;
if _STATE=11 then _STATE_DC=1; else _STATE_DC=0;
if _STATE=12 then _STATE_Florida=1; else _STATE_Florida=0;
if _STATE=13 then _STATE_Georgia=1; else _STATE_Georgia=0;
if _STATE=15 then _STATE_Hawaii=1; else _STATE_Hawaii=0;
if _STATE=16 then _STATE_Idaho=1; else _STATE_Idaho=0;
if _STATE=17 then _STATE_Illinois=1; else
_STATE_Illinois=0;
if _STATE=18 then _STATE_Indiana=1; else _STATE_Indiana=0;
if _STATE=19 then _STATE_Iowa=1; else _STATE_Iowa=0;
if _STATE=20 then _STATE_Kansas=1; else _STATE_Kansas=0;
```

```
if _STATE=21 then _STATE_Kentucky=1; else
_STATE_Kentucky=0;
if _STATE=22 then _STATE_Louisiana=1; else
_STATE_Louisiana=0;
if _STATE=23 then _STATE_Maine=1; else _STATE_Maine=0;
if _STATE=24 then _STATE_Maryland=1; else
_STATE_Maryland=0;
if _STATE=25 then _STATE_Massachusetts=1; else
_STATE_Massachusetts=0;
if _STATE=26 then _STATE_Michigan=1; else
_STATE_Michigan=0;
if _STATE=27 then _STATE_Minnesota=1; else
_STATE_Minnesota=0;
if _STATE=28 then _STATE_Mississippi=1; else
_STATE_Mississippi=0;
if _STATE=29 then _STATE_Missouri=1; else
_STATE_Missouri=0;
if _STATE=30 then _STATE_Montana=1; else _STATE_Montana=0;
if _STATE=31 then _STATE_Nebraska=1; else
_STATE_Nebraska=0;
if _STATE=32 then _STATE_Nevada=1; else _STATE_Nevada=0;
if _STATE=33 then _STATE_NewHampshire=1; else
_STATE_NewHampshire=0;
if _STATE=34 then _STATE_NewJersey=1; else
_STATE_NewJersey=0;
if _STATE=35 then _STATE_NewMexico=1; else
_STATE_NewMexico=0;
if _STATE=36 then _STATE_NewYork=1; else _STATE_NewYork=0;
if _STATE=37 then _STATE_NorthCarolina=1; else
_STATE_NorthCarolina=0;
if _STATE=38 then _STATE_NorthDakota=1; else
_STATE_NorthDakota=0;
if _STATE=39 then _STATE_Ohio=1; else _STATE_Ohio=0;
if _STATE=40 then _STATE_Oklahoma=1; else
_STATE_Oklahoma=0;
if _STATE=41 then _STATE_Oregon=1; else _STATE_Oregon=0;
if _STATE=42 then _STATE_Pennsylvania=1; else
_STATE_Pennsylvania=0;
if _STATE=44 then _STATE_RhodeIsland=1; else
_STATE_RhodeIsland=0;
```

```
if _STATE=45 then _STATE_SouthCarolina=1; else
_STATE_SouthCarolina=0;
if _STATE=46 then _STATE_SouthDakota=1; else
_STATE_SouthDakota=0;
if _STATE=47 then _STATE_Tennesee=1; else
_STATE_Tennesee=0;
if _STATE=48 then _STATE_Texas=1; else _STATE_Texas=0;
if _STATE=49 then _STATE_Utah=1; else _STATE_Utah=0;
if _STATE=50 then _STATE_Vermont=1; else _STATE_Vermont=0;
if _STATE=51 then _STATE_Virginia=1; else
_STATE_Virginia=0;
if _STATE=53 then _STATE_Washington=1; else
_STATE_Washington=0;
if _STATE=54 then _STATE_WestVirginia=1; else
_STATE_WestVirginia=0;
if _STATE=55 then _STATE_Wisconsin=1; else
_STATE_Wisconsin=0;
if _STATE=56 then _STATE_Wyoming=1; else _STATE_Wyoming=0;

*VETERAN3 Base category is VETERAN3=7 and 9;
if VETERAN3=1 then VETERAN3_Y=1; else VETERAN3_Y=0;
if VETERAN3=2 then VETERAN3_N=1; else VETERAN3_N=0;

*MARITAL Base category is MARITAL=9;
if MARITAL=1 then MARITAL_MARRIED=1; else
MARITAL_MARRIED=0;
if MARITAL=2 then MARITAL_DIVORCED=1; else
MARITAL_DIVORCED=0;
if MARITAL=3 then MARITAL_WIDOWED=1; else
MARITAL_WIDOWED=0;
if MARITAL=4 then MARITAL_SEPARATED=1; else
MARITAL_SEPARATED=0;
if MARITAL=5 then MARITAL_NEVERMARRIED=1; else
MARITAL_NEVERMARRIED=0;
if MARITAL=6 then MARITAL_UNMARRIEDCOUPLE=1; else
MARITAL_UNMARRIEDCOUPLE=0;

*CVDSTRK3 Base category is CVDSTRK3=7 and 9;
if CVDSTRK3=1 then CVDSTRK3_Y=1; else CVDSTRK3_Y=0;
if CVDSTRK3=2 then CVDSTRK3_N=1; else CVDSTRK3_N=0;
```

```
*ASTHMA3 Base category is ASTHMA3=7 and 9;
if ASTHMA3=1 then ASTHMA3_Y=1; else ASTHMA3_Y=0;
if ASTHMA3=2 then ASTHMA3_N=1; else ASTHMA3_N=0;

*CHCSCNCR Base category is CHCSCNCR=7 and 9;
if CHCSCNCR=1 then CHCSCNCR_Y=1; else CHCSCNCR_Y=0;
if CHCSCNCR=2 then CHCSCNCR_N=1; else CHCSCNCR_N=0;

*CHCOCNCR Base category is CHCOCNCR=7 and 9;
if CHCOCNCR=1 then CHCOCNCR_Y=1; else CHCOCNCR_Y=0;
if CHCOCNCR=2 then CHCOCNCR_N=1; else CHCOCNCR_N=0;

*CHCCOPD1 Base category is CHCCOPD1=7 and 9;
if CHCCOPD1=1 then CHCCOPD1_Y=1; else CHCCOPD1_Y=0;
if CHCCOPD1=2 then CHCCOPD1_N=1; else CHCCOPD1_N=0;

*CHECKUP1 Base category is CHECKUP1=7, 8 and 9;
if CHECKUP1=1 then CHECKUP1_LT1=1; else CHECKUP1_LT1=0;
if CHECKUP1=2 or CHECKUP1=3 then CHECKUP1_1T5=1; else
CHECKUP1_1T5=0;
if CHECKUP1=4 then CHECKUP1_GT5=1; else CHECKUP1_GT5=0;

*EXERANY2 Base category is EXERANY2=7 and 9;
if EXERANY2=1 then EXERANY2_Y=1; else EXERANY2_Y=0;
if EXERANY2=2 then EXERANY2_N=1; else EXERANY2_N=0;

*CVDINFR4 Base category is CVDINFR4=7 and 9;
if CVDINFR4=1 then CVDINFR4_Y=1; else CVDINFR4_Y=0;
if CVDINFR4=2 then CVDINFR4_N=1; else CVDINFR4_N=0;

*CVDCRHD4 Base category is CVDCRHD4=7 and 9;
if CVDCRHD4=1 then CVDCRHD4_Y=1; else CVDCRHD4_Y=0;
if CVDCRHD4=2 then CVDCRHD4_N=1; else CVDCRHD4_N=0;

*EDUCA Base category is EDUCA=9;
if EDUCA=1 then EDUCA_None=1; else EDUCA_None=0;
if EDUCA=2 then EDUCA_Elem=1; else EDUCA_Elem=0;
if EDUCA=3 then EDUCA_SomeHS=1; else EDUCA_SomeHS=0;
if EDUCA=4 then EDUCA_GradHS=1; else EDUCA_GradHS=0;
```

```
if EDUCA=5 then EDUCA_SomeColl=1; else EDUCA_SomeColl=0;
if EDUCA=6 then EDUCA_GradColl=1; else EDUCA_GradColl=0;

*EMPLOY Base category is EMPLOY=9;
if EMPLOY=1 or EMPLOY=2 then EMPLOY_SelfOrWages=1; else
EMPLOY_SelfOrWages=0;
if EMPLOY=3 or EMPLOY=4 then EMPLOY_Unemployed=1; else
EMPLOY_Unemployed=0;
if EMPLOY=5 then EMPLOY_Homemaker=1; else
EMPLOY_Homemaker=0;
if EMPLOY=6 then EMPLOY_Student=1; else EMPLOY_Student=0;
if EMPLOY=7 then EMPLOY_Retired=1; else EMPLOY_Retired=0;
if EMPLOY=8 then EMPLOY_Unable=1; else EMPLOY_Unable=0;

*INCOME2 Base category is INCOME2=77 and 99;
if INCOME2=01 then INCOME2_LT10K=1; else INCOME2_LT10K=0;
if INCOME2=02 or INCOME2=03 or INCOME2=04 then
INCOME2_10KLT25K=1; else INCOME2_10KLT25K=0;
if INCOME2=05 or INCOME2=06 then INCOME2_25KLT50K=1; else
INCOME2_25KLT50K=0;
if INCOME2=07 or INCOME2=08 then INCOME2_50KPLUS=1; else
INCOME2_50KPLUS=0;

*PREGNANT Base category is PREGNANT=7 and 9;
if PREGNANT=1 then PREGNANT_Y=1; else PREGNANT_Y=0;
if PREGNANT=2 then PREGNANT_N=1; else PREGNANT_N=0;
if PREGNANT=99 then PREGNANT_NA=1; else PREGNANT_NA=0;

*QLACTLM2 Base category is QLACTLM2=7 and 9;
if QLACTLM2=1 then QLACTLM2_Y=1; else QLACTLM2_Y=0;
if QLACTLM2=2 then QLACTLM2_N=1; else QLACTLM2_N=0;

*USEEQUIP Base category is USEEQUIP=7 and 9;
if USEEQUIP=1 then USEEQUIP_Y=1; else USEEQUIP_Y=0;
if USEEQUIP=2 then USEEQUIP_N=1; else USEEQUIP_N=0;

*USENOW3 Base category is USENOW3=7 and 9;
if USENOW3=1 then USENOW3_Daily=1; else USENOW3_Daily=0;
if USENOW3=2 then USENOW3_SomeDays=1; else
USENOW3_SomeDays=0;
```

```
if USENOW3=3 then USENOW3_None=1; else USENOW3_None=0;

*FLUSHOT5 Base category is FLUSHOT5=7 and 9;
if FLUSHOT5=1 then FLUSHOT5_Y=1; else FLUSHOT5_Y=0;
if FLUSHOT5=2 then FLUSHOT5_N=1; else FLUSHOT5_N=0;

*PNEUVAC3 Base category is PNEUVAC3=7 and 9;
if PNEUVAC3=1 then PNEUVAC3_Y=1; else PNEUVAC3_Y=0;
if PNEUVAC3=2 then PNEUVAC3_N=1; else PNEUVAC3_N=0;

*SEATBELT Base category is SEATBELT=7 and 9;
if SEATBELT=1 then SEATBELT_Always=1; else
SEATBELT_Always=0;
if SEATBELT=2 or SEATBELT=3 then SEATBELT_Sometimes=1; else
SEATBELT_Sometimes=0;
if SEATBELT=4 or SEATBELT=5 then SEATBELT_SeldomNever=1;
else SEATBELT_SeldomNever=0;

*HAVARTH3 Base category is HAVARTH3=7 and 9;
if HAVARTH3=1 then HAVARTH3_Y=1; else HAVARTH3_Y=0;
if HAVARTH3=2 then HAVARTH3_N=1; else HAVARTH3_N=0;

*ADDEPEV2 Base category is ADDEPEV2=7 and 9;
if ADDEPEV2=1 then ADDEPEV2_Y=1; else ADDEPEV2_Y=0;
if ADDEPEV2=2 then ADDEPEV2_N=1; else ADDEPEV2_N=0;

*CHCKIDNY Base category is CHCKIDNY=7 and 9;
if CHCKIDNY=1 then CHCKIDNY_Y=1; else CHCKIDNY_Y=0;
if CHCKIDNY=2 then CHCKIDNY_N=1; else CHCKIDNY_N=0;

*CHCVISN1 Base category is CHCVISN1=7 and 9;
if CHCVISN1=1 or CHCVISN1=3 then CHCVISN1_Y=1; else
CHCVISN1_Y=0;
if CHCVISN1=2 then CHCVISN1_N=1; else CHCVISN1_N=0;

*DIABETE3 Base category is DIABETE3=7 and 9;
if DIABETE3=1 or DIABETE3=2 then DIABETE3_Y=1; else
DIABETE3_Y=0;
if DIABETE3=3 or DIABETE3=4 then DIABETE3_N=1; else
DIABETE3_N=0;
```

```
*LASTDEN3 Base category is LASTDEN3=7 and 9;
if LASTDEN3=1 then LASTDEN3_LT1=1; else LASTDEN3_LT1=0;
if LASTDEN3=2 or LASTDEN3=3 then LASTDEN3_1LT5=1; else
LASTDEN3_1LT5=0;
if LASTDEN3=4 then LASTDEN3_GTE5=1; else LASTDEN3_GTE5=0;
if LASTDEN3=8 then LASTDEN3_Never=1; else LASTDEN3_Never=0;

*Now, subset the training set to do the analysis;
data LogTraining;
     set LogModel;
     where role='TRAINING';
run;

title "Logistic Model";
PROC LOGISTIC Data=LogTraining PLOTS(MAXPOINTS=NONE)
PLOTS(ONLY)=ROC;
     MODEL GENHLTH_LogD(EVENT='1')=HLTHPLN1_Y HLTHPLN1_N
PERSDOC2_Y PERSDOC2_N MEDCOST_Y MEDCOST_N
     HADMAM_Y HADMAM_N HADMAM_NA PROFEXAM_Y PROFEXAM_N
PROFEXAM_NA HADPAP2_Y HADPAP2_N HADPAP2_NA
     PSATEST1_Y PSATEST1_N PSATEST1_NA HIVTST6_Y HIVTST6_N
_STATE_Alabama _STATE_Alaska _STATE_Arizona
     _STATE_Arkansas _STATE_California _STATE_Colorado
_STATE_Connecticut _STATE_Delaware _STATE_DC
     _STATE_Florida _STATE_Georgia _STATE_Hawaii
_STATE_Idaho _STATE_Illinois _STATE_Indiana _STATE_Iowa
     _STATE_Kansas _STATE_Kentucky _STATE_Louisiana
_STATE_Maine _STATE_Maryland _STATE_Massachusetts
     _STATE_Michigan _STATE_Minnesota _STATE_Mississippi
_STATE_Missouri _STATE_Montana _STATE_Nebraska
     _STATE_Nevada _STATE_NewHampshire _STATE_NewJersey
_STATE_NewMexico _STATE_NewYork _STATE_NorthCarolina
     _STATE_NorthDakota _STATE_Ohio _STATE_Oklahoma
_STATE_Oregon _STATE_Pennsylvania _STATE_RhodeIsland
     _STATE_SouthCarolina _STATE_SouthDakota _STATE_Tennesee
_STATE_Texas _STATE_Utah _STATE_Vermont
     _STATE_Virginia _STATE_Washington _STATE_WestVirginia
_STATE_Wisconsin _STATE_Wyoming VETERAN3_Y
```

```
     VETERAN3_N MARITAL_MARRIED MARITAL_DIVORCED
MARITAL_WIDOWED MARITAL_SEPARATED MARITAL_NEVERMARRIED
     MARITAL_UNMARRIEDCOUPLE CVDSTRK3_Y CVDSTRK3_N ASTHMA3_Y
ASTHMA3_N CHCSCNCR_Y CHCSCNCR_N
     CHCOCNCR_Y CHCOCNCR_N CHCCOPD1_Y CHCCOPD1_N
CHECKUP1_LT1 CHECKUP1_1T5 CHECKUP1_GT5 EXERANY2_Y
     EXERANY2_N CVDINFR4_Y CVDINFR4_N CVDCRHD4_Y CVDCRHD4_N
EDUCA_None EDUCA_Elem EDUCA_SomeHS
     EDUCA_GradHS EDUCA_SomeColl EDUCA_GradColl
EMPLOY_SelfOrWages EMPLOY_Unemployed EMPLOY_Homemaker
     EMPLOY_Student EMPLOY_Retired EMPLOY_Unable
INCOME2_LT10K INCOME2_10KLT25K INCOME2_25KLT50K
     INCOME2_50KPLUS PREGNANT_Y PREGNANT_N PREGNANT_NA
QLACTLM2_Y QLACTLM2_N USEEQUIP_Y USEEQUIP_N
     USENOW3_Daily USENOW3_SomeDays USENOW3_None FLUSHOT5_Y
FLUSHOT5_N PNEUVAC3_Y PNEUVAC3_N
     SEATBELT_Always SEATBELT_Sometimes SEATBELT_SeldomNever
HAVARTH3_Y HAVARTH3_N ADDEPEV2_Y
     ADDEPEV2_N CHCKIDNY_Y CHCKIDNY_N CHCVISN1_Y CHCVISN1_N
DIABETE3_Y DIABETE3_N LASTDEN3_LT1
     LASTDEN3_1LT5 LASTDEN3_GTE5 LASTDEN3_Never AGE CHILDREN
WEIGHT2 FALL12MN / LACKFIT CTABLE PPROB=.5
SELECTION=BACKWARD;
RUN;
```

SAS Code for Ordinal Logistic Regression Analysis

```
data OrdLogModel;
set THESIS.BECKYDATA;

*Create a role variable to split the dataset into train/
test sets;
u=uniform(123);
if u < 0.7 then role="TRAINING";
else role = "TESTING";

*Create dummy variables for the 35 categorical explanatory
variables;

*HLTHPLN1 Base category is HLTHPLN1=7 and 9;
```

```
if HLTHPLN1=1 then HLTHPLN1_Y=1; else HLTHPLN1_Y=0;
if HLTHPLN1=2 then HLTHPLN1_N=1; else HLTHPLN1_N=0;

*PERSDOC2 Base category is PERSDOC2=7 and 9;
if PERSDOC2=1 or PERSDOC2=2 then PERSDOC2_Y=1; else
PERSDOC2_Y=0;
if PERSDOC2=3 then PERSDOC2_N=1; else PERSDOC2_N=0;

*MEDCOST Base category is MEDCOST=7 and 9;
if MEDCOST=1 then MEDCOST_Y=1; else MEDCOST_Y=0;
if MEDCOST=2 then MEDCOST_N=1; else MEDCOST_N=0;

*HADMAM Base category is HADMAM=7 and 9;
if HADMAM=1 then HADMAM_Y=1; else HADMAM_Y=0;
if HADMAM=2 then HADMAM_N=1; else HADMAM_N=0;
if HADMAM=99 then HADMAM_NA=1; else HADMAM_NA=0;

*PROFEXAM Base category is PROFEXAM=7 and 9;
if PROFEXAM=1 then PROFEXAM_Y=1; else PROFEXAM_Y=0;
if PROFEXAM=2 then PROFEXAM_N=1; else PROFEXAM_N=0;
if PROFEXAM=99 then PROFEXAM_NA=1; else PROFEXAM_NA=0;

*HADPAP2 Base category is HADPAP2=7 and 9;
if HADPAP2=1 then HADPAP2_Y=1; else HADPAP2_Y=0;
if HADPAP2=2 then HADPAP2_N=1; else HADPAP2_N=0;
if HADPAP2=99 then HADPAP2_NA=1; else HADPAP2_NA=0;

*PSATEST1 Base category is PSATEST1=7 and 9;
if PSATEST1=1 then PSATEST1_Y=1; else PSATEST1_Y=0;
if PSATEST1=2 then PSATEST1_N=1; else PSATEST1_N=0;
if PSATEST1=99 then PSATEST1_NA=1; else PSATEST1_NA=0;

*HIVTST6 Base category is HIVTST6=7 and 9;
if HIVTST6=1 then HIVTST6_Y=1; else HIVTST6_Y=0;
if HIVTST6=2 then HIVTST6_N=1; else HIVTST6_N=0;

*_STATE Base category is _STATE=66 and 72;
if _STATE=1 then _STATE_Alabama=1; else _STATE_Alabama=0;
if _STATE=2 then _STATE_Alaska=1; else _STATE_Alaska=0;
if _STATE=4 then _STATE_Arizona=1; else _STATE_Arizona=0;
```

```
if _STATE=5 then _STATE_Arkansas=1; else _STATE_Arkansas=0;
if _STATE=6 then _STATE_California=1; else
_STATE_California=0;
if _STATE=8 then _STATE_Colorado=1; else _STATE_Colorado=0;
if _STATE=9 then _STATE_Connecticut=1; else
_STATE_Connecticut=0;
if _STATE=10 then _STATE_Delaware=1; else
_STATE_Delaware=0;
if _STATE=11 then _STATE_DC=1; else _STATE_DC=0;
if _STATE=12 then _STATE_Florida=1; else _STATE_Florida=0;
if _STATE=13 then _STATE_Georgia=1; else _STATE_Georgia=0;
if _STATE=15 then _STATE_Hawaii=1; else _STATE_Hawaii=0;
if _STATE=16 then _STATE_Idaho=1; else _STATE_Idaho=0;
if _STATE=17 then _STATE_Illinois=1; else
_STATE_Illinois=0;
if _STATE=18 then _STATE_Indiana=1; else _STATE_Indiana=0;
if _STATE=19 then _STATE_Iowa=1; else _STATE_Iowa=0;
if _STATE=20 then _STATE_Kansas=1; else _STATE_Kansas=0;
if _STATE=21 then _STATE_Kentucky=1; else
_STATE_Kentucky=0;
if _STATE=22 then _STATE_Louisiana=1; else
_STATE_Louisiana=0;
if _STATE=23 then _STATE_Maine=1; else _STATE_Maine=0;
if _STATE=24 then _STATE_Maryland=1; else
_STATE_Maryland=0;
if _STATE=25 then _STATE_Massachusetts=1; else
_STATE_Massachusetts=0;
if _STATE=26 then _STATE_Michigan=1; else
_STATE_Michigan=0;
if _STATE=27 then _STATE_Minnesota=1; else
_STATE_Minnesota=0;
if _STATE=28 then _STATE_Mississippi=1; else
_STATE_Mississippi=0;
if _STATE=29 then _STATE_Missouri=1; else
_STATE_Missouri=0;
if _STATE=30 then _STATE_Montana=1; else _STATE_Montana=0;
if _STATE=31 then _STATE_Nebraska=1; else
_STATE_Nebraska=0;
if _STATE=32 then _STATE_Nevada=1; else _STATE_Nevada=0;
```

```
if _STATE=33 then _STATE_NewHampshire=1; else
_STATE_NewHampshire=0;
if _STATE=34 then _STATE_NewJersey=1; else
_STATE_NewJersey=0;
if _STATE=35 then _STATE_NewMexico=1; else
_STATE_NewMexico=0;
if _STATE=36 then _STATE_NewYork=1; else _STATE_NewYork=0;
if _STATE=37 then _STATE_NorthCarolina=1; else
_STATE_NorthCarolina=0;
if _STATE=38 then _STATE_NorthDakota=1; else
_STATE_NorthDakota=0;
if _STATE=39 then _STATE_Ohio=1; else _STATE_Ohio=0;
if _STATE=40 then _STATE_Oklahoma=1; else
_STATE_Oklahoma=0;
if _STATE=41 then _STATE_Oregon=1; else _STATE_Oregon=0;
if _STATE=42 then _STATE_Pennsylvania=1; else
_STATE_Pennsylvania=0;
if _STATE=44 then _STATE_RhodeIsland=1; else
_STATE_RhodeIsland=0;
if _STATE=45 then _STATE_SouthCarolina=1; else
_STATE_SouthCarolina=0;
if _STATE=46 then _STATE_SouthDakota=1; else
_STATE_SouthDakota=0;
if _STATE=47 then _STATE_Tennesee=1; else
_STATE_Tennesee=0;
if _STATE=48 then _STATE_Texas=1; else _STATE_Texas=0;
if _STATE=49 then _STATE_Utah=1; else _STATE_Utah=0;
if _STATE=50 then _STATE_Vermont=1; else _STATE_Vermont=0;
if _STATE=51 then _STATE_Virginia=1; else
_STATE_Virginia=0;
if _STATE=53 then _STATE_Washington=1; else
_STATE_Washington=0;
if _STATE=54 then _STATE_WestVirginia=1; else
_STATE_WestVirginia=0;
if _STATE=55 then _STATE_Wisconsin=1; else
_STATE_Wisconsin=0;
if _STATE=56 then _STATE_Wyoming=1; else _STATE_Wyoming=0;

*VETERAN3 Base category is VETERAN3=7 and 9;
if VETERAN3=1 then VETERAN3_Y=1; else VETERAN3_Y=0;
```

```
if VETERAN3=2 then VETERAN3_N=1; else VETERAN3_N=0;

*MARITAL Base category is MARITAL=9;
if MARITAL=1 then MARITAL_MARRIED=1; else
MARITAL_MARRIED=0;
if MARITAL=2 then MARITAL_DIVORCED=1; else
MARITAL_DIVORCED=0;
if MARITAL=3 then MARITAL_WIDOWED=1; else
MARITAL_WIDOWED=0;
if MARITAL=4 then MARITAL_SEPARATED=1; else
MARITAL_SEPARATED=0;
if MARITAL=5 then MARITAL_NEVERMARRIED=1; else
MARITAL_NEVERMARRIED=0;
if MARITAL=6 then MARITAL_UNMARRIEDCOUPLE=1; else
MARITAL_UNMARRIEDCOUPLE=0;

*CVDSTRK3 Base category is CVDSTRK3=7 and 9;
if CVDSTRK3=1 then CVDSTRK3_Y=1; else CVDSTRK3_Y=0;
if CVDSTRK3=2 then CVDSTRK3_N=1; else CVDSTRK3_N=0;

*ASTHMA3 Base category is ASTHMA3=7 and 9;
if ASTHMA3=1 then ASTHMA3_Y=1; else ASTHMA3_Y=0;
if ASTHMA3=2 then ASTHMA3_N=1; else ASTHMA3_N=0;

*CHCSCNCR Base category is CHCSCNCR=7 and 9;
if CHCSCNCR=1 then CHCSCNCR_Y=1; else CHCSCNCR_Y=0;
if CHCSCNCR=2 then CHCSCNCR_N=1; else CHCSCNCR_N=0;

*CHCOCNCR Base category is CHCOCNCR=7 and 9;
if CHCOCNCR=1 then CHCOCNCR_Y=1; else CHCOCNCR_Y=0;
if CHCOCNCR=2 then CHCOCNCR_N=1; else CHCOCNCR_N=0;

*CHCCOPD1 Base category is CHCCOPD1=7 and 9;
if CHCCOPD1=1 then CHCCOPD1_Y=1; else CHCCOPD1_Y=0;
if CHCCOPD1=2 then CHCCOPD1_N=1; else CHCCOPD1_N=0;

*CHECKUP1 Base category is CHECKUP1=7, 8 and 9;
if CHECKUP1=1 then CHECKUP1_LT1=1; else CHECKUP1_LT1=0;
if CHECKUP1=2 or CHECKUP1=3 then CHECKUP1_1T5=1; else
CHECKUP1_1T5=0;
```

```
if CHECKUP1=4 then CHECKUP1_GT5=1; else CHECKUP1_GT5=0;

*EXERANY2 Base category is EXERANY2=7 and 9;
if EXERANY2=1 then EXERANY2_Y=1; else EXERANY2_Y=0;
if EXERANY2=2 then EXERANY2_N=1; else EXERANY2_N=0;

*CVDINFR4 Base category is CVDINFR4=7 and 9;
if CVDINFR4=1 then CVDINFR4_Y=1; else CVDINFR4_Y=0;
if CVDINFR4=2 then CVDINFR4_N=1; else CVDINFR4_N=0;

*CVDCRHD4 Base category is CVDCRHD4=7 and 9;
if CVDCRHD4=1 then CVDCRHD4_Y=1; else CVDCRHD4_Y=0;
if CVDCRHD4=2 then CVDCRHD4_N=1; else CVDCRHD4_N=0;

*EDUCA Base category is EDUCA=9;
if EDUCA=1 then EDUCA_None=1; else EDUCA_None=0;
if EDUCA=2 then EDUCA_Elem=1; else EDUCA_Elem=0;
if EDUCA=3 then EDUCA_SomeHS=1; else EDUCA_SomeHS=0;
if EDUCA=4 then EDUCA_GradHS=1; else EDUCA_GradHS=0;
if EDUCA=5 then EDUCA_SomeColl=1; else EDUCA_SomeColl=0;
if EDUCA=6 then EDUCA_GradColl=1; else EDUCA_GradColl=0;

*EMPLOY Base category is EMPLOY=9;
if EMPLOY=1 or EMPLOY=2 then EMPLOY_SelfOrWages=1; else
EMPLOY_SelfOrWages=0;
if EMPLOY=3 or EMPLOY=4 then EMPLOY_Unemployed=1; else
EMPLOY_Unemployed=0;
if EMPLOY=5 then EMPLOY_Homemaker=1; else
EMPLOY_Homemaker=0;
if EMPLOY=6 then EMPLOY_Student=1; else EMPLOY_Student=0;
if EMPLOY=7 then EMPLOY_Retired=1; else EMPLOY_Retired=0;
if EMPLOY=8 then EMPLOY_Unable=1; else EMPLOY_Unable=0;

*INCOME2 Base category is INCOME2=77 and 99;
if INCOME2=01 then INCOME2_LT10K=1; else INCOME2_LT10K=0;
if INCOME2=02 or INCOME2=03 or INCOME2=04 then
INCOME2_10KLT25K=1; else INCOME2_10KLT25K=0;
if INCOME2=05 or INCOME2=06 then INCOME2_25KLT50K=1; else
INCOME2_25KLT50K=0;
```

```
if INCOME2=07 or INCOME2=08 then INCOME2_50KPLUS=1; else
INCOME2_50KPLUS=0;

*PREGNANT Base category is PREGNANT=7 and 9;
if PREGNANT=1 then PREGNANT_Y=1; else PREGNANT_Y=0;
if PREGNANT=2 then PREGNANT_N=1; else PREGNANT_N=0;
if PREGNANT=99 then PREGNANT_NA=1; else PREGNANT_NA=0;

*QLACTLM2 Base category is QLACTLM2=7 and 9;
if QLACTLM2=1 then QLACTLM2_Y=1; else QLACTLM2_Y=0;
if QLACTLM2=2 then QLACTLM2_N=1; else QLACTLM2_N=0;

*USEEQUIP Base category is USEEQUIP=7 and 9;
if USEEQUIP=1 then USEEQUIP_Y=1; else USEEQUIP_Y=0;
if USEEQUIP=2 then USEEQUIP_N=1; else USEEQUIP_N=0;

*USENOW3 Base category is USENOW3=7 and 9;
if USENOW3=1 then USENOW3_Daily=1; else USENOW3_Daily=0;
if USENOW3=2 then USENOW3_SomeDays=1; else
USENOW3_SomeDays=0;
if USENOW3=3 then USENOW3_None=1; else USENOW3_None=0;

*FLUSHOT5 Base category is FLUSHOT5=7 and 9;
if FLUSHOT5=1 then FLUSHOT5_Y=1; else FLUSHOT5_Y=0;
if FLUSHOT5=2 then FLUSHOT5_N=1; else FLUSHOT5_N=0;

*PNEUVAC3 Base category is PNEUVAC3=7 and 9;
if PNEUVAC3=1 then PNEUVAC3_Y=1; else PNEUVAC3_Y=0;
if PNEUVAC3=2 then PNEUVAC3_N=1; else PNEUVAC3_N=0;

*SEATBELT Base category is SEATBELT=7 and 9;
if SEATBELT=1 then SEATBELT_Always=1; else
SEATBELT_Always=0;
if SEATBELT=2 or SEATBELT=3 then SEATBELT_Sometimes=1; else
SEATBELT_Sometimes=0;
if SEATBELT=4 or SEATBELT=5 then SEATBELT_SeldomNever=1;
else SEATBELT_SeldomNever=0;

*HAVEARTH3 Base category is HAVARTH3=7 and 9;
if HAVARTH3=1 then HAVARTH3_Y=1; else HAVARTH3_Y=0;
```

```
if HAVARTH3=2 then HAVARTH3_N=1; else HAVARTH3_N=0;

*ADDEPEV2 Base category is ADDEPEV2=7 and 9;
if ADDEPEV2=1 then ADDEPEV2_Y=1; else ADDEPEV2_Y=0;
if ADDEPEV2=2 then ADDEPEV2_N=1; else ADDEPEV2_N=0;

*CHCKIDNY Base category is CHCKIDNY=7 and 9;
if CHCKIDNY=1 then CHCKIDNY_Y=1; else CHCKIDNY_Y=0;
if CHCKIDNY=2 then CHCKIDNY_N=1; else CHCKIDNY_N=0;

*CHCVISN1 Base category is CHCVISN1=7 and 9;
if CHCVISN1=1 or CHCVISN1=3 then CHCVISN1_Y=1; else
CHCVISN1_Y=0;
if CHCVISN1=2 then CHCVISN1_N=1; else CHCVISN1_N=0;

*DIABETE3 Base category is DIABETE3=7 and 9;
if DIABETE3=1 or DIABETE3=2 then DIABETE3_Y=1; else
DIABETE3_Y=0;
if DIABETE3=3 or DIABETE3=4 then DIABETE3_N=1; else
DIABETE3_N=0;

*LASTDEN3 Base category is LASTDEN3=7 and 9;
if LASTDEN3=1 then LASTDEN3_LT1=1; else LASTDEN3_LT1=0;
if LASTDEN3=2 or LASTDEN3=3 then LASTDEN3_1LT5=1; else
LASTDEN3_1LT5=0;
if LASTDEN3=4 then LASTDEN3_GTE5=1; else LASTDEN3_GTE5=0;
if LASTDEN3=8 then LASTDEN3_Never=1; else LASTDEN3_Never=0;

*Now, subset the training set to do the analysis;
data OrdLogTraining;
     set OrdLogModel;
     where role='TRAINING';
run;

title "Ordinal Logistic Model";
PROC LOGISTIC Data=OrdLogTraining;
     MODEL GENHLTH=HLTHPLN1_Y HLTHPLN1_N PERSDOC2_Y
PERSDOC2_N MEDCOST_Y MEDCOST_N
     HADMAM_Y HADMAM_N HADMAM_NA PROFEXAM_Y PROFEXAM_N
PROFEXAM_NA HADPAP2_Y HADPAP2_N HADPAP2_NA
```

```
     PSATEST1_Y PSATEST1_N PSATEST1_NA HIVTST6_Y HIVTST6_N
_STATE_Alabama _STATE_Alaska _STATE_Arizona
     _STATE_Arkansas _STATE_California _STATE_Colorado
_STATE_Connecticut _STATE_Delaware _STATE_DC
     _STATE_Florida _STATE_Georgia _STATE_Hawaii
_STATE_Idaho _STATE_Illinois _STATE_Indiana _STATE_Iowa
     _STATE_Kansas _STATE_Kentucky _STATE_Louisiana
_STATE_Maine _STATE_Maryland _STATE_Massachusetts
     _STATE_Michigan _STATE_Minnesota _STATE_Mississippi
_STATE_Missouri _STATE_Montana _STATE_Nebraska
     _STATE_Nevada _STATE_NewHampshire _STATE_NewJersey
_STATE_NewMexico _STATE_NewYork _STATE_NorthCarolina
     _STATE_NorthDakota _STATE_Ohio _STATE_Oklahoma
_STATE_Oregon _STATE_Pennsylvania _STATE_RhodeIsland
     _STATE_SouthCarolina _STATE_SouthDakota _STATE_Tennesee
_STATE_Texas _STATE_Utah _STATE_Vermont
     _STATE_Virginia _STATE_Washington _STATE_WestVirginia
_STATE_Wisconsin _STATE_Wyoming VETERAN3_Y
     VETERAN3_N MARITAL_MARRIED MARITAL_DIVORCED
MARITAL_WIDOWED MARITAL_SEPARATED MARITAL_NEVERMARRIED
     MARITAL_UNMARRIEDCOUPLE CVDSTRK3_Y CVDSTRK3_N ASTHMA3_Y
ASTHMA3_N CHCSCNCR_Y CHCSCNCR_N
     CHCOCNCR_Y CHCOCNCR_N CHCCOPD1_Y CHCCOPD1_N
CHECKUP1_LT1 CHECKUP1_1T5 CHECKUP1_GT5 EXERANY2_Y
     EXERANY2_N CVDINFR4_Y CVDINFR4_N CVDCRHD4_Y CVDCRHD4_N
EDUCA_None EDUCA_Elem EDUCA_SomeHS
     EDUCA_GradHS EDUCA_SomeColl EDUCA_GradColl
EMPLOY_SelfOrWages EMPLOY_Unemployed EMPLOY_Homemaker
     EMPLOY_Student EMPLOY_Retired EMPLOY_Unable
INCOME2_LT10K INCOME2_10KLT25K INCOME2_25KLT50K
     INCOME2_50KPLUS PREGNANT_Y PREGNANT_N PREGNANT_NA
QLACTLM2_Y QLACTLM2_N USEEQUIP_Y USEEQUIP_N
     USENOW3_Daily USENOW3_SomeDays USENOW3_None FLUSHOT5_Y
FLUSHOT5_N PNEUVAC3_Y PNEUVAC3_N
     SEATBELT_Always SEATBELT_Sometimes SEATBELT_SeldomNever
HAVARTH3_Y HAVARTH3_N ADDEPEV2_Y
     ADDEPEV2_N CHCKIDNY_Y CHCKIDNY_N CHCVISN1_Y CHCVISN1_N
DIABETE3_Y DIABETE3_N LASTDEN3_LT1
     LASTDEN3_1LT5 LASTDEN3_GTE5 LASTDEN3_Never AGE CHILDREN
WEIGHT2 FALL12MN/SELECTION=BACKWARD;
```

```
RUN;
```

SAS Code for OLS Regression Analysis

```
data OLSModel;
set THESIS.BECKYDATA;

*Create a role variable to split the dataset into train/
test sets;
u=uniform(123);
if u < 0.7 then role="TRAINING";
else role = "TESTING";

*Create dummy variables for the 35 categorical explanatory
variables;

*HLTHPLN1 Base category is HLTHPLN1=7 and 9;
if HLTHPLN1=1 then HLTHPLN1_Y=1; else HLTHPLN1_Y=0;
if HLTHPLN1=2 then HLTHPLN1_N=1; else HLTHPLN1_N=0;

*PERSDOC2 Base category is PERSDOC2=7 and 9;
if PERSDOC2=1 or PERSDOC2=2 then PERSDOC2_Y=1; else
PERSDOC2_Y=0;
if PERSDOC2=3 then PERSDOC2_N=1; else PERSDOC2_N=0;

*MEDCOST Base category is MEDCOST=7 and 9;
if MEDCOST=1 then MEDCOST_Y=1; else MEDCOST_Y=0;
if MEDCOST=2 then MEDCOST_N=1; else MEDCOST_N=0;

*HADMAM Base category is HADMAM=7 and 9;
if HADMAM=1 then HADMAM_Y=1; else HADMAM_Y=0;
if HADMAM=2 then HADMAM_N=1; else HADMAM_N=0;
if HADMAM=99 then HADMAM_NA=1; else HADMAM_NA=0;

*PROFEXAM Base category is PROFEXAM=7 and 9;
if PROFEXAM=1 then PROFEXAM_Y=1; else PROFEXAM_Y=0;
if PROFEXAM=2 then PROFEXAM_N=1; else PROFEXAM_N=0;
if PROFEXAM=99 then PROFEXAM_NA=1; else PROFEXAM_NA=0;

*HADPAP2 Base category is HADPAP2=7 and 9;
```

```
if HADPAP2=1 then HADPAP2_Y=1; else HADPAP2_Y=0;
if HADPAP2=2 then HADPAP2_N=1; else HADPAP2_N=0;
if HADPAP2=99 then HADPAP2_NA=1; else HADPAP2_NA=0;

*PSATEST1 Base category is PSATEST1=7 and 9;
if PSATEST1=1 then PSATEST1_Y=1; else PSATEST1_Y=0;
if PSATEST1=2 then PSATEST1_N=1; else PSATEST1_N=0;
if PSATEST1=99 then PSATEST1_NA=1; else PSATEST1_NA=0;

*HIVTST6 Base category is HIVTST6=7 and 9;
if HIVTST6=1 then HIVTST6_Y=1; else HIVTST6_Y=0;
if HIVTST6=2 then HIVTST6_N=1; else HIVTST6_N=0;

*_STATE Base category is _STATE=66 and 72;
if _STATE=1 then _STATE_Alabama=1; else _STATE_Alabama=0;
if _STATE=2 then _STATE_Alaska=1; else _STATE_Alaska=0;
if _STATE=4 then _STATE_Arizona=1; else _STATE_Arizona=0;
if _STATE=5 then _STATE_Arkansas=1; else _STATE_Arkansas=0;
if _STATE=6 then _STATE_California=1; else
_STATE_California=0;
if _STATE=8 then _STATE_Colorado=1; else _STATE_Colorado=0;
if _STATE=9 then _STATE_Connecticut=1; else
_STATE_Connecticut=0;
if _STATE=10 then _STATE_Delaware=1; else
_STATE_Delaware=0;
if _STATE=11 then _STATE_DC=1; else _STATE_DC=0;
if _STATE=12 then _STATE_Florida=1; else _STATE_Florida=0;
if _STATE=13 then _STATE_Georgia=1; else _STATE_Georgia=0;
if _STATE=15 then _STATE_Hawaii=1; else _STATE_Hawaii=0;
if _STATE=16 then _STATE_Idaho=1; else _STATE_Idaho=0;
if _STATE=17 then _STATE_Illinois=1; else
_STATE_Illinois=0;
if _STATE=18 then _STATE_Indiana=1; else _STATE_Indiana=0;
if _STATE=19 then _STATE_Iowa=1; else _STATE_Iowa=0;
if _STATE=20 then _STATE_Kansas=1; else _STATE_Kansas=0;
if _STATE=21 then _STATE_Kentucky=1; else
_STATE_Kentucky=0;
if _STATE=22 then _STATE_Louisiana=1; else
_STATE_Louisiana=0;
if _STATE=23 then _STATE_Maine=1; else _STATE_Maine=0;
```

```
if _STATE=24 then _STATE_Maryland=1; else
_STATE_Maryland=0;
if _STATE=25 then _STATE_Massachusetts=1; else
_STATE_Massachusetts=0;
if _STATE=26 then _STATE_Michigan=1; else
_STATE_Michigan=0;
if _STATE=27 then _STATE_Minnesota=1; else
_STATE_Minnesota=0;
if _STATE=28 then _STATE_Mississippi=1; else
_STATE_Mississippi=0;
if _STATE=29 then _STATE_Missouri=1; else
_STATE_Missouri=0;
if _STATE=30 then _STATE_Montana=1; else _STATE_Montana=0;
if _STATE=31 then _STATE_Nebraska=1; else
_STATE_Nebraska=0;
if _STATE=32 then _STATE_Nevada=1; else _STATE_Nevada=0;
if _STATE=33 then _STATE_NewHampshire=1; else
_STATE_NewHampshire=0;
if _STATE=34 then _STATE_NewJersey=1; else
_STATE_NewJersey=0;
if _STATE=35 then _STATE_NewMexico=1; else
_STATE_NewMexico=0;
if _STATE=36 then _STATE_NewYork=1; else _STATE_NewYork=0;
if _STATE=37 then _STATE_NorthCarolina=1; else
_STATE_NorthCarolina=0;
if _STATE=38 then _STATE_NorthDakota=1; else
_STATE_NorthDakota=0;
if _STATE=39 then _STATE_Ohio=1; else _STATE_Ohio=0;
if _STATE=40 then _STATE_Oklahoma=1; else
_STATE_Oklahoma=0;
if _STATE=41 then _STATE_Oregon=1; else _STATE_Oregon=0;
if _STATE=42 then _STATE_Pennsylvania=1; else
_STATE_Pennsylvania=0;
if _STATE=44 then _STATE_RhodeIsland=1; else
_STATE_RhodeIsland=0;
if _STATE=45 then _STATE_SouthCarolina=1; else
_STATE_SouthCarolina=0;
if _STATE=46 then _STATE_SouthDakota=1; else
_STATE_SouthDakota=0;
```

```
if _STATE=47 then _STATE_Tennesee=1; else
_STATE_Tennesee=0;
if _STATE=48 then _STATE_Texas=1; else _STATE_Texas=0;
if _STATE=49 then _STATE_Utah=1; else _STATE_Utah=0;
if _STATE=50 then _STATE_Vermont=1; else _STATE_Vermont=0;
if _STATE=51 then _STATE_Virginia=1; else
_STATE_Virginia=0;
if _STATE=53 then _STATE_Washington=1; else
_STATE_Washington=0;
if _STATE=54 then _STATE_WestVirginia=1; else
_STATE_WestVirginia=0;
if _STATE=55 then _STATE_Wisconsin=1; else
_STATE_Wisconsin=0;
if _STATE=56 then _STATE_Wyoming=1; else _STATE_Wyoming=0;

*VETERAN3 Base category is VETERAN3=7 and 9;
if VETERAN3=1 then VETERAN3_Y=1; else VETERAN3_Y=0;
if VETERAN3=2 then VETERAN3_N=1; else VETERAN3_N=0;

*MARITAL Base category is MARITAL=9;
if MARITAL=1 then MARITAL_MARRIED=1; else
MARITAL_MARRIED=0;
if MARITAL=2 then MARITAL_DIVORCED=1; else
MARITAL_DIVORCED=0;
if MARITAL=3 then MARITAL_WIDOWED=1; else
MARITAL_WIDOWED=0;
if MARITAL=4 then MARITAL_SEPARATED=1; else
MARITAL_SEPARATED=0;
if MARITAL=5 then MARITAL_NEVERMARRIED=1; else
MARITAL_NEVERMARRIED=0;
if MARITAL=6 then MARITAL_UNMARRIEDCOUPLE=1; else
MARITAL_UNMARRIEDCOUPLE=0;

*CVDSTRK3 Base category is CVDSTRK3=7 and 9;
if CVDSTRK3=1 then CVDSTRK3_Y=1; else CVDSTRK3_Y=0;
if CVDSTRK3=2 then CVDSTRK3_N=1; else CVDSTRK3_N=0;

*ASTHMA3 Base category is ASTHMA3=7 and 9;
if ASTHMA3=1 then ASTHMA3_Y=1; else ASTHMA3_Y=0;
if ASTHMA3=2 then ASTHMA3_N=1; else ASTHMA3_N=0;
```

```
*CHCSCNCR Base category is CHCSCNCR=7 and 9;
if CHCSCNCR=1 then CHCSCNCR_Y=1; else CHCSCNCR_Y=0;
if CHCSCNCR=2 then CHCSCNCR_N=1; else CHCSCNCR_N=0;

*CHCOCNCR Base category is CHCOCNCR=7 and 9;
if CHCOCNCR=1 then CHCOCNCR_Y=1; else CHCOCNCR_Y=0;
if CHCOCNCR=2 then CHCOCNCR_N=1; else CHCOCNCR_N=0;

*CHCCOPD1 Base category is CHCCOPD1=7 and 9;
if CHCCOPD1=1 then CHCCOPD1_Y=1; else CHCCOPD1_Y=0;
if CHCCOPD1=2 then CHCCOPD1_N=1; else CHCCOPD1_N=0;

*CHECKUP1 Base category is CHECKUP1=7, 8 and 9;
if CHECKUP1=1 then CHECKUP1_LT1=1; else CHECKUP1_LT1=0;
if CHECKUP1=2 or CHECKUP1=3 then CHECKUP1_1T5=1; else
CHECKUP1_1T5=0;
if CHECKUP1=4 then CHECKUP1_GT5=1; else CHECKUP1_GT5=0;

*EXERANY2 Base category is EXERANY2=7 and 9;
if EXERANY2=1 then EXERANY2_Y=1; else EXERANY2_Y=0;
if EXERANY2=2 then EXERANY2_N=1; else EXERANY2_N=0;

*CVDINFR4 Base category is CVDINFR4=7 and 9;
if CVDINFR4=1 then CVDINFR4_Y=1; else CVDINFR4_Y=0;
if CVDINFR4=2 then CVDINFR4_N=1; else CVDINFR4_N=0;

*CVDCRHD4 Base category is CVDCRHD4=7 and 9;
if CVDCRHD4=1 then CVDCRHD4_Y=1; else CVDCRHD4_Y=0;
if CVDCRHD4=2 then CVDCRHD4_N=1; else CVDCRHD4_N=0;

*EDUCA Base category is EDUCA=9;
if EDUCA=1 then EDUCA_None=1; else EDUCA_None=0;
if EDUCA=2 then EDUCA_Elem=1; else EDUCA_Elem=0;
if EDUCA=3 then EDUCA_SomeHS=1; else EDUCA_SomeHS=0;
if EDUCA=4 then EDUCA_GradHS=1; else EDUCA_GradHS=0;
if EDUCA=5 then EDUCA_SomeColl=1; else EDUCA_SomeColl=0;
if EDUCA=6 then EDUCA_GradColl=1; else EDUCA_GradColl=0;

*EMPLOY Base category is EMPLOY=9;
```

```
if EMPLOY=1 or EMPLOY=2 then EMPLOY_SelfOrWages=1; else
EMPLOY_SelfOrWages=0;
if EMPLOY=3 or EMPLOY=4 then EMPLOY_Unemployed=1; else
EMPLOY_Unemployed=0;
if EMPLOY=5 then EMPLOY_Homemaker=1; else
EMPLOY_Homemaker=0;
if EMPLOY=6 then EMPLOY_Student=1; else EMPLOY_Student=0;
if EMPLOY=7 then EMPLOY_Retired=1; else EMPLOY_Retired=0;
if EMPLOY=8 then EMPLOY_Unable=1; else EMPLOY_Unable=0;

*INCOME2 Base category is INCOME2=77 and 99;
if INCOME2=01 then INCOME2_LT10K=1; else INCOME2_LT10K=0;
if INCOME2=02 or INCOME2=03 or INCOME2=04 then
INCOME2_10KLT25K=1; else INCOME2_10KLT25K=0;
if INCOME2=05 or INCOME2=06 then INCOME2_25KLT50K=1; else
INCOME2_25KLT50K=0;
if INCOME2=07 or INCOME2=08 then INCOME2_50KPLUS=1; else
INCOME2_50KPLUS=0;

*PREGNANT Base category is PREGNANT=7 and 9;
if PREGNANT=1 then PREGNANT_Y=1; else PREGNANT_Y=0;
if PREGNANT=2 then PREGNANT_N=1; else PREGNANT_N=0;
if PREGNANT=99 then PREGNANT_NA=1; else PREGNANT_NA=0;

*QLACTLM2 Base category is QLACTLM2=7 and 9;
if QLACTLM2=1 then QLACTLM2_Y=1; else QLACTLM2_Y=0;
if QLACTLM2=2 then QLACTLM2_N=1; else QLACTLM2_N=0;

*USEEQUIP Base category is USEEQUIP=7 and 9;
if USEEQUIP=1 then USEEQUIP_Y=1; else USEEQUIP_Y=0;
if USEEQUIP=2 then USEEQUIP_N=1; else USEEQUIP_N=0;

*USENOW3 Base category is USENOW3=7 and 9;
if USENOW3=1 then USENOW3_Daily=1; else USENOW3_Daily=0;
if USENOW3=2 then USENOW3_SomeDays=1; else
USENOW3_SomeDays=0;
if USENOW3=3 then USENOW3_None=1; else USENOW3_None=0;

*FLUSHOT5 Base category is FLUSHOT5=7 and 9;
if FLUSHOT5=1 then FLUSHOT5_Y=1; else FLUSHOT5_Y=0;
```

```
if FLUSHOT5=2 then FLUSHOT5_N=1; else FLUSHOT5_N=0;

*PNEUVAC3 Base category is PNEUVAC3=7 and 9;
if PNEUVAC3=1 then PNEUVAC3_Y=1; else PNEUVAC3_Y=0;
if PNEUVAC3=2 then PNEUVAC3_N=1; else PNEUVAC3_N=0;

*SEATBELT Base category is SEATBELT=7 and 9;
if SEATBELT=1 then SEATBELT_Always=1; else
SEATBELT_Always=0;
if SEATBELT=2 or SEATBELT=3 then SEATBELT_Sometimes=1; else
SEATBELT_Sometimes=0;
if SEATBELT=4 or SEATBELT=5 then SEATBELT_SeldomNever=1;
else SEATBELT_SeldomNever=0;

*HAVARTH3 Base category is HAVARTH3=7 and 9;
if HAVARTH3=1 then HAVARTH3_Y=1; else HAVARTH3_Y=0;
if HAVARTH3=2 then HAVARTH3_N=1; else HAVARTH3_N=0;

*ADDEPEV2 Base category is ADDEPEV2=7 and 9;
if ADDEPEV2=1 then ADDEPEV2_Y=1; else ADDEPEV2_Y=0;
if ADDEPEV2=2 then ADDEPEV2_N=1; else ADDEPEV2_N=0;

*CHCKIDNY Base category is CHCKIDNY=7 and 9;
if CHCKIDNY=1 then CHCKIDNY_Y=1; else CHCKIDNY_Y=0;
if CHCKIDNY=2 then CHCKIDNY_N=1; else CHCKIDNY_N=0;

*CHCVISN1 Base category is CHCVISN1=7 and 9;
if CHCVISN1=1 or CHCVISN1=3 then CHCVISN1_Y=1; else
CHCVISN1_Y=0;
if CHCVISN1=2 then CHCVISN1_N=1; else CHCVISN1_N=0;

*DIABETE3 Base category is DIABETE3=7 and 9;
if DIABETE3=1 or DIABETE3=2 then DIABETE3_Y=1; else
DIABETE3_Y=0;
if DIABETE3=3 or DIABETE3=4 then DIABETE3_N=1; else
DIABETE3_N=0;

*LASTDEN3 Base category is LASTDEN3=7 and 9;
if LASTDEN3=1 then LASTDEN3_LT1=1; else LASTDEN3_LT1=0;
```

```
if LASTDEN3=2 or LASTDEN3=3 then LASTDEN3_1LT5=1; else
LASTDEN3_1LT5=0;
if LASTDEN3=4 then LASTDEN3_GTE5=1; else LASTDEN3_GTE5=0;
if LASTDEN3=8 then LASTDEN3_Never=1; else LASTDEN3_Never=0;

*Now, subset the training set to do the analysis;
data OLSTraining;
     set OLSModel;
     where role='TRAINING';
run;

title "OLS Model";
PROC REG Data=OLSTraining
PLOTS(MAXPOINTS=NONE)=DIAGNOSTICS;
     MODEL PHYSHLTH=HLTHPLN1_Y HLTHPLN1_N PERSDOC2_Y
PERSDOC2_N MEDCOST_Y MEDCOST_N
     HADMAM_Y HADMAM_N HADMAM_NA PROFEXAM_Y PROFEXAM_N
PROFEXAM_NA HADPAP2_Y HADPAP2_N HADPAP2_NA
     PSATEST1_Y PSATEST1_N PSATEST1_NA HIVTST6_Y HIVTST6_N
_STATE_Alabama _STATE_Alaska _STATE_Arizona
     _STATE_Arkansas _STATE_California _STATE_Colorado
_STATE_Connecticut _STATE_Delaware _STATE_DC
     _STATE_Florida _STATE_Georgia _STATE_Hawaii
_STATE_Idaho _STATE_Illinois _STATE_Indiana _STATE_Iowa
     _STATE_Kansas _STATE_Kentucky _STATE_Louisiana
_STATE_Maine _STATE_Maryland _STATE_Massachusetts
     _STATE_Michigan _STATE_Minnesota _STATE_Mississippi
_STATE_Missouri _STATE_Montana _STATE_Nebraska
     _STATE_Nevada _STATE_NewHampshire _STATE_NewJersey
_STATE_NewMexico _STATE_NewYork _STATE_NorthCarolina
     _STATE_NorthDakota _STATE_Ohio _STATE_Oklahoma
_STATE_Oregon _STATE_Pennsylvania _STATE_RhodeIsland
     _STATE_SouthCarolina _STATE_SouthDakota _STATE_Tennesee
_STATE_Texas _STATE_Utah _STATE_Vermont
     _STATE_Virginia _STATE_Washington _STATE_WestVirginia
_STATE_Wisconsin _STATE_Wyoming VETERAN3_Y
     VETERAN3_N MARITAL_MARRIED MARITAL_DIVORCED
MARITAL_WIDOWED MARITAL_SEPARATED MARITAL_NEVERMARRIED
     MARITAL_UNMARRIEDCOUPLE CVDSTRK3_Y CVDSTRK3_N ASTHMA3_Y
ASTHMA3_N CHCSCNCR_Y CHCSCNCR_N
```

```
     CHCOCNCR_Y CHCOCNCR_N CHCCOPD1_Y CHCCOPD1_N
CHECKUP1_LT1 CHECKUP1_1T5 CHECKUP1_GT5 EXERANY2_Y
     EXERANY2_N CVDINFR4_Y CVDINFR4_N CVDCRHD4_Y CVDCRHD4_N
EDUCA_None EDUCA_Elem EDUCA_SomeHS
     EDUCA_GradHS EDUCA_SomeColl EDUCA_GradColl
EMPLOY_SelfOrWages EMPLOY_Unemployed EMPLOY_Homemaker
     EMPLOY_Student EMPLOY_Retired EMPLOY_Unable
INCOME2_LT10K INCOME2_10KLT25K INCOME2_25KLT50K
     INCOME2_50KPLUS PREGNANT_Y PREGNANT_N PREGNANT_NA
QLACTLM2_Y QLACTLM2_N USEEQUIP_Y USEEQUIP_N
     USENOW3_Daily USENOW3_SomeDays USENOW3_None FLUSHOT5_Y
FLUSHOT5_N PNEUVAC3_Y PNEUVAC3_N
     SEATBELT_Always SEATBELT_Sometimes SEATBELT_SeldomNever
HAVARTH3_Y HAVARTH3_N ADDEPEV2_Y
     ADDEPEV2_N CHCKIDNY_Y CHCKIDNY_N CHCVISN1_Y CHCVISN1_N
DIABETE3_Y DIABETE3_N LASTDEN3_LT1
     LASTDEN3_1LT5 LASTDEN3_GTE5 LASTDEN3_Never AGE CHILDREN
WEIGHT2 FALL12MN/vif collin SELECTION=BACKWARD;
RUN;
```

SAS Code for Poisson Regression Analysis

```
data PoissModel;
set THESIS.BECKYDATA;

*Create a role variable to split the dataset into train/
test sets;
u=uniform(123);
if u < 0.7 then role="TRAINING";
else role = "TESTING";

*Create dummy variables for the 35 categorical explanatory
variables;

*HLTHPLN1 Base category is HLTHPLN1=7 and 9;
if HLTHPLN1=1 then HLTHPLN1_Y=1; else HLTHPLN1_Y=0;
if HLTHPLN1=2 then HLTHPLN1_N=1; else HLTHPLN1_N=0;

*PERSDOC2 Base category is PERSDOC2=7 and 9;
```

```
if PERSDOC2=1 or PERSDOC2=2 then PERSDOC2_Y=1; else
PERSDOC2_Y=0;
if PERSDOC2=3 then PERSDOC2_N=1; else PERSDOC2_N=0;

*MEDCOST Base category is MEDCOST=7 and 9;
if MEDCOST=1 then MEDCOST_Y=1; else MEDCOST_Y=0;
if MEDCOST=2 then MEDCOST_N=1; else MEDCOST_N=0;

*HADMAM Base category is HADMAM=7 and 9;
if HADMAM=1 then HADMAM_Y=1; else HADMAM_Y=0;
if HADMAM=2 then HADMAM_N=1; else HADMAM_N=0;
if HADMAM=99 then HADMAM_NA=1; else HADMAM_NA=0;

*PROFEXAM Base category is PROFEXAM=7 and 9;
if PROFEXAM=1 then PROFEXAM_Y=1; else PROFEXAM_Y=0;
if PROFEXAM=2 then PROFEXAM_N=1; else PROFEXAM_N=0;
if PROFEXAM=99 then PROFEXAM_NA=1; else PROFEXAM_NA=0;

*HADPAP2 Base category is HADPAP2=7 and 9;
if HADPAP2=1 then HADPAP2_Y=1; else HADPAP2_Y=0;
if HADPAP2=2 then HADPAP2_N=1; else HADPAP2_N=0;
if HADPAP2=99 then HADPAP2_NA=1; else HADPAP2_NA=0;

*PSATEST1 Base category is PSATEST1=7 and 9;
if PSATEST1=1 then PSATEST1_Y=1; else PSATEST1_Y=0;
if PSATEST1=2 then PSATEST1_N=1; else PSATEST1_N=0;
if PSATEST1=99 then PSATEST1_NA=1; else PSATEST1_NA=0;

*HIVTST6 Base category is HIVTST6=7 and 9;
if HIVTST6=1 then HIVTST6_Y=1; else HIVTST6_Y=0;
if HIVTST6=2 then HIVTST6_N=1; else HIVTST6_N=0;

*_STATE Base category is _STATE=66 and 72;
if _STATE=1 then _STATE_Alabama=1; else _STATE_Alabama=0;
if _STATE=2 then _STATE_Alaska=1; else _STATE_Alaska=0;
if _STATE=4 then _STATE_Arizona=1; else _STATE_Arizona=0;
if _STATE=5 then _STATE_Arkansas=1; else _STATE_Arkansas=0;
if _STATE=6 then _STATE_California=1; else
_STATE_California=0;
if _STATE=8 then _STATE_Colorado=1; else _STATE_Colorado=0;
```

```
if _STATE=9 then _STATE_Connecticut=1; else
_STATE_Connecticut=0;
if _STATE=10 then _STATE_Delaware=1; else
_STATE_Delaware=0;
if _STATE=11 then _STATE_DC=1; else _STATE_DC=0;
if _STATE=12 then _STATE_Florida=1; else _STATE_Florida=0;
if _STATE=13 then _STATE_Georgia=1; else _STATE_Georgia=0;
if _STATE=15 then _STATE_Hawaii=1; else _STATE_Hawaii=0;
if _STATE=16 then _STATE_Idaho=1; else _STATE_Idaho=0;
if _STATE=17 then _STATE_Illinois=1; else
_STATE_Illinois=0;
if _STATE=18 then _STATE_Indiana=1; else _STATE_Indiana=0;
if _STATE=19 then _STATE_Iowa=1; else _STATE_Iowa=0;
if _STATE=20 then _STATE_Kansas=1; else _STATE_Kansas=0;
if _STATE=21 then _STATE_Kentucky=1; else
_STATE_Kentucky=0;
if _STATE=22 then _STATE_Louisiana=1; else
_STATE_Louisiana=0;
if _STATE=23 then _STATE_Maine=1; else _STATE_Maine=0;
if _STATE=24 then _STATE_Maryland=1; else
_STATE_Maryland=0;
if _STATE=25 then _STATE_Massachusetts=1; else
_STATE_Massachusetts=0;
if _STATE=26 then _STATE_Michigan=1; else
_STATE_Michigan=0;
if _STATE=27 then _STATE_Minnesota=1; else
_STATE_Minnesota=0;
if _STATE=28 then _STATE_Mississippi=1; else
_STATE_Mississippi=0;
if _STATE=29 then _STATE_Missouri=1; else
_STATE_Missouri=0;
if _STATE=30 then _STATE_Montana=1; else _STATE_Montana=0;
if _STATE=31 then _STATE_Nebraska=1; else
_STATE_Nebraska=0;
if _STATE=32 then _STATE_Nevada=1; else _STATE_Nevada=0;
if _STATE=33 then _STATE_NewHampshire=1; else
_STATE_NewHampshire=0;
if _STATE=34 then _STATE_NewJersey=1; else
_STATE_NewJersey=0;
```

```
if _STATE=35 then _STATE_NewMexico=1; else
_STATE_NewMexico=0;
if _STATE=36 then _STATE_NewYork=1; else _STATE_NewYork=0;
if _STATE=37 then _STATE_NorthCarolina=1; else
_STATE_NorthCarolina=0;
if _STATE=38 then _STATE_NorthDakota=1; else
_STATE_NorthDakota=0;
if _STATE=39 then _STATE_Ohio=1; else _STATE_Ohio=0;
if _STATE=40 then _STATE_Oklahoma=1; else
_STATE_Oklahoma=0;
if _STATE=41 then _STATE_Oregon=1; else _STATE_Oregon=0;
if _STATE=42 then _STATE_Pennsylvania=1; else
_STATE_Pennsylvania=0;
if _STATE=44 then _STATE_RhodeIsland=1; else
_STATE_RhodeIsland=0;
if _STATE=45 then _STATE_SouthCarolina=1; else
_STATE_SouthCarolina=0;
if _STATE=46 then _STATE_SouthDakota=1; else
_STATE_SouthDakota=0;
if _STATE=47 then _STATE_Tennesee=1; else
_STATE_Tennesee=0;
if _STATE=48 then _STATE_Texas=1; else _STATE_Texas=0;
if _STATE=49 then _STATE_Utah=1; else _STATE_Utah=0;
if _STATE=50 then _STATE_Vermont=1; else _STATE_Vermont=0;
if _STATE=51 then _STATE_Virginia=1; else
_STATE_Virginia=0;
if _STATE=53 then _STATE_Washington=1; else
_STATE_Washington=0;
if _STATE=54 then _STATE_WestVirginia=1; else
_STATE_WestVirginia=0;
if _STATE=55 then _STATE_Wisconsin=1; else
_STATE_Wisconsin=0;
if _STATE=56 then _STATE_Wyoming=1; else _STATE_Wyoming=0;

*VETERAN3 Base category is VETERAN3=7 and 9;
if VETERAN3=1 then VETERAN3_Y=1; else VETERAN3_Y=0;
if VETERAN3=2 then VETERAN3_N=1; else VETERAN3_N=0;

*MARITAL Base category is MARITAL=9;
```

```
if MARITAL=1 then MARITAL_MARRIED=1; else
MARITAL_MARRIED=0;
if MARITAL=2 then MARITAL_DIVORCED=1; else
MARITAL_DIVORCED=0;
if MARITAL=3 then MARITAL_WIDOWED=1; else
MARITAL_WIDOWED=0;
if MARITAL=4 then MARITAL_SEPARATED=1; else
MARITAL_SEPARATED=0;
if MARITAL=5 then MARITAL_NEVERMARRIED=1; else
MARITAL_NEVERMARRIED=0;
if MARITAL=6 then MARITAL_UNMARRIEDCOUPLE=1; else
MARITAL_UNMARRIEDCOUPLE=0;

*CVDSTRK3 Base category is CVDSTRK3=7 and 9;
if CVDSTRK3=1 then CVDSTRK3_Y=1; else CVDSTRK3_Y=0;
if CVDSTRK3=2 then CVDSTRK3_N=1; else CVDSTRK3_N=0;

*ASTHMA3 Base category is ASTHMA3=7 and 9;
if ASTHMA3=1 then ASTHMA3_Y=1; else ASTHMA3_Y=0;
if ASTHMA3=2 then ASTHMA3_N=1; else ASTHMA3_N=0;

*CHCSCNCR Base category is CHCSCNCR=7 and 9;
if CHCSCNCR=1 then CHCSCNCR_Y=1; else CHCSCNCR_Y=0;
if CHCSCNCR=2 then CHCSCNCR_N=1; else CHCSCNCR_N=0;

*CHCOCNCR Base category is CHCOCNCR=7 and 9;
if CHCOCNCR=1 then CHCOCNCR_Y=1; else CHCOCNCR_Y=0;
if CHCOCNCR=2 then CHCOCNCR_N=1; else CHCOCNCR_N=0;

*CHCCOPD1 Base category is CHCCOPD1=7 and 9;
if CHCCOPD1=1 then CHCCOPD1_Y=1; else CHCCOPD1_Y=0;
if CHCCOPD1=2 then CHCCOPD1_N=1; else CHCCOPD1_N=0;

*CHECKUP1 Base category is CHECKUP1=7, 8 and 9;
if CHECKUP1=1 then CHECKUP1_LT1=1; else CHECKUP1_LT1=0;
if CHECKUP1=2 or CHECKUP1=3 then CHECKUP1_1T5=1; else
CHECKUP1_1T5=0;
if CHECKUP1=4 then CHECKUP1_GT5=1; else CHECKUP1_GT5=0;

*EXERANY2 Base category is EXERANY2=7 and 9;
```

```
if EXERANY2=1 then EXERANY2_Y=1; else EXERANY2_Y=0;
if EXERANY2=2 then EXERANY2_N=1; else EXERANY2_N=0;

*CVDINFR4 Base category is CVDINFR4=7 and 9;
if CVDINFR4=1 then CVDINFR4_Y=1; else CVDINFR4_Y=0;
if CVDINFR4=2 then CVDINFR4_N=1; else CVDINFR4_N=0;

*CVDCRHD4 Base category is CVDCRHD4=7 and 9;
if CVDCRHD4=1 then CVDCRHD4_Y=1; else CVDCRHD4_Y=0;
if CVDCRHD4=2 then CVDCRHD4_N=1; else CVDCRHD4_N=0;

*EDUCA Base category is EDUCA=9;
if EDUCA=1 then EDUCA_None=1; else EDUCA_None=0;
if EDUCA=2 then EDUCA_Elem=1; else EDUCA_Elem=0;
if EDUCA=3 then EDUCA_SomeHS=1; else EDUCA_SomeHS=0;
if EDUCA=4 then EDUCA_GradHS=1; else EDUCA_GradHS=0;
if EDUCA=5 then EDUCA_SomeColl=1; else EDUCA_SomeColl=0;
if EDUCA=6 then EDUCA_GradColl=1; else EDUCA_GradColl=0;

*EMPLOY Base category is EMPLOY=9;
if EMPLOY=1 or EMPLOY=2 then EMPLOY_SelfOrWages=1; else
EMPLOY_SelfOrWages=0;
if EMPLOY=3 or EMPLOY=4 then EMPLOY_Unemployed=1; else
EMPLOY_Unemployed=0;
if EMPLOY=5 then EMPLOY_Homemaker=1; else
EMPLOY_Homemaker=0;
if EMPLOY=6 then EMPLOY_Student=1; else EMPLOY_Student=0;
if EMPLOY=7 then EMPLOY_Retired=1; else EMPLOY_Retired=0;
if EMPLOY=8 then EMPLOY_Unable=1; else EMPLOY_Unable=0;

*INCOME2 Base category is INCOME2=77 and 99;
if INCOME2=01 then INCOME2_LT10K=1; else INCOME2_LT10K=0;
if INCOME2=02 or INCOME2=03 or INCOME2=04 then
INCOME2_10KLT25K=1; else INCOME2_10KLT25K=0;
if INCOME2=05 or INCOME2=06 then INCOME2_25KLT50K=1; else
INCOME2_25KLT50K=0;
if INCOME2=07 or INCOME2=08 then INCOME2_50KPLUS=1; else
INCOME2_50KPLUS=0;

*PREGNANT Base category is PREGNANT=7 and 9;
```

```
if PREGNANT=1 then PREGNANT_Y=1; else PREGNANT_Y=0;
if PREGNANT=2 then PREGNANT_N=1; else PREGNANT_N=0;
if PREGNANT=99 then PREGNANT_NA=1; else PREGNANT_NA=0;

*QLACTLM2 Base category is QLACTLM2=7 and 9;
if QLACTLM2=1 then QLACTLM2_Y=1; else QLACTLM2_Y=0;
if QLACTLM2=2 then QLACTLM2_N=1; else QLACTLM2_N=0;

*USEEQUIP Base category is USEEQUIP=7 and 9;
if USEEQUIP=1 then USEEQUIP_Y=1; else USEEQUIP_Y=0;
if USEEQUIP=2 then USEEQUIP_N=1; else USEEQUIP_N=0;

*USENOW3 Base category is USENOW3=7 and 9;
if USENOW3=1 then USENOW3_Daily=1; else USENOW3_Daily=0;
if USENOW3=2 then USENOW3_SomeDays=1; else
USENOW3_SomeDays=0;
if USENOW3=3 then USENOW3_None=1; else USENOW3_None=0;

*FLUSHOT5 Base category is FLUSHOT5=7 and 9;
if FLUSHOT5=1 then FLUSHOT5_Y=1; else FLUSHOT5_Y=0;
if FLUSHOT5=2 then FLUSHOT5_N=1; else FLUSHOT5_N=0;

*PNEUVAC3 Base category is PNEUVAC3=7 and 9;
if PNEUVAC3=1 then PNEUVAC3_Y=1; else PNEUVAC3_Y=0;
if PNEUVAC3=2 then PNEUVAC3_N=1; else PNEUVAC3_N=0;

*SEATBELT Base category is SEATBELT=7 and 9;
if SEATBELT=1 then SEATBELT_Always=1; else
SEATBELT_Always=0;
if SEATBELT=2 or SEATBELT=3 then SEATBELT_Sometimes=1; else
SEATBELT_Sometimes=0;
if SEATBELT=4 or SEATBELT=5 then SEATBELT_SeldomNever=1;
else SEATBELT_SeldomNever=0;

*HAVARTH3 Base category is HAVARTH3=7 and 9;
if HAVARTH3=1 then HAVARTH3_Y=1; else HAVARTH3_Y=0;
if HAVARTH3=2 then HAVARTH3_N=1; else HAVARTH3_N=0;

*ADDEPEV2 Base category is ADDEPEV2=7 and 9;
if ADDEPEV2=1 then ADDEPEV2_Y=1; else ADDEPEV2_Y=0;
```

```
if ADDEPEV2=2 then ADDEPEV2_N=1; else ADDEPEV2_N=0;

*CHCKIDNY Base category is CHCKIDNY=7 and 9;
if CHCKIDNY=1 then CHCKIDNY_Y=1; else CHCKIDNY_Y=0;
if CHCKIDNY=2 then CHCKIDNY_N=1; else CHCKIDNY_N=0;

*CHCVISN1 Base category is CHCVISN1=7 and 9;
if CHCVISN1=1 or CHCVISN1=3 then CHCVISN1_Y=1; else
CHCVISN1_Y=0;
if CHCVISN1=2 then CHCVISN1_N=1; else CHCVISN1_N=0;

*DIABETE3 Base category is DIABETE3=7 and 9;
if DIABETE3=1 or DIABETE3=2 then DIABETE3_Y=1; else
DIABETE3_Y=0;
if DIABETE3=3 or DIABETE3=4 then DIABETE3_N=1; else
DIABETE3_N=0;

*LASTDEN3 Base category is LASTDEN3=7 and 9;
if LASTDEN3=1 then LASTDEN3_LT1=1; else LASTDEN3_LT1=0;
if LASTDEN3=2 or LASTDEN3=3 then LASTDEN3_1LT5=1; else
LASTDEN3_1LT5=0;
if LASTDEN3=4 then LASTDEN3_GTE5=1; else LASTDEN3_GTE5=0;
if LASTDEN3=8 then LASTDEN3_Never=1; else LASTDEN3_Never=0;

*Now, subset the training set to do the analysis;
data PoissTraining;
     set PoissModel;
     where role='TRAINING';
run;

title "Poisson Model";
PROC GENMOD Data=PoissTraining;
     MODEL PHYSHLTH=HLTHPLN1_Y HLTHPLN1_N PERSDOC2_Y
PERSDOC2_N MEDCOST_Y MEDCOST_N
     HADMAM_Y HADMAM_N HADMAM_NA PROFEXAM_Y PROFEXAM_N
PROFEXAM_NA HADPAP2_Y HADPAP2_N HADPAP2_NA
     PSATEST1_Y PSATEST1_N PSATEST1_NA HIVTST6_Y HIVTST6_N
_STATE_Alabama _STATE_Alaska _STATE_Arizona
     _STATE_Arkansas _STATE_California _STATE_Colorado
_STATE_Connecticut _STATE_Delaware _STATE_DC
```

```
     _STATE_Florida _STATE_Georgia _STATE_Hawaii
_STATE_Idaho _STATE_Illinois _STATE_Indiana _STATE_Iowa
     _STATE_Kansas _STATE_Kentucky _STATE_Louisiana
_STATE_Maine _STATE_Maryland _STATE_Massachusetts
     _STATE_Michigan _STATE_Minnesota _STATE_Mississippi
_STATE_Missouri _STATE_Montana _STATE_Nebraska
     _STATE_Nevada _STATE_NewHampshire _STATE_NewJersey
_STATE_NewMexico _STATE_NewYork _STATE_NorthCarolina
     _STATE_NorthDakota _STATE_Ohio _STATE_Oklahoma
_STATE_Oregon _STATE_Pennsylvania _STATE_RhodeIsland
     _STATE_SouthCarolina _STATE_SouthDakota _STATE_Tennesee
_STATE_Texas _STATE_Utah _STATE_Vermont
     _STATE_Virginia _STATE_Washington _STATE_WestVirginia
_STATE_Wisconsin _STATE_Wyoming VETERAN3_Y
     VETERAN3_N MARITAL_MARRIED MARITAL_DIVORCED
MARITAL_WIDOWED MARITAL_SEPARATED MARITAL_NEVERMARRIED
     MARITAL_UNMARRIEDCOUPLE CVDSTRK3_Y CVDSTRK3_N ASTHMA3_Y
ASTHMA3_N CHCSCNCR_Y CHCSCNCR_N
     CHCOCNCR_Y CHCOCNCR_N CHCCOPD1_Y CHCCOPD1_N
CHECKUP1_LT1 CHECKUP1_1T5 CHECKUP1_GT5 EXERANY2_Y
     EXERANY2_N CVDINFR4_Y CVDINFR4_N CVDCRHD4_Y CVDCRHD4_N
EDUCA_None EDUCA_Elem EDUCA_SomeHS
     EDUCA_GradHS EDUCA_SomeColl EDUCA_GradColl
EMPLOY_SelfOrWages EMPLOY_Unemployed EMPLOY_Homemaker
     EMPLOY_Student EMPLOY_Retired EMPLOY_Unable
INCOME2_LT10K INCOME2_10KLT25K INCOME2_25KLT50K
     INCOME2_50KPLUS PREGNANT_Y PREGNANT_N PREGNANT_NA
QLACTLM2_Y QLACTLM2_N USEEQUIP_Y USEEQUIP_N
     USENOW3_Daily USENOW3_SomeDays USENOW3_None FLUSHOT5_Y
FLUSHOT5_N PNEUVAC3_Y PNEUVAC3_N
     SEATBELT_Always SEATBELT_Sometimes SEATBELT_SeldomNever
HAVARTH3_Y HAVARTH3_N ADDEPEV2_Y
     ADDEPEV2_N CHCKIDNY_Y CHCKIDNY_N CHCVISN1_Y CHCVISN1_N
DIABETE3_Y DIABETE3_N LASTDEN3_LT1
     LASTDEN3_1LT5 LASTDEN3_GTE5 LASTDEN3_Never AGE CHILDREN
WEIGHT2 FALL12MN/dist=poisson SELECTION=BACKWARD;
RUN;
```

SAS Code for Negative Binomial Regression Analysis

```
data NBModel;
set THESIS.BECKYDATA;

*Create a role variable to split the dataset into train/
test sets;
u=uniform(123);
if u < 0.7 then role="TRAINING";
else role = "TESTING";

*Create dummy variables for the 35 categorical explanatory
variables;

*HLTHPLN1 Base category is HLTHPLN1=7 and 9;
if HLTHPLN1=1 then HLTHPLN1_Y=1; else HLTHPLN1_Y=0;
if HLTHPLN1=2 then HLTHPLN1_N=1; else HLTHPLN1_N=0;

*PERSDOC2 Base category is PERSDOC2=7 and 9;
if PERSDOC2=1 or PERSDOC2=2 then PERSDOC2_Y=1; else
PERSDOC2_Y=0;
if PERSDOC2=3 then PERSDOC2_N=1; else PERSDOC2_N=0;

*MEDCOST Base category is MEDCOST=7 and 9;
if MEDCOST=1 then MEDCOST_Y=1; else MEDCOST_Y=0;
if MEDCOST=2 then MEDCOST_N=1; else MEDCOST_N=0;

*HADMAM Base category is HADMAM=7 and 9;
if HADMAM=1 then HADMAM_Y=1; else HADMAM_Y=0;
if HADMAM=2 then HADMAM_N=1; else HADMAM_N=0;
if HADMAM=99 then HADMAM_NA=1; else HADMAM_NA=0;

*PROFEXAM Base category is PROFEXAM=7 and 9;
if PROFEXAM=1 then PROFEXAM_Y=1; else PROFEXAM_Y=0;
if PROFEXAM=2 then PROFEXAM_N=1; else PROFEXAM_N=0;
if PROFEXAM=99 then PROFEXAM_NA=1; else PROFEXAM_NA=0;

*HADPAP2 Base category is HADPAP2=7 and 9;
if HADPAP2=1 then HADPAP2_Y=1; else HADPAP2_Y=0;
if HADPAP2=2 then HADPAP2_N=1; else HADPAP2_N=0;
```

```
if HADPAP2=99 then HADPAP2_NA=1; else HADPAP2_NA=0;

*PSATEST1 Base category is PSATEST1=7 and 9;
if PSATEST1=1 then PSATEST1_Y=1; else PSATEST1_Y=0;
if PSATEST1=2 then PSATEST1_N=1; else PSATEST1_N=0;
if PSATEST1=99 then PSATEST1_NA=1; else PSATEST1_NA=0;

*HIVTST6 Base category is HIVTST6=7 and 9;
if HIVTST6=1 then HIVTST6_Y=1; else HIVTST6_Y=0;
if HIVTST6=2 then HIVTST6_N=1; else HIVTST6_N=0;

*_STATE Base category is _STATE=66 and 72;
if _STATE=1 then _STATE_Alabama=1; else _STATE_Alabama=0;
if _STATE=2 then _STATE_Alaska=1; else _STATE_Alaska=0;
if _STATE=4 then _STATE_Arizona=1; else _STATE_Arizona=0;
if _STATE=5 then _STATE_Arkansas=1; else _STATE_Arkansas=0;
if _STATE=6 then _STATE_California=1; else
_STATE_California=0;
if _STATE=8 then _STATE_Colorado=1; else _STATE_Colorado=0;
if _STATE=9 then _STATE_Connecticut=1; else
_STATE_Connecticut=0;
if _STATE=10 then _STATE_Delaware=1; else
_STATE_Delaware=0;
if _STATE=11 then _STATE_DC=1; else _STATE_DC=0;
if _STATE=12 then _STATE_Florida=1; else _STATE_Florida=0;
if _STATE=13 then _STATE_Georgia=1; else _STATE_Georgia=0;
if _STATE=15 then _STATE_Hawaii=1; else _STATE_Hawaii=0;
if _STATE=16 then _STATE_Idaho=1; else _STATE_Idaho=0;
if _STATE=17 then _STATE_Illinois=1; else
_STATE_Illinois=0;
if _STATE=18 then _STATE_Indiana=1; else _STATE_Indiana=0;
if _STATE=19 then _STATE_Iowa=1; else _STATE_Iowa=0;
if _STATE=20 then _STATE_Kansas=1; else _STATE_Kansas=0;
if _STATE=21 then _STATE_Kentucky=1; else
_STATE_Kentucky=0;
if _STATE=22 then _STATE_Louisiana=1; else
_STATE_Louisiana=0;
if _STATE=23 then _STATE_Maine=1; else _STATE_Maine=0;
if _STATE=24 then _STATE_Maryland=1; else
_STATE_Maryland=0;
```

```
if _STATE=25 then _STATE_Massachusetts=1; else
_STATE_Massachusetts=0;
if _STATE=26 then _STATE_Michigan=1; else
_STATE_Michigan=0;
if _STATE=27 then _STATE_Minnesota=1; else
_STATE_Minnesota=0;
if _STATE=28 then _STATE_Mississippi=1; else
_STATE_Mississippi=0;
if _STATE=29 then _STATE_Missouri=1; else
_STATE_Missouri=0;
if _STATE=30 then _STATE_Montana=1; else _STATE_Montana=0;
if _STATE=31 then _STATE_Nebraska=1; else
_STATE_Nebraska=0;
if _STATE=32 then _STATE_Nevada=1; else _STATE_Nevada=0;
if _STATE=33 then _STATE_NewHampshire=1; else
_STATE_NewHampshire=0;
if _STATE=34 then _STATE_NewJersey=1; else
_STATE_NewJersey=0;
if _STATE=35 then _STATE_NewMexico=1; else
_STATE_NewMexico=0;
if _STATE=36 then _STATE_NewYork=1; else _STATE_NewYork=0;
if _STATE=37 then _STATE_NorthCarolina=1; else
_STATE_NorthCarolina=0;
if _STATE=38 then _STATE_NorthDakota=1; else
_STATE_NorthDakota=0;
if _STATE=39 then _STATE_Ohio=1; else _STATE_Ohio=0;
if _STATE=40 then _STATE_Oklahoma=1; else
_STATE_Oklahoma=0;
if _STATE=41 then _STATE_Oregon=1; else _STATE_Oregon=0;
if _STATE=42 then _STATE_Pennsylvania=1; else
_STATE_Pennsylvania=0;
if _STATE=44 then _STATE_RhodeIsland=1; else
_STATE_RhodeIsland=0;
if _STATE=45 then _STATE_SouthCarolina=1; else
_STATE_SouthCarolina=0;
if _STATE=46 then _STATE_SouthDakota=1; else
_STATE_SouthDakota=0;
if _STATE=47 then _STATE_Tennesee=1; else
_STATE_Tennesee=0;
if _STATE=48 then _STATE_Texas=1; else _STATE_Texas=0;
```

```
if _STATE=49 then _STATE_Utah=1; else _STATE_Utah=0;
if _STATE=50 then _STATE_Vermont=1; else _STATE_Vermont=0;
if _STATE=51 then _STATE_Virginia=1; else
_STATE_Virginia=0;
if _STATE=53 then _STATE_Washington=1; else
_STATE_Washington=0;
if _STATE=54 then _STATE_WestVirginia=1; else
_STATE_WestVirginia=0;
if _STATE=55 then _STATE_Wisconsin=1; else
_STATE_Wisconsin=0;
if _STATE=56 then _STATE_Wyoming=1; else _STATE_Wyoming=0;

*VETERAN3 Base category is VETERAN3=7 and 9;
if VETERAN3=1 then VETERAN3_Y=1; else VETERAN3_Y=0;
if VETERAN3=2 then VETERAN3_N=1; else VETERAN3_N=0;

*MARITAL Base category is MARITAL=9;
if MARITAL=1 then MARITAL_MARRIED=1; else
MARITAL_MARRIED=0;
if MARITAL=2 then MARITAL_DIVORCED=1; else
MARITAL_DIVORCED=0;
if MARITAL=3 then MARITAL_WIDOWED=1; else
MARITAL_WIDOWED=0;
if MARITAL=4 then MARITAL_SEPARATED=1; else
MARITAL_SEPARATED=0;
if MARITAL=5 then MARITAL_NEVERMARRIED=1; else
MARITAL_NEVERMARRIED=0;
if MARITAL=6 then MARITAL_UNMARRIEDCOUPLE=1; else
MARITAL_UNMARRIEDCOUPLE=0;

*CVDSTRK3 Base category is CVDSTRK3=7 and 9;
if CVDSTRK3=1 then CVDSTRK3_Y=1; else CVDSTRK3_Y=0;
if CVDSTRK3=2 then CVDSTRK3_N=1; else CVDSTRK3_N=0;

*ASTHMA3 Base category is ASTHMA3=7 and 9;
if ASTHMA3=1 then ASTHMA3_Y=1; else ASTHMA3_Y=0;
if ASTHMA3=2 then ASTHMA3_N=1; else ASTHMA3_N=0;

*CHCSCNCR Base category is CHCSCNCR=7 and 9;
if CHCSCNCR=1 then CHCSCNCR_Y=1; else CHCSCNCR_Y=0;
```

```
if CHCSCNCR=2 then CHCSCNCR_N=1; else CHCSCNCR_N=0;

*CHCOCNCR Base category is CHCOCNCR=7 and 9;
if CHCOCNCR=1 then CHCOCNCR_Y=1; else CHCOCNCR_Y=0;
if CHCOCNCR=2 then CHCOCNCR_N=1; else CHCOCNCR_N=0;

*CHCCOPD1 Base category is CHCCOPD1=7 and 9;
if CHCCOPD1=1 then CHCCOPD1_Y=1; else CHCCOPD1_Y=0;
if CHCCOPD1=2 then CHCCOPD1_N=1; else CHCCOPD1_N=0;

*CHECKUP1 Base category is CHECKUP1=7, 8 and 9;
if CHECKUP1=1 then CHECKUP1_LT1=1; else CHECKUP1_LT1=0;
if CHECKUP1=2 or CHECKUP1=3 then CHECKUP1_1T5=1; else
CHECKUP1_1T5=0;
if CHECKUP1=4 then CHECKUP1_GT5=1; else CHECKUP1_GT5=0;

*EXERANY2 Base category is EXERANY2=7 and 9;
if EXERANY2=1 then EXERANY2_Y=1; else EXERANY2_Y=0;
if EXERANY2=2 then EXERANY2_N=1; else EXERANY2_N=0;

*CVDINFR4 Base category is CVDINFR4=7 and 9;
if CVDINFR4=1 then CVDINFR4_Y=1; else CVDINFR4_Y=0;
if CVDINFR4=2 then CVDINFR4_N=1; else CVDINFR4_N=0;

*CVDCRHD4 Base category is CVDCRHD4=7 and 9;
if CVDCRHD4=1 then CVDCRHD4_Y=1; else CVDCRHD4_Y=0;
if CVDCRHD4=2 then CVDCRHD4_N=1; else CVDCRHD4_N=0;

*EDUCA Base category is EDUCA=9;
if EDUCA=1 then EDUCA_None=1; else EDUCA_None=0;
if EDUCA=2 then EDUCA_Elem=1; else EDUCA_Elem=0;
if EDUCA=3 then EDUCA_SomeHS=1; else EDUCA_SomeHS=0;
if EDUCA=4 then EDUCA_GradHS=1; else EDUCA_GradHS=0;
if EDUCA=5 then EDUCA_SomeColl=1; else EDUCA_SomeColl=0;
if EDUCA=6 then EDUCA_GradColl=1; else EDUCA_GradColl=0;

*EMPLOY Base category is EMPLOY=9;
if EMPLOY=1 or EMPLOY=2 then EMPLOY_SelfOrWages=1; else
EMPLOY_SelfOrWages=0;
```

```
if EMPLOY=3 or EMPLOY=4 then EMPLOY_Unemployed=1; else
EMPLOY_Unemployed=0;
if EMPLOY=5 then EMPLOY_Homemaker=1; else
EMPLOY_Homemaker=0;
if EMPLOY=6 then EMPLOY_Student=1; else EMPLOY_Student=0;
if EMPLOY=7 then EMPLOY_Retired=1; else EMPLOY_Retired=0;
if EMPLOY=8 then EMPLOY_Unable=1; else EMPLOY_Unable=0;

*INCOME2 Base category is INCOME2=77 and 99;
if INCOME2=01 then INCOME2_LT10K=1; else INCOME2_LT10K=0;
if INCOME2=02 or INCOME2=03 or INCOME2=04 then
INCOME2_10KLT25K=1; else INCOME2_10KLT25K=0;
if INCOME2=05 or INCOME2=06 then INCOME2_25KLT50K=1; else
INCOME2_25KLT50K=0;
if INCOME2=07 or INCOME2=08 then INCOME2_50KPLUS=1; else
INCOME2_50KPLUS=0;

*PREGNANT Base category is PREGNANT=7 and 9;
if PREGNANT=1 then PREGNANT_Y=1; else PREGNANT_Y=0;
if PREGNANT=2 then PREGNANT_N=1; else PREGNANT_N=0;
if PREGNANT=99 then PREGNANT_NA=1; else PREGNANT_NA=0;

*QLACTLM2 Base category is QLACTLM2=7 and 9;
if QLACTLM2=1 then QLACTLM2_Y=1; else QLACTLM2_Y=0;
if QLACTLM2=2 then QLACTLM2_N=1; else QLACTLM2_N=0;

*USEEQUIP Base category is USEEQUIP=7 and 9;
if USEEQUIP=1 then USEEQUIP_Y=1; else USEEQUIP_Y=0;
if USEEQUIP=2 then USEEQUIP_N=1; else USEEQUIP_N=0;

*USENOW3 Base category is USENOW3=7 and 9;
if USENOW3=1 then USENOW3_Daily=1; else USENOW3_Daily=0;
if USENOW3=2 then USENOW3_SomeDays=1; else
USENOW3_SomeDays=0;
if USENOW3=3 then USENOW3_None=1; else USENOW3_None=0;

*FLUSHOT5 Base category is FLUSHOT5=7 and 9;
if FLUSHOT5=1 then FLUSHOT5_Y=1; else FLUSHOT5_Y=0;
if FLUSHOT5=2 then FLUSHOT5_N=1; else FLUSHOT5_N=0;
```

```
*PNEUVAC3 Base category is PNEUVAC3=7 and 9;
if PNEUVAC3=1 then PNEUVAC3_Y=1; else PNEUVAC3_Y=0;
if PNEUVAC3=2 then PNEUVAC3_N=1; else PNEUVAC3_N=0;

*SEATBELT Base category is SEATBELT=7 and 9;
if SEATBELT=1 then SEATBELT_Always=1; else
SEATBELT_Always=0;
if SEATBELT=2 or SEATBELT=3 then SEATBELT_Sometimes=1; else
SEATBELT_Sometimes=0;
if SEATBELT=4 or SEATBELT=5 then SEATBELT_SeldomNever=1;
else SEATBELT_SeldomNever=0;

*HAVARTH3 Base category is HAVARTH3=7 and 9;
if HAVARTH3=1 then HAVARTH3_Y=1; else HAVARTH3_Y=0;
if HAVARTH3=2 then HAVARTH3_N=1; else HAVARTH3_N=0;

*ADDEPEV2 Base category is ADDEPEV2=7 and 9;
if ADDEPEV2=1 then ADDEPEV2_Y=1; else ADDEPEV2_Y=0;
if ADDEPEV2=2 then ADDEPEV2_N=1; else ADDEPEV2_N=0;

*CHCKIDNY Base category is CHCKIDNY=7 and 9;
if CHCKIDNY=1 then CHCKIDNY_Y=1; else CHCKIDNY_Y=0;
if CHCKIDNY=2 then CHCKIDNY_N=1; else CHCKIDNY_N=0;

*CHCVISN1 Base category is CHCVISN1=7 and 9;
if CHCVISN1=1 or CHCVISN1=3 then CHCVISN1_Y=1; else
CHCVISN1_Y=0;
if CHCVISN1=2 then CHCVISN1_N=1; else CHCVISN1_N=0;

*DIABETE3 Base category is DIABETE3=7 and 9;
if DIABETE3=1 or DIABETE3=2 then DIABETE3_Y=1; else
DIABETE3_Y=0;
if DIABETE3=3 or DIABETE3=4 then DIABETE3_N=1; else
DIABETE3_N=0;

*LASTDEN3 Base category is LASTDEN3=7 and 9;
if LASTDEN3=1 then LASTDEN3_LT1=1; else LASTDEN3_LT1=0;
if LASTDEN3=2 or LASTDEN3=3 then LASTDEN3_1LT5=1; else
LASTDEN3_1LT5=0;
if LASTDEN3=4 then LASTDEN3_GTE5=1; else LASTDEN3_GTE5=0;
```

```
if LASTDEN3=8 then LASTDEN3_Never=1; else LASTDEN3_Never=0;

*Now, subset the training set to do the analysis;
data NBTraining;
     set NBModel;
     where role='TRAINING';
run;

title "Negative Binomial Model";
PROC GENMOD Data=NBTraining;
     MODEL PHYSHLTH=HLTHPLN1_Y HLTHPLN1_N PERSDOC2_Y
PERSDOC2_N MEDCOST_Y MEDCOST_N
     HADMAM_Y HADMAM_N HADMAM_NA PROFEXAM_Y PROFEXAM_N
PROFEXAM_NA HADPAP2_Y HADPAP2_N HADPAP2_NA
     PSATEST1_Y PSATEST1_N PSATEST1_NA HIVTST6_Y HIVTST6_N
_STATE_Alabama _STATE_Alaska _STATE_Arizona
     _STATE_Arkansas _STATE_California _STATE_Colorado
_STATE_Connecticut _STATE_Delaware _STATE_DC
     _STATE_Florida _STATE_Georgia _STATE_Hawaii
_STATE_Idaho _STATE_Illinois _STATE_Indiana _STATE_Iowa
     _STATE_Kansas _STATE_Kentucky _STATE_Louisiana
_STATE_Maine _STATE_Maryland _STATE_Massachusetts
     _STATE_Michigan _STATE_Minnesota _STATE_Mississippi
_STATE_Missouri _STATE_Montana _STATE_Nebraska
     _STATE_Nevada _STATE_NewHampshire _STATE_NewJersey
_STATE_NewMexico _STATE_NewYork _STATE_NorthCarolina
     _STATE_NorthDakota _STATE_Ohio _STATE_Oklahoma
_STATE_Oregon _STATE_Pennsylvania _STATE_RhodeIsland
     _STATE_SouthCarolina _STATE_SouthDakota _STATE_Tennesee
_STATE_Texas _STATE_Utah _STATE_Vermont
     _STATE_Virginia _STATE_Washington _STATE_WestVirginia
_STATE_Wisconsin _STATE_Wyoming VETERAN3_Y
     VETERAN3_N MARITAL_MARRIED MARITAL_DIVORCED
MARITAL_WIDOWED MARITAL_SEPARATED MARITAL_NEVERMARRIED
     MARITAL_UNMARRIEDCOUPLE CVDSTRK3_Y CVDSTRK3_N ASTHMA3_Y
ASTHMA3_N CHCSCNCR_Y CHCSCNCR_N
     CHCOCNCR_Y CHCOCNCR_N CHCCOPD1_Y CHCCOPD1_N
CHECKUP1_LT1 CHECKUP1_1T5 CHECKUP1_GT5 EXERANY2_Y
     EXERANY2_N CVDINFR4_Y CVDINFR4_N CVDCRHD4_Y CVDCRHD4_N
EDUCA_None EDUCA_Elem EDUCA_SomeHS
```

```
     EDUCA_GradHS EDUCA_SomeColl EDUCA_GradColl
EMPLOY_SelfOrWages EMPLOY_Unemployed EMPLOY_Homemaker
     EMPLOY_Student EMPLOY_Retired EMPLOY_Unable
INCOME2_LT10K INCOME2_10KLT25K INCOME2_25KLT50K
     INCOME2_50KPLUS PREGNANT_Y PREGNANT_N PREGNANT_NA
QLACTLM2_Y QLACTLM2_N USEEQUIP_Y USEEQUIP_N
     USENOW3_Daily USENOW3_SomeDays USENOW3_None FLUSHOT5_Y
FLUSHOT5_N PNEUVAC3_Y PNEUVAC3_N
     SEATBELT_Always SEATBELT_Sometimes SEATBELT_SeldomNever
HAVARTH3_Y HAVARTH3_N ADDEPEV2_Y
     ADDEPEV2_N CHCKIDNY_Y CHCKIDNY_N CHCVISN1_Y CHCVISN1_N
DIABETE3_Y DIABETE3_N LASTDEN3_LT1
     LASTDEN3_1LT5 LASTDEN3_GTE5 LASTDEN3_Never AGE CHILDREN
WEIGHT2 FALL12MN/dist=negbin SELECTION=BACKWARD;
RUN;
```