Zach, Collin, Yiqun, Siyu

12/14/2018

Final Project

For our final project we considered multiple variables to be our response variable, and ultimately ended on 'The Number of Prescription Medicines You are Supposed to Take'. We chose this variable because we were intrigued to see the results of regression analysis on it, to see what actually affected the estimator. There are many different variables that are classically associated with our response variable and while some of them may be grounded in real statistical evidence, none of us had done any actual research previously to see if they were fact. There was also grounds to consider potential concerning factors our predictors introduced to an ever growing senior population. Taking prescription medications can become an arduous and taxing process in one's everyday life, especially in old age. It can also become exceedingly more expensive depending on the specialization of the drug and the sheer number of medications that an individual needs to take. If any of our predictors show a possibility of an increased number of prescription medications in our model, then we can address such predictors and suggest emphasis on characteristics or habits that can benefit our population. This project provided us with the tools we would need to test these hypotheses and would also allow us to practice our SAS and analysis skills.

Exploratory Analysis:

Our first step in our exploratory analysis was to choose which data set we were interested in. We had a selection of four separate sets, and did a very quick skim through the summary and data contained in each of them. After looking through each of the sets we decided that we wanted to use the wave 2 survivor data. Our reasoning was that it had the largest sample size and variables that we were interested. There was an initial debate that a decedent data analysis would be more telling of an individual's health, seeing as it is essentially a retrospective analysis of their habits prior to passing away. We had not chosen a response variable at this point. Once we had selected our data set we went to work with going through each of the variables contained in

the set and trying to come to a group consensus on which variable to choose as our response variable. After some deliberation over which one would be most interesting to our group as a whole we chose 'The Number of Prescription Medicines You are Supposed to Take' (RV). The data provided in the variable was sufficient enough for us to work with. Additionally, we had the added motivation to observe trends in the senior population that would yield an increase to the necessary number of medications they need to take; a potential reflection of their ongoing health.

Since we had already gone through most of the variables in the set at this point the next step of choosing our 15 independent variables was much quicker and we ended with these 15 (See Figure 1). Some variables provided a more explicit relationship to our response variable. Health status, health and veteran status are examples of such variables with these relationships. A veteran could have experienced any form of trauma due to war and may need medications to supplement them; a given war affecting some veterans differently or more than others. Education level, for instance, provided an implicit relationship, where some more physically demanding trade jobs were more taxing on an individual than someone with a degree-based career.

There was a variety of other variables that we had been interested in, but unfortunately they contained mostly missing values and would lower our sample size by too great of an amount to include in our final regression analysis. Of our 15 variables, we additionally decided to highlight 4 variables of special interest within this set (See Figure 2). These were 4 variables that we felt especially strong about, for better or worse. We chose these 4 variables based off of general intuition of our subject matter; essentially if we could gleam an inherent relationship between our response and the given predictor. We deemed 2 variables to be ones that would persist through each generation of our models up to our final result. For instance, health status felt like a given in our model, since it potentially captures poor health, and in turn, a necessity to take multiple types of prescription medication. Meanwhile, we predicted the other 2 variables to eventually vanish from our final model due to the lack of an inherent relationship. Region and time married carried no clear relationship to our response and were some of the variables we chose last in our exploratory analysis. As such, we felt not nearly as strong about them than the variables we chose at the beginning. In selecting these 4 variables, we felt that our analysis could inspire conversation about our resulting final model. For instance, we could observe whether our

intuition about our choice of predictors was relevant, or if the variables we felt had little impact on the final model would eventually have relevancy.

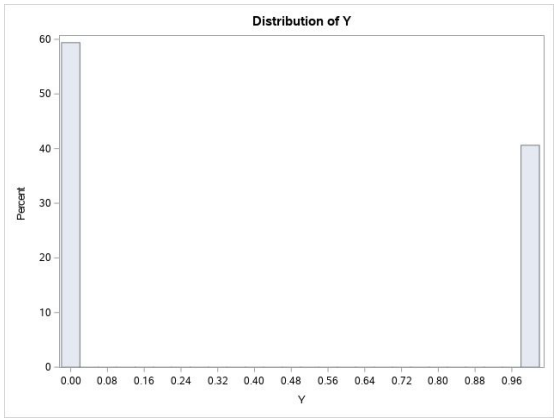
Variable
Veteran Status
Type of Residence
Current Smoker
Health Status
Health Comparison to last Survey
Education level of individual
Family Income
Length of Time Married
Sex
Main Racial Background
Class of Worker
NHIS Poverty Index
Condition List Assigned
Region
Activity level compared to a year ago

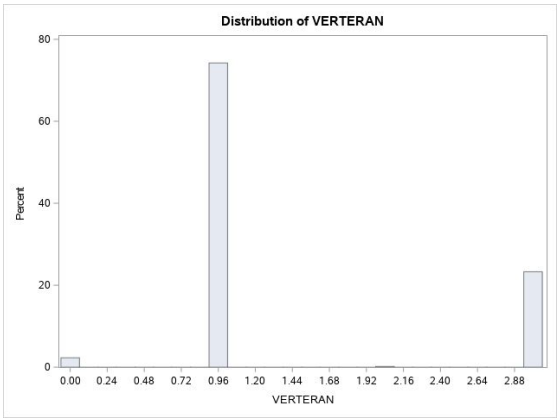
Figure 1: Predictor Variables Chosen.

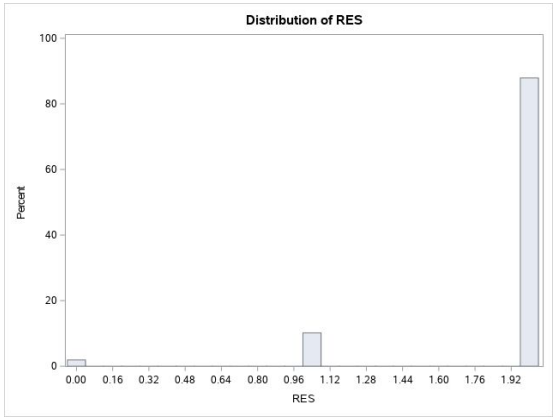
Variables of Interest
Health Status
Family Income
Region
Length of Time Married to Spouse

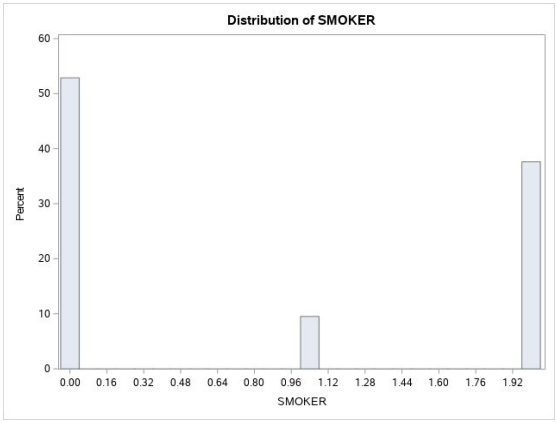
Figure 2: Predictors of Particular Interest. Green are predicted to remain, while red predicted likely to disappear.

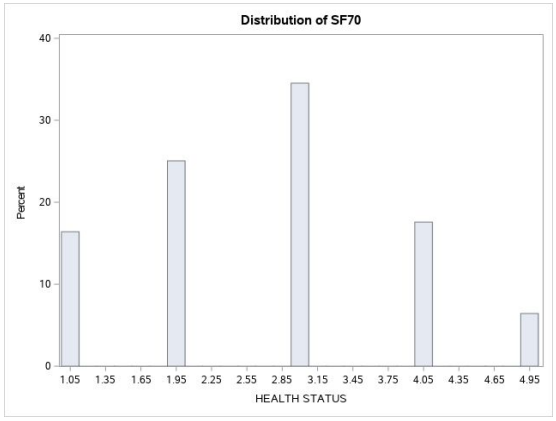
Our next step was to clean the data and then form our initial analysis: creating histograms, looking at the covariance, and looking at the correlation for each of the variables. These are our results:

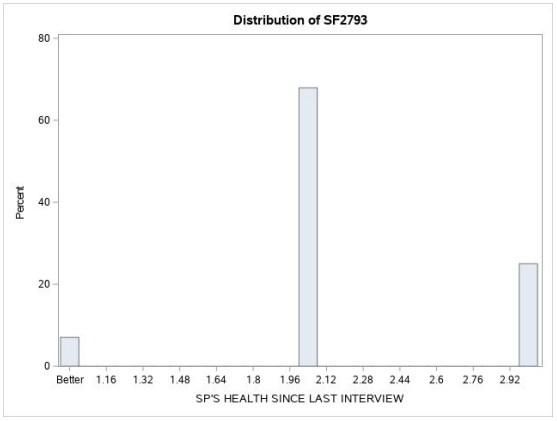

Distribution of Y
Percent
60
50
40
30
20
10
0
0.00 0.08 0.16 0.24 0.32 0.40 0.48 0.56 0.64 0.72 0.80 0.88 0.96
Y

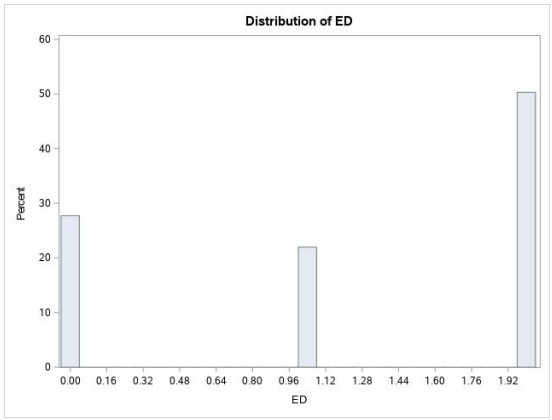

Distribution of VERTERAN
Percent
80
60
40
20
0
0.00 0.24 0.48 0.72 0.96 1.20 1.44 1.68 1.92 2.16 2.40 2.64 2.88
VERTERAN

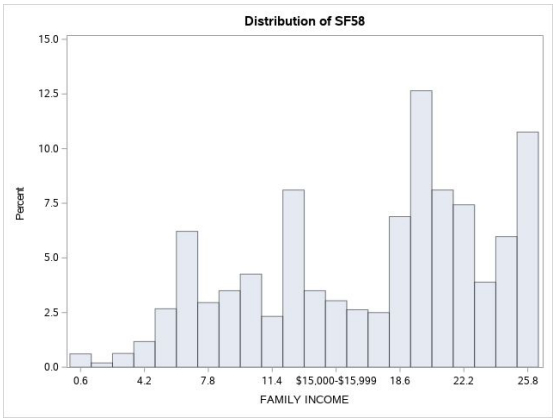

Distribution of RES
Percent
100
80
60
40
20
0
0.00 0.16 0.32 0.48 0.64 0.80 0.96 1.12 1.28 1.44 1.60 1.76 1.92
RES

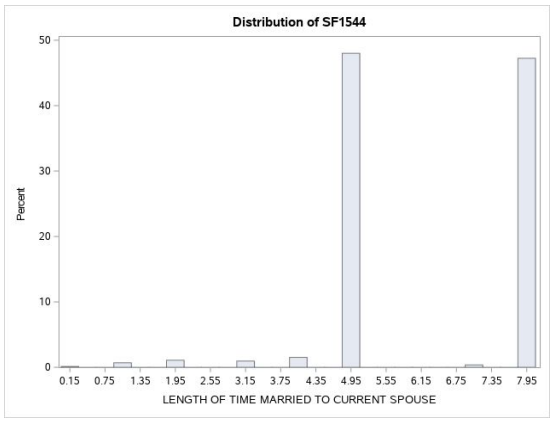

Distribution of SMOKER
Percent
60
50
40
30
20
10
0
0.00 0.16 0.32 0.48 0.64 0.80 0.96 1.12 1.28 1.44 1.60 1.76 1.92
SMOKER

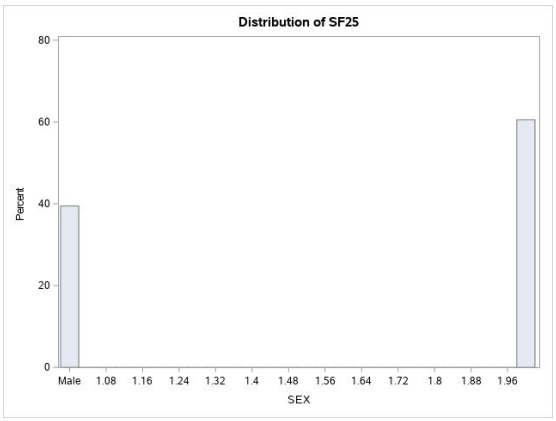

Distribution of SF70
Percent
40
30
20
10
0
1.05 1.35 1.65 1.95 2.25 2.55 2.85 3.15 3.45 3.75 4.05 4.35 4.65 4.95
HEALTH STATUS

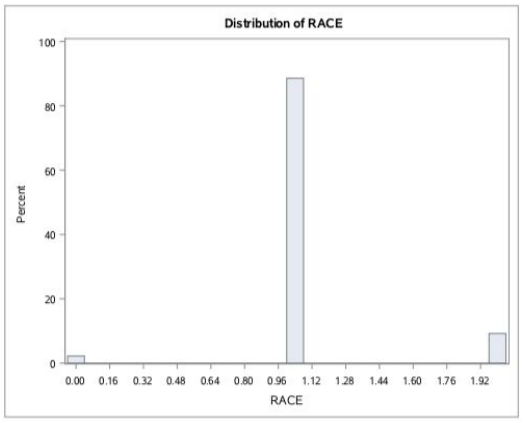

Distribution of SF2793
Percent
80
60
40
20
0
Better 1.16 1.32 1.48 1.64 1.8 1.96 2.12 2.28 2.44 2.6 2.76 2.92
SP'S HEALTH SINCE LAST INTERVIEW

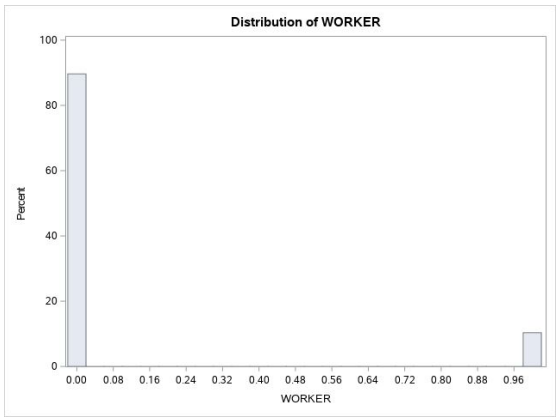

Distribution of ED
Percent
60
50
40
30
20
10
0
0.00 0.16 0.32 0.48 0.64 0.80 0.96 1.12 1.28 1.44 1.60 1.76 1.92
ED

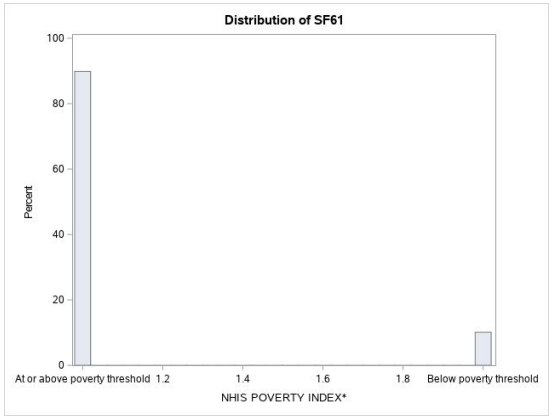

Distribution of SF58
Percent
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.6 4.2 7.8 11.4 $15,000-$15,999 18.6 22.2 25.8
FAMILY INCOME

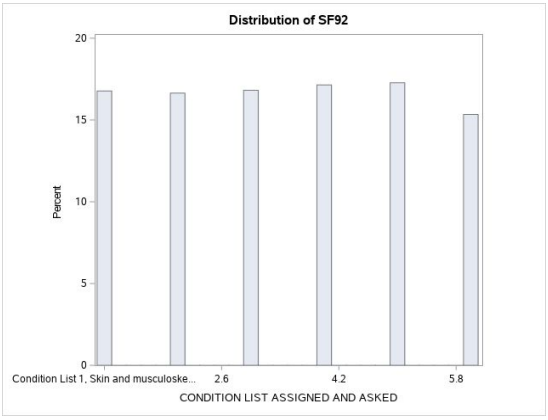

Distribution of SF1544
Percent
50
40
30
20
10
0
0.15 0.75 1.35 1.95 2.55 3.15 3.75 4.35 4.95 5.55 6.15 6.75 7.35 7.95
LENGTH OF TIME MARRIED TO CURRENT SPOUSE

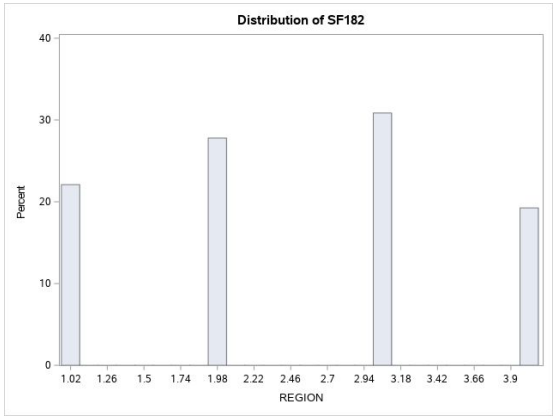

Distribution of SF25
Percent
80
60
40
20
0
Male 1.08 1.16 1.24 1.32 1.4 1.48 1.56 1.64 1.72 1.8 1.88 1.96
SEX

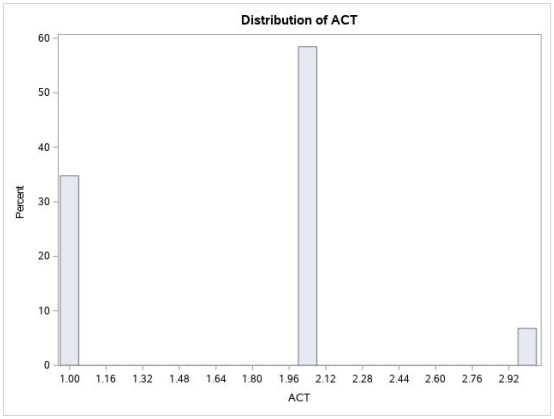

Distribution of RACE
Percent
100
80
60
40
20
0
0.00 0.16 0.32 0.48 0.64 0.80 0.96 1.12 1.28 1.44 1.60 1.76 1.92
RACE


Distribution of WORKER
Percent
100
80
60
40
20
0
0.00 0.08 0.16 0.24 0.32 0.40 0.48 0.56 0.64 0.72 0.80 0.88 0.96
WORKER









Pearson Correlation Coefficients, N = 4602 Prob > |r| under H0: Rho=0	SF29	VERTERAN	RES	SMOKER	SF70	SF2793	ED	SF58	SF1544
SF29 AGE	1.00000	-0.13340 <.0001	-0.08798 <.0001	-0.10114 <.0001	0.02811 0.0565	0.07872 <.0001	-0.00384 0.7947	-0.14117 <.0001	0.19213 <.0001
VERTERAN	-0.13340 <.0001	1.00000	0.07035 <.0001	0.23063 <.0001	-0.03658 0.0131	-0.02193 0.1368	-0.07271 <.0001	0.19154 <.0001	-0.26792 <.0001
RES	-0.08798 <.0001	0.07035 <.0001	1.00000	0.03299 0.0252	-0.01727 0.2413	-0.01972 0.1811	-0.00614 0.6773	0.15768 <.0001	-0.18220 <.0001
SMOKER	-0.10114 <.0001	0.23063 <.0001	0.03299 0.0252	1.00000	0.00339 0.8180	0.00388 0.7926	-0.01884 0.2014	0.12031 <.0001	-0.13778 <.0001
SF70 HEALTH STATUS	0.02811 0.0565	-0.03658 0.0131	-0.01727 0.2413	0.00339 0.8180	1.00000	0.23401 <.0001	0.11311 <.0001	-0.18886 <.0001	0.03274 0.0263
SF2793 SF'S HEALTH SINCE LAST INTERVIEW	0.07872 <.0001	-0.02193 0.1368	-0.01972 0.1811	0.00388 0.7926	0.23401 <.0001	1.00000	0.00024 0.9869	-0.06909 <.0001	0.02387 0.1054
ED	-0.00384 0.7947	-0.07271 <.0001	-0.00614 0.6773	-0.01884 0.2014	0.11311 <.0001	0.00024 0.9869	1.00000	-0.21725 <.0001	0.07362 <.0001
SF58 FAMILY INCOME	-0.14117 <.0001	0.19154 <.0001	0.15768 <.0001	0.12031 <.0001	-0.18886 <.0001	-0.06909 <.0001	-0.21725 <.0001	1.00000	-0.38423 <.0001
SF1544 LENGTH OF TIME MARRIED TO CURRENT SPOUSE	0.19213 <.0001	-0.26792 <.0001	-0.18220 <.0001	-0.13778 <.0001	0.03274 0.0263	0.02387 0.1054	0.07362 <.0001	-0.38423 <.0001	1.00000
SF25 SEX	0.06272 <.0001	-0.58058 <.0001	-0.08610 <.0001	-0.30542 <.0001	0.03795 0.0100	0.00575 0.6965	0.09475 <.0001	-0.23432 <.0001	0.38990 <.0001
RACE	0.01208 0.4127	-0.01628 0.2696	-0.02769 0.0603	-0.00680 0.6446	0.08906 <.0001	-0.00608 0.6801	0.05617 0.0001	-0.18352 <.0001	0.08715 <.0001
WORKER	-0.14324 <.0001	0.10843 <.0001	0.01722 0.2428	0.02402 0.1032	-0.11261 <.0001	-0.04575 0.0019	-0.08645 <.0001	0.11432 <.0001	-0.05189 0.0004
SF61 NHIS POVERTY INDEX*	0.05097 0.0005	-0.11260 <.0001	-0.09367 <.0001	-0.07512 <.0001	0.13245 <.0001	0.04473 0.0024	0.05380 0.0003	-0.58036 <.0001	0.20638 <.0001
SF92 CONDITION LIST ASSIGNED AND ASKED	-0.01396 0.3437	-0.01850 0.2095	-0.00063 0.9659	-0.00498 0.7353	-0.01821 0.2167	0.00131 0.9293	-0.01036 0.4825	0.01791 0.2244	-0.01029 0.4852
SF182 REGION	0.00893 0.5447	0.01756 0.2337	0.09644 <.0001	0.00597 0.6854	0.01945 0.1870	0.00882 0.5498	-0.07187 <.0001	0.01442 0.3281	-0.02527 0.0865
ACT	-0.12711 <.0001	0.05827 <.0001	0.01105 0.4536	0.00260 0.8600	-0.20407 <.0001	-0.04668 <.0001	-0.00053 0.9714	0.08273 <.0001	-0.06174 <.0001
Y	0.03482 0.0182	-0.04517 0.0022	-0.02384 0.1059	0.01607 0.2757	0.28782 <.0001	0.08941 <.0001	0.02385 0.1057	-0.04786 0.0012	0.05191 0.0004

Pearson Correlation Coefficients, N = 4602 Prob > |r| under H0: Rho=0	SF25	RACE	WORKER	SF61	SF92	SF182	ACT	Y
SF29 AGE	0.06272 <.0001	0.01208 0.4127	-0.14324 <.0001	0.05097 0.0005	-0.01396 0.3437	0.00893 0.5447	-0.12711 <.0001	0.03482 0.0182
VERTERAN	-0.58058 <.0001	-0.01628 0.2696	0.10843 <.0001	-0.11260 <.0001	-0.01850 0.2095	0.01756 0.2337	0.05827 <.0001	-0.04517 0.0022
RES	-0.08610 <.0001	-0.02769 0.0603	0.01722 0.2428	-0.09367 <.0001	-0.00063 0.9659	0.09644 <.0001	0.01105 0.4536	-0.02384 0.1059
SMOKER	-0.30542 <.0001	-0.00680 0.6446	0.02402 0.1032	-0.07512 <.0001	-0.00498 0.7353	0.00597 0.6854	0.00260 0.8600	0.01607 0.2757
SF70 HEALTH STATUS	0.03795 0.0100	0.08906 <.0001	-0.11261 <.0001	0.13245 <.0001	-0.01821 0.2167	0.01945 0.1870	-0.20407 <.0001	0.28782 <.0001
SF2793 SF'S HEALTH SINCE LAST INTERVIEW	0.00575 0.6965	-0.00608 0.6801	-0.04575 0.0019	0.04473 0.0024	0.00131 0.9293	0.00882 0.5498	-0.46668 <.0001	0.08941 <.0001
ED	0.09475 <.0001	0.05617 0.0001	-0.08645 <.0001	0.05380 0.0003	-0.01036 0.4825	-0.07187 <.0001	-0.00053 0.9714	0.02385 0.1057
SF58 FAMILY INCOME	-0.23432 <.0001	-0.18352 <.0001	0.11432 <.0001	-0.58036 <.0001	0.01791 0.2244	0.01442 0.3281	0.08273 <.0001	-0.04786 0.0012
SF1544 LENGTH OF TIME MARRIED TO CURRENT SPOUSE	0.38990 <.0001	0.08715 <.0001	-0.05189 0.0004	0.20638 <.0001	-0.01029 0.4852	-0.02527 0.0865	-0.06174 <.0001	0.05191 0.0004
SF25 SEX	1.00000	0.04359 0.0031	-0.13017 <.0001	0.13689 <.0001	0.00395 0.7887	-0.02474 0.0934	-0.00862 0.5588	0.07018 <.0001
RACE	0.04359 0.0031	1.00000	-0.01166 0.4291	0.18341 <.0001	0.01107 0.4527	-0.01434 0.3309	-0.00068 0.9634	0.03718 0.0116
WORKER	-0.13017 <.0001	-0.01166 0.4291	1.00000	-0.05488 0.0002	0.03285 0.0258	0.00209 0.8875	0.03960 0.0072	-0.07166 <.0001
SF61 NHIS POVERTY INDEX*	0.13689 <.0001	0.18341 <.0001	-0.05488 0.0002	1.00000	-0.03217 0.0291	0.04913 0.0009	-0.04289 0.0036	0.03776 0.0104
SF92 CONDITION LIST ASSIGNED AND ASKED	0.00395 0.7887	0.01107 0.4527	0.03285 0.0258	-0.03217 0.0291	1.00000	-0.00925 0.5304	-0.01958 0.1843	0.00309 0.8340
SF182 REGION	-0.02474 0.0934	-0.01434 0.3309	0.00209 0.8875	0.04913 0.0009	-0.00925 0.5304	1.00000	-0.01958 0.1843	0.00413 0.7797
ACT	-0.00862 0.5588	-0.00068 0.9634	0.03960 0.0072	-0.04289 0.0036	0.01397 0.3432	-0.01958 0.1843	1.00000	-0.08451 <.0001
Y	0.07018 <.0001	0.03718 0.0116	-0.07166 <.0001	0.03776 0.0104	0.00309 0.8340	0.00413 0.7797	-0.08451 <.0001	1.00000

Pearson's correlation coefficient:

r(SF70,Y)=0.28782 P<0.0001, So Health Status and Y have significant correlation.

r(SF58,Y)=-0.04786 P=0.0012, So Family Income and Y have significant correlation.

r(SF182, Y)=0.00413 P=0.7797>0.05, So Region and Y are independent.

r(SF1544,Y)=0.05191 P=0.0004, Surprisingly, Length of time married and Y have significant correlation.

Regression Analysis:

After we had finished our exploratory analysis we began to work on our regression of the model. We tried a variety of different regression methods: binomial with logit link, normal with identity link, negative binomial with log link, poisson with log link, Forward logit binomial regression and backward logit binomial regression. Most of our AICs came out to be very large numbers, but the smallest one that we found was through logit regression using the backwards method. These are the results of all of our regression:

Binomial Regression

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-2863.2333	
Full Log Likelihood		-2863.2333	
AIC (smaller is better)		5786.4666	
AICC (smaller is better)		5786.8735	
BIC (smaller is better)		5979.4940	

Normal Regression

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4572	996.4366	0.2179
Scaled Deviance	4572	4602.0000	1.0066
Pearson Chi-Square	4572	996.4366	0.2179
Scaled Pearson X2	4572	4602.0000	1.0066
Log Likelihood		-3009.2854	
Full Log Likelihood		-3009.2854	
AIC (smaller is better)		6080.5707	
AICC (smaller is better)		6081.0049	
BIC (smaller is better)		6280.0324	

Poisson Regression

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4572	3091.5615	0.6762
Scaled Deviance	4572	3091.5615	0.6762
Pearson Chi-Square	4572	2710.0525	0.5927
Scaled Pearson X2	4572	2710.0525	0.5927
Log Likelihood		-3414.7807	
Full Log Likelihood		-3414.7807	
AIC (smaller is better)		6889.5615	
AICC (smaller is better)		6889.9684	
BIC (smaller is better)		7082.5889	

Negative Binomial

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4572	3091.5615	0.6762
Scaled Deviance	4572	3091.5615	0.6762
Pearson Chi-Square	4572	2710.0524	0.5927
Scaled Pearson X2	4572	2710.0524	0.5927
Log Likelihood		-3414.7807	
Full Log Likelihood		-3414.7807	
AIC (smaller is better)		6891.5615	
AICC (smaller is better)		6891.9956	
BIC (smaller is better)		7091.0231	

Forwards Regression

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6218.549	5763.853
SC	6224.983	5815.326
-2 Log L	6216.549	5747.853

Backwards Regression

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	6107.415	5535.962
SC	6113.836	5600.171
-2 Log L	6105.415	5515.962

Conclusion:

After adding multiple iterations and drop some data whose standard residual greater than 2, we were finally able to come to a final model:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
INTERCEPT		1	-2.2469	0.2597	74.8458	<.0001
SMOKER	1	1	-0.5051	0.1209	17.4523	<.0001
SMOKER	2	1	0.2012	0.0725	7.7044	0.0055
SF70		1	0.6229	0.0324	370.0567	<.0001
SF58		1	0.0132	0.00546	5.8193	0.0159
SF1544		1	0.0497	0.0224	4.9265	0.0264
SF25	Male	1	-0.3527	0.0764	21.3255	<.0001
WORKER	1	1	-0.3458	0.1159	8.8937	0.0029
ACT	1	1	-0.1517	0.1337	1.2862	0.2568
ACT	2	1	-0.4706	0.1290	13.3063	0.0003

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SMOKER 1 vs 0	0.603	0.476	0.765
SMOKER 2 vs 0	1.223	1.061	1.410
SF70	1.864	1.750	1.987
SF58	1.013	1.002	1.024
SF1544	1.051	1.006	1.098
SF25 Male vs Female	0.703	0.605	0.816
WORKER 1 vs 0	0.708	0.564	0.888
ACT 1 vs 3	0.859	0.661	1.117
ACT 2 vs 3	0.625	0.485	0.804

*Y^=-2.2469-0.5051*SMOK0ER1+0.2012*SMOKER2+0.6229*SF70+0.0132*SF58+0.0497*SF1544-0.3527*SF25 (male)- 0.3458*WORKER-0.1517*ACT1-.47*ACT2*

(Smoker1, Smoker2)=(1.0) Current Smoker (Smoker1, Smoker2)=(0,1) Former Smoker

(Smoker1, Smoker2)=(0,0) Non-Smoker (Act1,Act2) =(1,0) Less activities

(Act1,Act2)=(0,1) The same activities Worker1=1 Others Worker1=0 Not in the labor force

This model, using backwards elimination with outlier data removed and appropriate categorical variables, gave us the lowest AIC value out of any of our models. Both of our expected variables to make the final model did in fact make the final model, and one of the variables that we anticipated not making it to the final model did also make it in. The variable SF1544, or length of time married to spouse, was ultimately added to our final model. We originally did not account for age when we did our regression which had also resulted in SF1544 being included; we then believed that it was the lack of the age variable that was causing this. We changed our code to include age, although it did not cause the SF1544 variable to leave the model. We have concluded then that it must be significant. The AIC of this final model is not only lower than the AIC for the other models, but compared to the changes in AIC that we had been observing previously, it is also significantly lower.

Looking back at the explanatory variable analysis, only smoker was included in the model with the R(Smoker,Y)=0.01607 P=0.2757 which has insignificant correlation. All the other variables shows significant correlation with Y. So basically, this model included all the useful variables.

In our model, compared with Non-Smokers, former smokers have larger odds to take more medicines while current smokers have lower odds. More married years, more income, Lower health status, female, not working in labor force and more activities than before will result in the odds of taking 3 or more prescription medicines increasing.

Predictive power:

Table of Y by Y_PRED			
	Y_PRED		
Y	1	0	Total
1	1194	614	1808
0	1016	1717	2733
Total	2210	2331	4541

Sensitivity=Pr(Y ^=1 | Y=1)=1194/1808=0.6604

Specificity=Pr(Y ^=0 | Y=0)=1717/2733=0.6283 (If Y^>0.40613 then Y^=1 else Y^=0)

Model Checking:

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4531	5515.9617	1.2174
Scaled Deviance	4531	5515.9617	1.2174
Pearson Chi-Square	4531	4478.7705	0.9885
Scaled Pearson X2	4531	4478.7705	0.9885
Log Likelihood		-2757.9809	
Full Log Likelihood		-2757.9809	
AIC (smaller is better)		5535.9617	
AICC (smaller is better)		5536.0103	
BIC (smaller is better)		5600.1708	

Deviance=5515.9617 with df=4531 P<0.00001

But according to the Pearson's Chi square P=0.70695 So, the model fits data well.

SAS Code:

```
  /* SAS Input statements for W2.SF.DATA.SF11.D020321.DAT */

options nocenter mergenoby = error validvarname = upcase;
options formchar='|-+'; /* Set Vertical, Horizontal, Intersections */

%let w2sf     =\\Client\C$\SAS\Data\w2_sf_data_sf11_d020321.dat ;



PROC FORMAT;

   VALUE P11L (default=40)
      1 ='Under 5 years'
      2 ='5-17 years'
      3 ='18-24 years'
      4 ='25-44 years'
      5 ='45-64 years'
      6 ='65-69 years'
      7 ='70-74 years'
      8 ='75 years and over'
;
   VALUE  P916L (default=40)
      0 = "None"
      1 = "One or Two"
      2 = "Three to five"
      3 = "Six to Nine"
      4 = "Ten or More"
      8 = "Not ascertained"
      9 = "DK or refused"

;
   VALUE P24L (default=40)
     1='Non-veteran'
     2='WW I'
     3='WW II'
     4='Korean War'
     5='Vietnam veteran'
     6='Post-Vietnam'
     7='Other service'
     8='Served in Armed Forces, unknown if wa...' /*r veteran*/
     9='Unknown if served in Armed Forces'
;
   VALUE P126L (default=40)
      01 ='Single family house or townhouse, not...' /* part of retirement community*/
      02 ='Single family house, townhouse, or ap...' /*artment that is part of retirement community*/
      03 ='Regular apartment'
      04 ='Supervised apartment'
      05 ='Group home'
      06 ='Halfway house'
      07 ='Personal care or board and care home'
      08 ='Developmental center'
      09 ='Other type of supervised group reside...' /*nce or facility*/
      10 ='Assisted living facility'
      11 ='Nursing or convalescent home'
      12 ='Retirement home'
      13 ='Center for Independent Living'
      14 ='Something else'
      98 ='Not ascertained'
      99 ='DK or refused'
;
```

```
  VALUE P1197L (default=40)
     0 ='Never smoked'
     1 ='Current smoker'
     2 ='Former smoker'
     3 ='Unknown if current/former smoker'
     9 ='Unknown if ever smoked'
;
  VALUE P40L (default=40)
     1 ='Excellent'
     2 ='Very Good'
     3 ='Good'
     4 ='Fair'
     5 ='Poor'
     6 ='Unknown'
;
   VALUE Q206Q
    93 = "SAQ Leg. Skip"
    94 = "Missing-Breakoff"
    95 = "Missing-CATI Error"
    96 = "Missing (SAQ)"
    97 = "Refused"
    98 = "Ineligible for Wave 2; Proxy Decedent interview"
    99 = "Dont Know"
     1 = "Better"
     2 = "About the Same"
     3 = "Worse"
;
  VALUE P27L (default=40)
    0='None; kindergarten only'
    1='1-8 years (elementary)'
    2='9-11 years (high school)'
    3='12 years (high school graduate)'
    4='1-3 years (college)'
    5='4 years (college graduate)'
    6='5 or more years (post-college)'
    7='Unknown'
;
  VALUE P31L (default=40)
     00 ='Less than  $1,000'
     01 ='$1,000-$1,999'
     02 ='$2,000-$2,999'
     03 ='$3,000-$3,999'
     04 ='$4,000-$4,999'
     05 ='$5,000-$5,999'
     06 ='$6,000-$6,999'
     07 ='$7,000-$7,999'
     08 ='$8,000-$8,999'
     09 ='$9,000-$9,999'
     10 ='$10,000-$10,999'
     11 ='$11,000-$11,999'
     12 ='$12,000-$12,999'
     13 ='$13,000-$13,999'
     14 ='$14,000-$14,999'
     15 ='$15,000-$15,999'
     16 ='$16,000-$16,999'
     17 ='$17,000-$17,999'
     18 ='$18,000-$18,999'
     19 ='$19,000-$19,999'
     20 ='$20,000-$24,999'
     21 ='$25,000-$29,999'
     22 ='$30,000-$34,999'
     23 ='$35,000-$39,999'
     24 ='$40,000-$44,999'
     25 ='$45,000-$49,999'
     26 ='$50,000 and over'
     27 ='Unknown'
```

```
;
  VALUE P1053L (default=40)
     0 ='Less than 1 year'
     1 ='1-4 years'
     2 ='5-9 years'
     3 ='10-14 years'
     4 ='15-19 years'
     5 ='20 years and over'
     7 ='Unknown how long married'
     8 ='Not married'
     9 ='Unknown current marital status'
;
  VALUE P9L (default=40)
     1 ='Male'
     2 ='Female'
;
  VALUE P18L (default=40)
     01 ='White'
     02 ='Black'
     03 ='Indian (American)'
     04 ='Eskimo'
     05 ='Aleut'
     06 ='Chinese'
     07 ='Filipino'
     08 ='Hawaiian'
     09 ='Korean'
     10 ='Vietnamese'
     11 ='Japanese'
     12 ='Asian Indian'
     13 ='Samoan'
     14 ='Guamanian'
     15 ='Other API'
     16 ='Other Race'
     17 ='Multiple Race'
     99 ='Unknown'
;
  VALUE P46L (default=40)
    0='Not in labor force'
    1='Private company'
    2='Federal Government employee'
    3='State Government employee'
    4='Local Government employee'
    5='Incorporated business'
    6='Self-employed'
    7='Without pay'
    8='Never worked'
    9='Unknown'
;
  VALUE P33L (default=40)
     1 ='At or above poverty threshold'
     2 ='Below poverty threshold'
     3 ='Unknown'
;
  VALUE P54L (default=40)
     1 ='Condition List 1, Skin and musculoske...' /*letal*/
     2 ='Condition List 2, Impairments'
     3 ='Condition List 3, Digestive'
     4 ='Condition List 4, Miscellaneous'
     5 ='Condition List 5, Circulatory'
     6 ='Condition List 6, Respiratory'
     7 ='Unknown'
;
  VALUE P65L (default=40)
     1 ='Northeast'
     2 ='Midwest'
     3 ='South'
```

```
      4 ='West'
;
   VALUE Q455Q
    93 = "SAQ Leg. Skip"
    94 = "Missing-Breakoff"
    95 = "Missing-CATI Error"
    96 = "Missing (SAQ)"
    97 = "Refused"
    98 = "Ineligible for Wave 2; Proxy Decedent interview"
    99 = "Dont Know"
     1 = "More Active"
     2 = "Less Active"
     3 = "About the Same"
;
   VALUE P1198L (default=40)
      0 ='Never smoked'
      1 ='Current smoker-everyday'
      2 ='Current smoker-some days'
      3 ='Former smoker'
      4 ='Unknown if current/former smoker'
      9 ='Unknown if ever smoked'
;
run;

data w2sf11 ;

  length
        SF29     3
        SF1351   3
        SF49     3
        SF431    3
        SF1770   3
        SF70     3
        SF2793   3
        SF53     3
        SF58     3
        SF1544   3
        SF25     3
        SF41     3
        SF76     3
        SF61     3
        SF92     3
        SF182    3
        SF2821   3

;
label
SF29     = "AGE"
SF1351   = "NUMBER OF PRESCRIPTION MEDICINES YOU ARE SUPPOSED TO TAKE "
SF49     = "VETERAN STATUS "
SF431    = "TYPE OF RESIDENCE "
SF1770   = "CURRENT SMOKER I "
SF70     = "HEALTH STATUS "
SF2793   = "SP'S HEALTH SINCE LAST INTERVIEW "
SF53     = "EDUCATION OF INDIVIDUAL RECODE "
SF58     = "FAMILY INCOME "
SF1544   = "LENGTH OF TIME MARRIED TO CURRENT SPOUSE "
SF25     = "SEX "
SF41     = "MAIN RACIAL BACKGROUND - REPORTED "
SF76     = "CLASS OF WORKER "
SF61     = "NHIS POVERTY INDEX* "
SF92     = "CONDITION LIST ASSIGNED AND ASKED "
SF182    = "REGION "
SF2821   = "PHYSICAL ACTIVITY MORE/LESS/SAME "
;
```

```
infile "&w2sf" lrecl=3660 pad ;

input
    SF29      29 - 29
    SF1351    1351- 1351
    SF49      49  - 49
    SF431     431 - 432
    SF1770    1770- 1770
    SF70      70  - 70
    SF2793    2793- 2794
    SF53      53  - 53
    SF58      58  - 59
    SF1544    1544- 1544
    SF25      25  - 25
    SF41      41  - 42
    SF76      76  - 76
    SF61      61  - 61
    SF92      92  - 92
    SF182     182 - 182
    SF2821    2821- 2822

              ;

FORMAT   SF29     p11L.    ;
FORMAT   SF1351   P916L.   ;
FORMAT   SF49     P24L.    ;
FORMAT   SF431    P126L.   ;
FORMAT   SF1770   P1198L.  ;
FORMAT   SF70     P40L.    ;
FORMAT   SF2793   Q206Q.   ;
FORMAT   SF53     P27L.    ;
FORMAT   SF58     P31L.    ;
FORMAT   SF1544   P1053L.  ;
FORMAT   SF25     P9L.     ;
FORMAT   SF41     P18L.    ;
FORMAT   SF76     P46L.    ;
FORMAT   SF61     P33L.    ;
FORMAT   SF92     P54L.    ;
FORMAT   SF182    P65L.    ;
FORMAT   SF2821   Q455Q.   ;

if SF1351=8 OR SF1351=9 OR SF1351="." THEN DELETE;
 if SF1351>=2 and SF1351<=4 then Y=1;
 else Y=0;

 if SF49=9 OR SF49="." THEN DELETE;
 If SF49=1 THEN Verteran=1;
 if SF49=2 then Verteran=2;
 IF SF49=3 then Verteran=3;
 IF SF49>=4 THEN Verteran=0;

 if SF431=98 OR SF431=99 OR SF431="." THEN DELETE;
 if SF431=01 or SF431=02 then Res=2;
 if SF431=03 OR SF431=04 then Res=1;
 if SF431>=05 then Res=0;

 if SF1770=4 OR SF1770=9 OR SF1770="." THEN DELETE;
 if SF1770=0 then Smoker=0;
 if SF1770=1 OR SF1770=2 then Smoker=1;
 if SF1770=3 then Smoker=2;

 if SF70=6 OR  SF70="." THEN DELETE;

 if SF2793<=99 AND SF2793>=93 OR SF2793="." THEN DELETE;

 if SF53=7 OR SF53="." THEN DELETE;
```

```
 if SF53<=1 then Ed=1;
 if SF53<=3 and SF53>=2 then Ed=2;
 if SF53>3 THEN Ed=0;


 if SF58=27 OR SF58="." THEN DELETE;

 if SF1544=9 OR SF1544="." THEN DELETE;

 if SF25="." THEN DELETE;

 if SF41=99 OR SF41="." THEN DELETE;
 if SF41=01 THEN Race=1;
 if SF41=02 then Race=2;
 if SF41>=03 then Race=0;

 if SF76=9 OR SF76="." THEN DELETE;
 if SF76=0 then Worker=0;
 if SF76>0 then Worker=1;

 if SF61=3 OR SF61="." THEN DELETE;

 if SF92=7 OR SF92="." THEN DELETE;

 if SF182="." THEN DELETE;

 if SF2821<=99 AND SF2821>=93 OR SF2821="." THEN DELETE;
 if SF2821=1 then Act=3;
 if SF2821=2 THEN Act=1;
 if SF2821=3 then Act=2;

 run;


 PROC SQL;
CREATE TABLE WORK.NEW AS
SELECT SF29, Verteran , Res , Smoker , SF70 , SF2793 , Ed , SF58 , SF1544 , SF25 , Race , Worker , SF61 , SF92 , SF182 , Act, Y FROM
WORK.W2SF11;
RUN;
QUIT;
/*Graph Histogram*/
ods graphics on;
ods pdf file="/folders/myfolders/final_pdf";

title 'Analysis of SF1351';
proc univariate data=WORK.NEW noprint;
  histogram Y;
run;
title 'Analysis of SF49';
proc univariate data=WORK.NEW noprint;
  histogram Verteran;
run;
title 'Analysis of SF431';
proc univariate data=WORK.NEW noprint;
  histogram Res;
run;
title 'Analysis of SF1770';
proc univariate data=WORK.NEW noprint;
  histogram Smoker;
run;
title 'Analysis of SF70';
proc univariate data=WORK.NEW noprint;
  histogram SF70;
run;
title 'Analysis of SF2793';
proc univariate data=WORK.NEW noprint;
```

```
  histogram SF2793;
run;
title 'Analysis of SF53';
proc univariate data=WORK.NEW noprint;
  histogram Ed;
run;
title 'Analysis of SF58';
proc univariate data=WORK.NEW noprint;
  histogram SF58;
run;
title 'Analysis of SF1544';
proc univariate data=WORK.NEW noprint;
  histogram SF1544;
run;
title 'Analysis of SF25';
proc univariate data=WORK.NEW noprint;
  histogram SF25;
run;
title 'Analysis of Race';
proc univariate data=WORK.NEW noprint;
  histogram Race;
run;
title 'Analysis of SF76';
proc univariate data=WORK.NEW noprint;
  histogram Worker;
run;
title 'Analysis of SF61';
proc univariate data=WORK.NEW noprint;
  histogram SF61;
run;
title 'Analysis of SF92';
proc univariate data=WORK.NEW noprint;
  histogram SF92;
run;
title 'Analysis of SF182';
proc univariate data=WORK.NEW noprint;
  histogram SF182;
run;
title 'Analysis of SF2821';
proc univariate data=WORK.NEW noprint;
  histogram Act;
run;

proc corr data=WORK.NEW plots(maxpoints=none);
run;
data w2sf11new;
set w2sf11(obs=1000);
run;
proc calis data=WORK.NEW  pcorr;
run;
ods pdf close;


/* binomial regression with logit link */
proc genmod data = work.NEW ;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0")  Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/  param = ref;
  model Y(EVENT="1") = SF29 Verteran  Res  Smoker  SF70  SF2793 Ed SF58  SF1544  SF25  Race  Worker  SF61  SF92  SF182  Act /
diat=binomial link=logit;
  run;

/*Normal regression with identity link*/
proc genmod data = work.NEW ;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0")  Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/  param = ref;
```

```
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182
Act/diat=normal link=identity ;
 run;

/*Poisson regression with log link*/
proc genmod data = work.NEW ;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0") Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/ param = ref;
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182
Act/diat=poisson link=log ;
 run;

/*negative binomial regression with log link*/
proc genmod data = work.NEW ;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0") Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/ param = ref;
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182
Act/diat=negbin link=log ;
 run;

/*Foreward binomial regression*/
proc logistic data=work.NEW;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0") Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/ param = ref;
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182 Act/
selection=foreward ;
run;

/*Backward binomial regression */
proc logistic data=work.NEW;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0") Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/ param = ref;
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182 Act /
selection=backward ;
run;

/*Do logistic binomial regression backward elimination by treating all the independent variables as numeric*/
proc logistic data=work.NEW;
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182 Act /
selection=backward ;
run;
/*AIC is higher and also against common sense*/

/*Trying to add some interaction terms and see whether they can be included in the mdel*/
proc logistic data=work.NEW;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0") Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/ param = ref;
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182 Act
SF61*SF2793 Verteran*SF29 Res*SF61 Res*Verteran Verteran*SF2793/ selection=backward ;
output out=new1 p=pred STDRESCHI=std_res RESCHI=Pearson_res;/*Residuals*/
run;
/*No Improvements, So we can just use the original backward elimination */


/*drop the outliers */
data work.new2;
set work.new1;
if std_res>2 then delete;
run;

/*Re-do the logistic backward elimination using the data without outliers*/
proc logistic data=work.NEW2;
Class Verteran(ref="0") Smoker(ref="0") Res(ref="0") Ed(ref="0") Worker(ref="0") Act(ref="3") Race(ref="0") SF25(ref="Female")
SF92(REF="Condition List 6, Respiratory") SF182(ref="West")/ param = ref;
 model Y(EVENT="1") = SF29 Verteran Res Smoker SF70 SF2793 Ed SF58 SF1544 SF25 Race Worker SF61 SF92 SF182 Act/
selection=backward ;
```

```
  output out=new1 p=pred STDRESCHI=std_res RESCHI=Pearson_res;
run;
/*AIC declined, So it is the optional model we could find*/


/* Classification table */
data new3;
 set new2;

 if pred > 0.40613 then Y_pred=1;/*The mean of Y is 0.40613*/
 else Y_pred=0;
run;

proc freq order=data;
 tables Y*Y_pred/nopercent norow nocol;
 title 'classification table';
run;
/*Sensitivity=1194/1808=0.6604   Specificity=1717/2733=0.6283*/

data new4 (keep=no SF29 Verteran  Res  Smoker  SF70  SF2793 Ed SF58  SF1544  SF25  Race  Worker  SF61  SF92  SF182  Act Y yes total);
  set NEW2;
    IF Y=1 THEN  yes=1;
    If Y=0 then yes=0;
    If Y=0 then no=1;
    IF Y=1 THEN no=0;
    total = yes+no;
run;

proc genmod data=new4 DESCENDING;* PLOTS=all;
 class  Smoker(ref="0")  Worker(ref="0") Act(ref="3") SF25(ref="Female")  /  param = ref;
 model yes/total = Smoker  SF70  SF58  SF1544  SF25   Worker  Act/ dist = binomial link = logit LRCI covb type1 type3;
 output out=out5 p=pred STDRESCHI=std_res RESCHI=Pearson_res;
run;
```