Team name - Data Explorers.

Data Set - US Mass Shooting

Title - US Mass Shooting

Team Mates - Diksha Joshi , Jimi George, FNU Preethi Prakash, Khizra Masood, Kubra Iqbal

Contents Page:

Abstract:

- The goal of this project is to determine if a potential male shooter can be predicted using different variables. We will analyze various variables such as race and location to see if there is a pattern to help us get a better understanding of our data.

Methodology:

- The US Mass Shooting data set was explored to see trends and correlations between dependent and independent variables. The dependent variable was “Males” and the independent variables were “mental health status”, “race” and “region”.

Introduction:

The data set compares shootings in the US with region, race and gender specifically focusing on males. The data set also gives detailed explanations about what happened during the shooting that took place and how many people were injured and killed. Our team conducted an analysis focusing on different regions and mental health and races amongst the male shooters. After analyzing five models – the best model picked was the one that focused on the Southwest region. The model specially showed that Caucasian Americans males conducted most of the shootings in that area. Even though shootings are largely happening in the US – gun control laws are trying to be implemented. This is one of the biggest debates that is occurring in the media these days and to stop it or partially make it a less of an issue in the country – gun laws should be supported. Thus, this analysis gives an overview of all the shootings that have occurred in the past few years with all the details. (DeLator 2014)

Methodology: The original data was pre processed and cleaned to add regions which were an extension to the original “location” variables. Additionally, the categorical variables such as : gender, race and mental health status was scrubbed into – U_gender , U_race, U_meantal_health_status. This was done so that the analysis could be accurate for our data.

URL of the dataset:

https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years/data

Data file:

Updatedshooting.xlx

Observations:

Data set consists of 12 variables and 320 observations. (Before data cleaning)

- Title - Name of mass shooting.
- Location - Where the shooting occurred. (States with Cities of the United States)
- Date - The date of the incident.
- Summary - Description of the incident.
- Fatalities - How many people were killed during the incident.
- Injured - How many people were injured during the incident.
- Total victims - How many people were killed and injured. (Fatalities + injured)
- Mental health issues - Presence or absence of a mental health problem. (Categorized as "no","unclear","unknown","yes").
- Race - Race of the shooter. (Categorized as Blank, Asian, Asian-American, asian-american/some other race, black, black American or African American, black American or African American/unknown, Latino, native American or Alaska native, other, some other race, two or more races, unknown, white, white American or European American, white American or European American/ some other race.
- Gender - Gender of the shooter. (Male/Female/Unknown/Both)
- Latitude - Coordinates of the incident.
- Longitude - Coordinates of the incident.

Observations:

New data set consists of 16 variables and 320 observations. (After data scrubbing)

- Title - Name of mass shooting.
- Region - Where the shooting occured. (Original Location variable was categorized into different US regions, including mid-atlantic, midwest, northeast, southeast, west, and unknown).
- Date - The date of the incident.
- Summary - Description of the incident.
- Fatalities - How many people were killed during the incident.
- Injured - How many people were injured during the incident.
- Total victims - How many people were killed and injured. (Fatalities + injured)
- U_Mental health - Presence or absence of a mental health problem (Mental health issues were categorized into "No"/"Yes"/"Unknown")
- U_Race - Race of the shooter. (Race was categorized into "asian","black american or african america" , "latino", "native american or alaska native", "other", "two or more races","white american or european american" and "unknown".)
- U_Gender - Gender of the shooter. (Gender was categorized as "male" , "female", "unknown" or "both".)
- Latitude - Coordinates of the incident.
- Longitude - Coordinates of the incident.

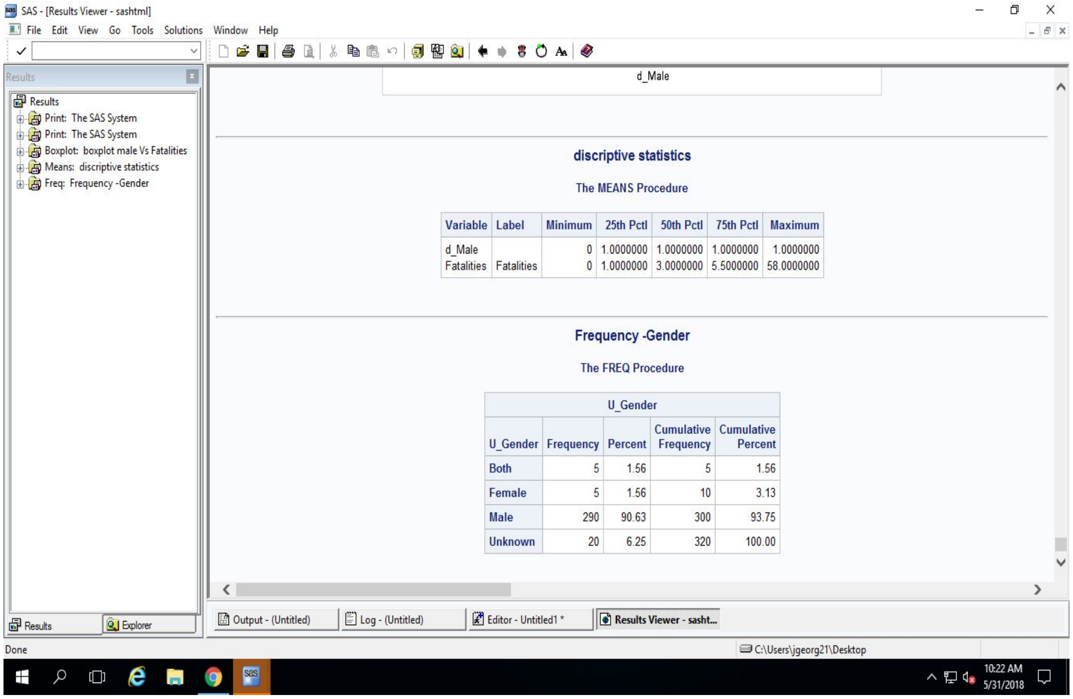
Dummy variables :

```
d_Midwest=(Region="Midwest");
d_West=(Region="West");
d_Northeast=(Region="Northeast");
d_Southeast=(Region="Southeast");
d_Southwest=(Region="Southwest");
d_Female=(U_Gender="Female");
d_Male=(U_Gender= "Male");
d_Bothgender=(U_Gender="Both");
d_White=(U_Race="White American or European American");
d_Black= (U_Race="Black American or African American");
d_Asian=(U_Race="Asian");
d_Latino= (U_Race="Latino");
d_OtherRaces=(U_Race= "Other");
d_MultipleRaces= (U_Race="Two or more races");
d_NativeAmerican= (U_Race="Native American or Alaska Native");
d_positivementalhealth=(u_mental_health="Yes");
d_negativementalhealth=(u_mental_health="No");
```

Frequency:

Data was explored by each member to verify the number of shooters in the data set. According to our data set 290 of the shooters were male and 5 were females.

```
Title"Frequency -Gender";
proc freq;
tables U_Gender;
run;
Title"Frequency -Gender VS fatalities ";
proc freq;
tables U_Gender*(Fatalities);
run;
```


SAS - [Results Viewer - sashtml]
File Edit View Go Tools Solutions Window Help
Results
Results
Print: The SAS System
Print: The SAS System
Boxplot: boxplot male Vs Fatalities
Means: discriptive statistics
Freq: Frequency -Gender
Results
Explorer
Output - (Untitled)
Log - (Untitled)
Editor - Untitled1 *
Results Viewer - sasht...
Done
C:\Users\jgeorg21\Desktop
10:22 AM
5/31/2018

d_Male

discriptive statistics

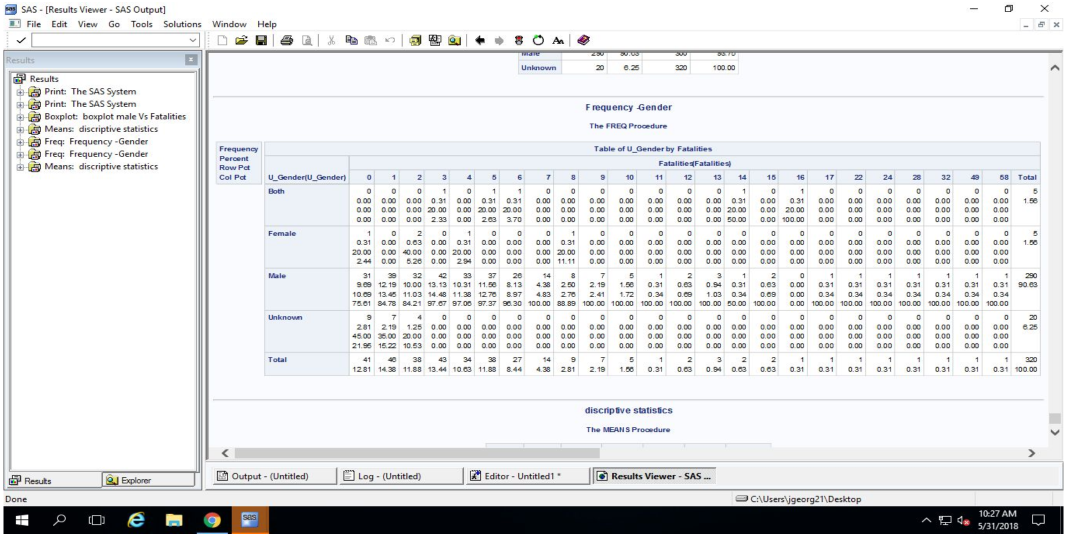
The MEANS Procedure

Variable	Label	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum
d_Male		0	1.0000000	1.0000000	1.0000000	1.0000000
Fatalities	Fatalities	0	1.0000000	3.0000000	5.5000000	58.0000000

Frequency -Gender

The FREQ Procedure

U_Gender				
U_Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Both	5	1.56	5	1.56
Female	5	1.56	10	3.13
Male	290	90.63	300	93.75
Unknown	20	6.25	320	100.00



Unknown	20	6.25	320	100.00

Frequency -Gender

The FREQ Procedure

Table of U_Gender by Fatalities

Frequency Percent Row Pct Col Pct	Fatalities(Fatalities)																								
U_Gender(U_Gender)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	22	24	28	32	49	58	Total
Both	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.31 20.00 2.33	0 0.00 0.00 0.00	1 0.31 20.00 2.63	1 0.31 20.00 3.70	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.31 20.00 50.00	0 0.00 0.00 0.00	1 0.31 20.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 1.56
Female	1 0.31 20.00 2.44	0 0.00 0.00 0.00	2 0.63 40.00 5.26	0 0.00 0.00 0.00	1 0.31 20.00 2.94	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.31 20.00 11.11	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 1.56
Male	31 9.69 10.69 75.61	39 12.19 13.45 84.78	32 10.00 11.03 84.21	42 13.13 14.48 97.67	33 10.31 11.38 97.06	37 11.56 12.76 97.37	26 8.13 8.97 96.30	14 4.38 4.83 100.00	8 2.50 2.76 88.89	7 2.19 2.41 100.00	5 1.56 1.72 100.00	1 0.31 0.34 100.00	2 0.63 0.69 100.00	3 0.94 1.03 100.00	1 0.31 0.34 50.00	2 0.63 0.69 100.00	0 0.00 0.00 0.00	1 0.31 0.34 100.00	1 0.31 0.34 100.00	1 0.31 0.34 100.00	1 0.31 0.34 100.00	1 0.31 0.34 100.00	1 0.31 0.34 100.00	1 0.31 0.34 100.00	290 90.63
Unknown	9 2.81 45.00 21.95	7 2.19 35.00 15.22	4 1.25 20.00 10.53	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	20 6.25
Total	41 12.81	46 14.38	38 11.88	43 13.44	34 10.63	38 11.88	27 8.44	14 4.38	9 2.81	7 2.19	5 1.56	1 0.31	2 0.63	3 0.94	2 0.63	2 0.63	1 0.31	1 0.31	1 0.31	1 0.31	1 0.31	1 0.31	1 0.31	1 0.31	320 100.00

discriptive statistics

The MEANS Procedure

Boxplots:

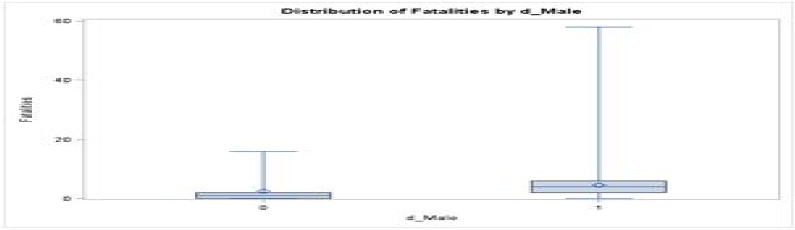
These boxplots indicates the distribution of male's vs injuries, male's vs fatalities and males vs total victims is significantly right skewed. The minimum number of fatalities that occurred due to a male shooter is 0 and the maximum number is approximately 58. The median fatalities that occurred due to a male shooter is 5. The mean for the male vs fatalities distribution is approximately 10. Since the distribution is significantly right skewed the median would be used to describe the spread of the data. The inter-quartile range would be used to describe the variability seen in the distribution.

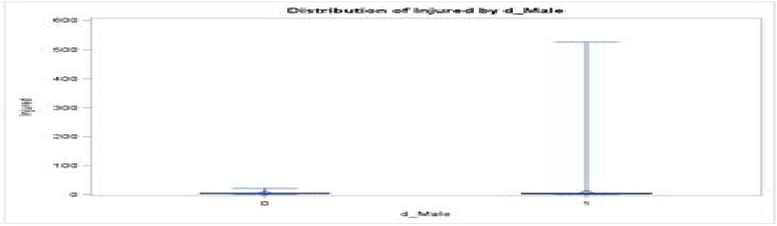
The minimum number of fatalities and people injured that occurred due to a male shooter is 0 whereas the minimum number of total victims occurred due to a male shooter is 1. The maximum number of injured people is approximately 550 by a male shooter. The maximum number of total victims is approximately 580.

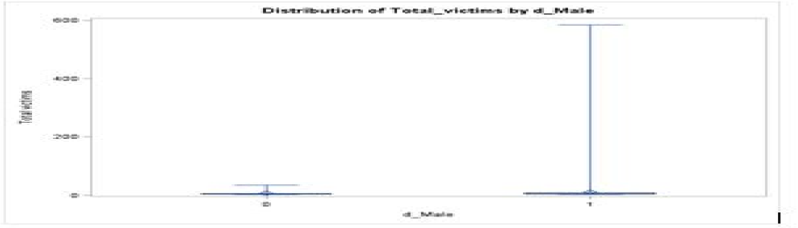
```
proc sort;
by d_Male;
run;
proc boxplot;
title"boxplot male Vs Fatalities";
plot Fatalities*d_Male;
run;

proc sort;
by d_Male;
run;
proc boxplot;
title"d_male VS injuried";
plot Injured*d_Male;
run;

proc sort;
by d_Male;
run;
proc boxplot;
title "boxplot totalvictims Vs males";
plot Total_victims*d_Male;
run;
```







Interaction variables:

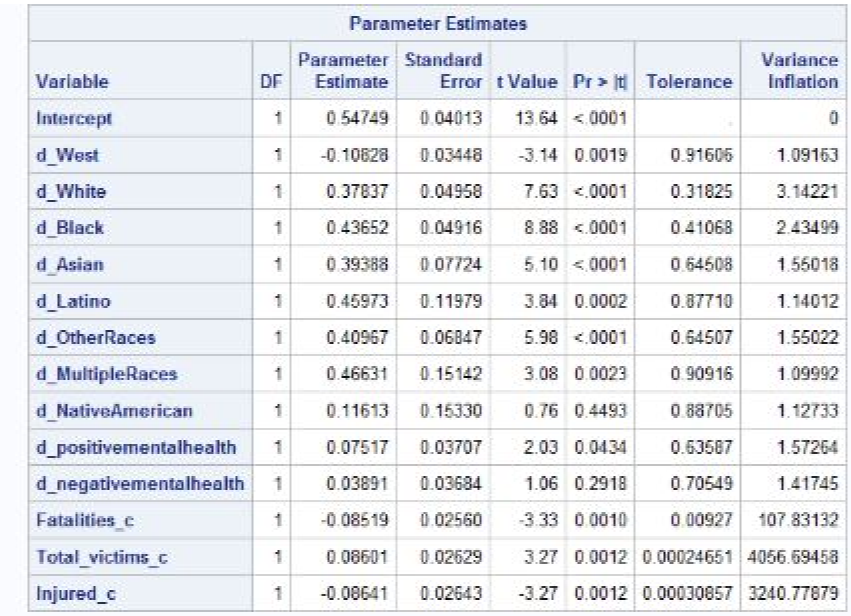
(This varies per model - rest included in Analysis section)

Each group member selected a specific region, different race and checked for positive mental health amongst males during this analysis.

```
data interactionterms;
set shootingproj;
Regionpostivemeantal health status _race=(d_region*d_positivementalhealth*d_race);
Positivementalhealthstatus_race=(d_positivementalhealth*d_race);
Southwest_white=(d_Southwest*d_black);
Positivemeantalhealthstatus_Fatalities=(d_positivementalhealth*Fatalities);
postivementalhealthstatus_race_fatalities=(d_positivementalhealth*d_race*fatalities);
run;
```

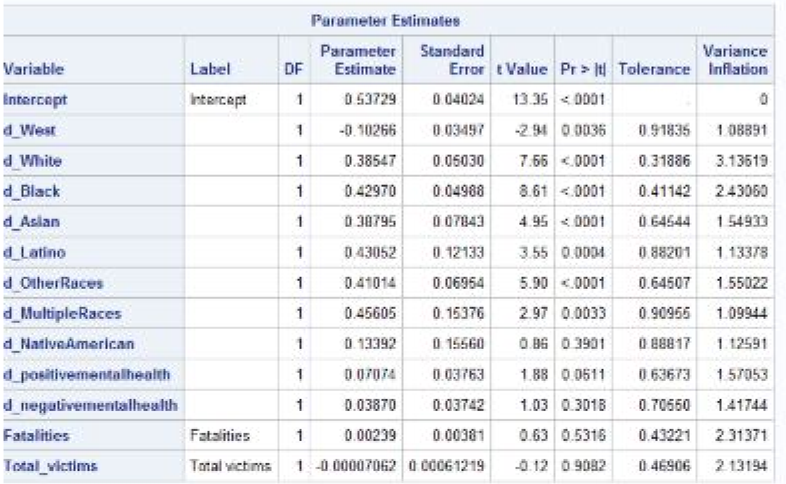
Multi Collinearity:

During our initial analysis it was found that quantitative variables – fatalities, injured and total number of victims had high multicollinearity between them. Since these variables were critical for a better analysis it was decided to use the centering technique. Although the centering technique was used, multicollinearity still existed between these variables. hence, it was decided to take the quantitative variable – injured, out of the model. Thus this helped to remove the multi collinearity.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	0.54749	0.04013	13.64	<.0001	.	0
d_West	1	-0.10628	0.03448	-3.14	0.0019	0.91606	1.09163
d_White	1	0.37837	0.04958	7.63	<.0001	0.31825	3.14221
d_Black	1	0.43652	0.04916	8.88	<.0001	0.41068	2.43499
d_Asian	1	0.39388	0.07724	5.10	<.0001	0.64508	1.55018
d_Latino	1	0.45973	0.11979	3.84	0.0002	0.87710	1.14012
d_OtherRaces	1	0.40967	0.06847	5.98	<.0001	0.64507	1.55022
d_MultipleRaces	1	0.46631	0.15142	3.08	0.0023	0.90916	1.09992
d_NativeAmerican	1	0.11613	0.15330	0.76	0.4493	0.88705	1.12733
d_positivementalhealth	1	0.07517	0.03707	2.03	0.0434	0.63587	1.57264
d_negativementalhealth	1	0.03891	0.03684	1.06	0.2918	0.70549	1.41745
Fatalities_c	1	-0.08519	0.02560	-3.33	0.0010	0.00927	107.83132
Total_victims_c	1	0.08601	0.02629	3.27	0.0012	0.00024651	4056.69458
Injured_c	1	-0.08641	0.02643	-3.27	0.0012	0.00030857	3240.77879



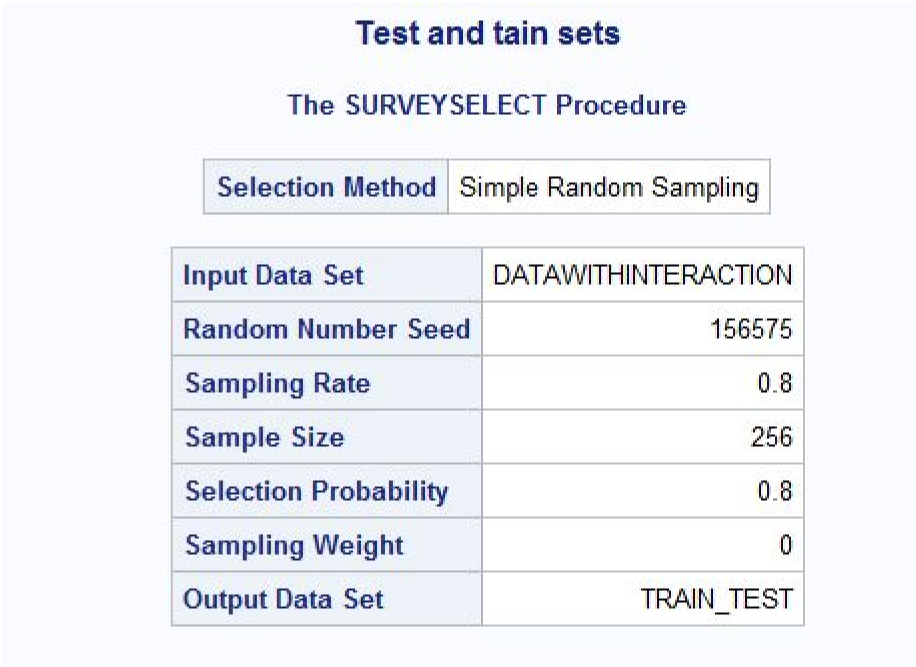
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.53729	0.04024	13.35	<.0001	.	0
d_West		1	-0.10266	0.03497	-2.94	0.0036	0.91835	1.08891
d_White		1	0.38547	0.05030	7.66	<.0001	0.31886	3.13619
d_Black		1	0.42970	0.04988	8.61	<.0001	0.41142	2.43060
d_Asian		1	0.38795	0.07843	4.95	<.0001	0.64544	1.54933
d_Latino		1	0.43052	0.12133	3.55	0.0004	0.88201	1.13378
d_OtherRaces		1	0.41014	0.06954	5.90	<.0001	0.64507	1.55022
d_MultipleRaces		1	0.45605	0.15376	2.97	0.0033	0.90955	1.09944
d_NativeAmerican		1	0.13392	0.15560	0.86	0.3901	0.88817	1.12591
d_positivementalhealth		1	0.07074	0.03763	1.88	0.0611	0.63673	1.57053
d_negativementalhealth		1	0.03870	0.03742	1.03	0.3018	0.70560	1.41744
Fatalities	Fatalities	1	0.00239	0.00381	0.63	0.5316	0.43221	2.31371
Total_victims	Total victims	1	-0.00007062	0.00061219	-0.12	0.9082	0.46906	2.13194

Model 1 – Jimi George

Analysis

The entire data set was divided into two set for training and testing. 80% of the data was used for model training and the remaining 20% was used model testing and prediction. After the split , the training set has 256 observation(see image 1). Since the response variable is binary , logistic regression was utilized to both fit and train the model. Initially a full model logistic regression was ran and then both forward and backward model selection techniques were utilized to fit the model accurately. The comparative analysis for two models fitted by forward and backward technique is given below .



Test and tain sets

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

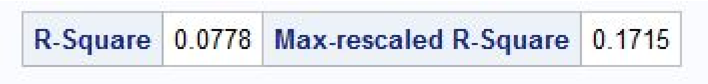
Input Data Set	DATAWITHINTERACTION
Random Number Seed	156575
Sampling Rate	0.8
Sample Size	256
Selection Probability	0.8
Sampling Weight	0
Output Data Set	TRAIN_TEST

Image **1**

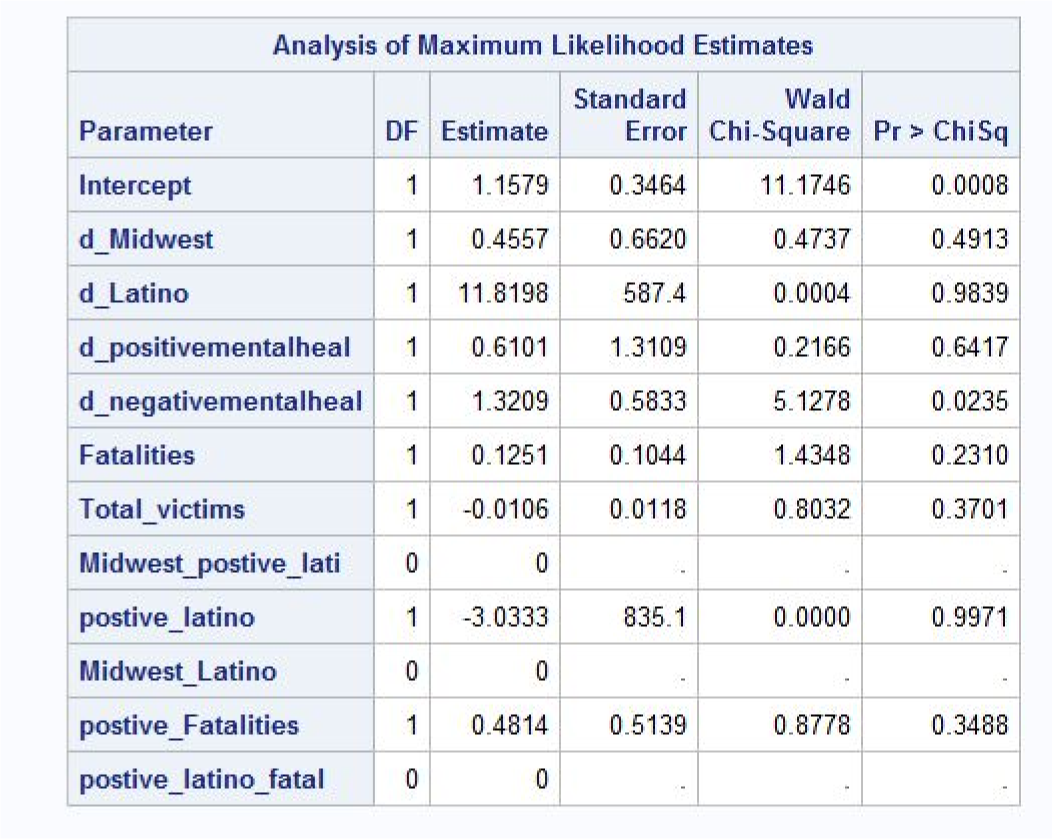
Full model

LogP(d_males =1)/1-P(d_males = 0) = 1.1579+ 0.4557 d_Midwest + 11.8191 d_Latino +0.6101positivementalhealth+1.3209d_negativementalhealth+0.1251fatalities-0.0106total_victims -3.0333positive_latinos+0.4814positive_fatalities+e

Where d_Latino = 1 when u_Race = Latino; otherwise = 0
Where d_Midwest= 1 when Region = midwest, otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= “Yes”, otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0



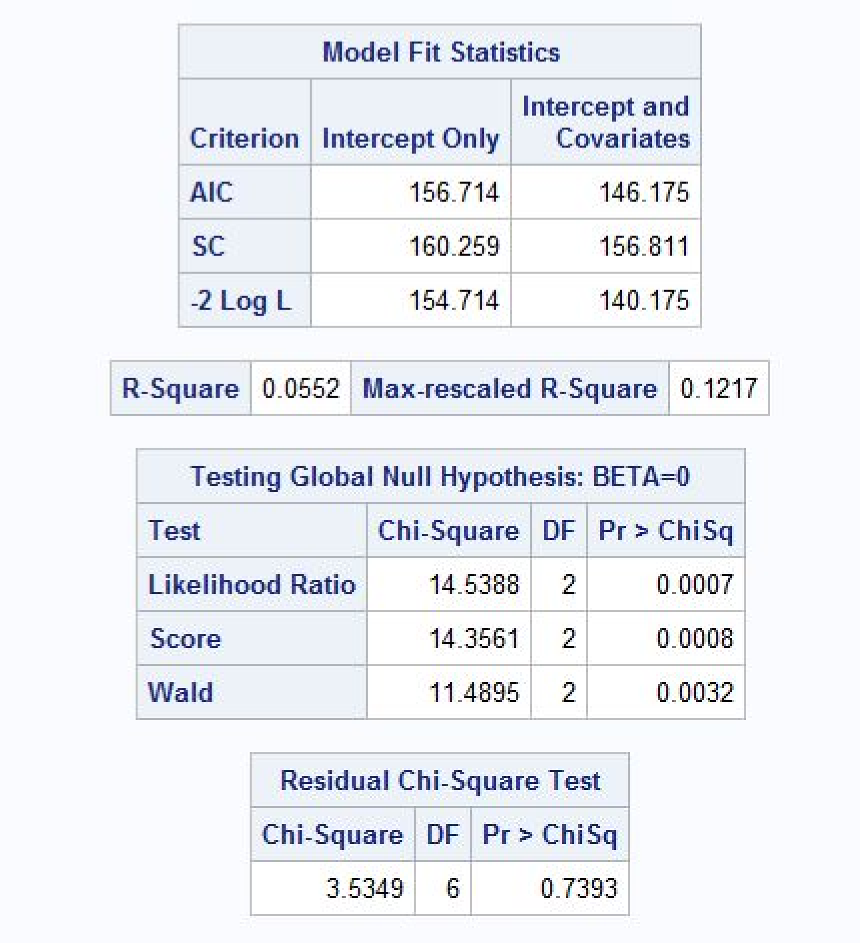
R-Square	0.0778	Max-rescaled R-Square	0.1715



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1579	0.3464	11.1746	0.0008
d_Midwest	1	0.4557	0.6620	0.4737	0.4913
d_Latino	1	11.8198	587.4	0.0004	0.9839
d_positivementalheal	1	0.6101	1.3109	0.2166	0.6417
d_negativementalheal	1	1.3209	0.5833	5.1278	0.0235
Fatalities	1	0.1251	0.1044	1.4348	0.2310
Total_victims	1	-0.0106	0.0118	0.8032	0.3701
Midwest_postive_lati	0	0	.	.	.
postive_latino	1	-3.0333	835.1	0.0000	0.9971
Midwest_Latino	0	0	.	.	.
postive_Fatalities	1	0.4814	0.5139	0.8778	0.3488
postive_latino_fatal	0	0	.	.	.

Image2

Forward selection



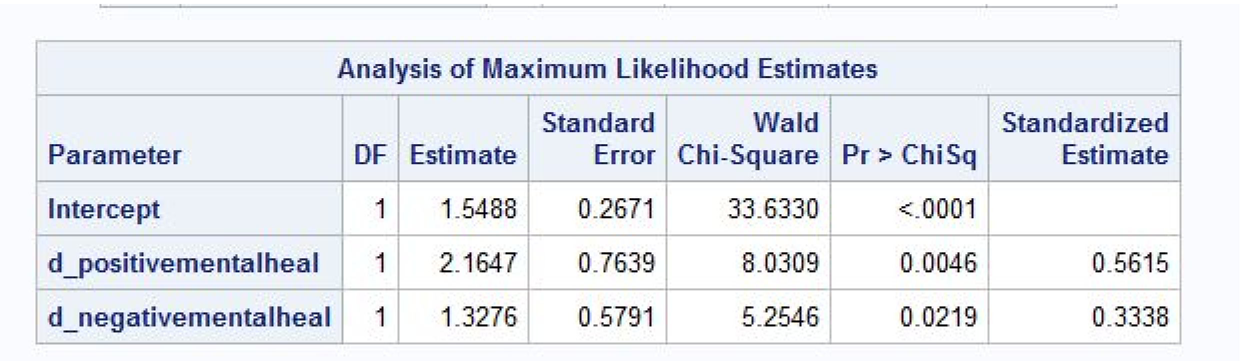
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	156.714	146.175
SC	160.259	156.811
-2 Log L	154.714	140.175

R-Square	0.0552	Max-rescaled R-Square	0.1217

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.5388	2	0.0007
Score	14.3561	2	0.0008
Wald	11.4895	2	0.0032

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
3.5349	6	0.7393

Image 3

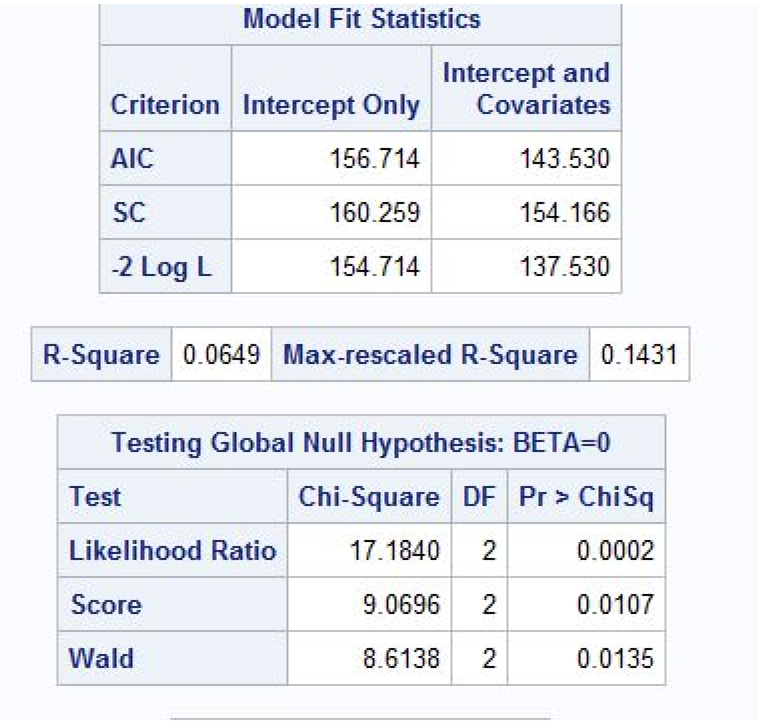


Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.5488	0.2671	33.6330	<.0001	
d_positivementalheal	1	2.1647	0.7639	8.0309	0.0046	0.5615
d_negativementalheal	1	1.3276	0.5791	5.2546	0.0219	0.3338

Model equation per forward selection

LogP(d_males =1)/1-P(d_males = 0) = 1.5488+
+2.1647 d_positivementalhealth+1.3276 d_negativementalhealth+ e

Backward Selection

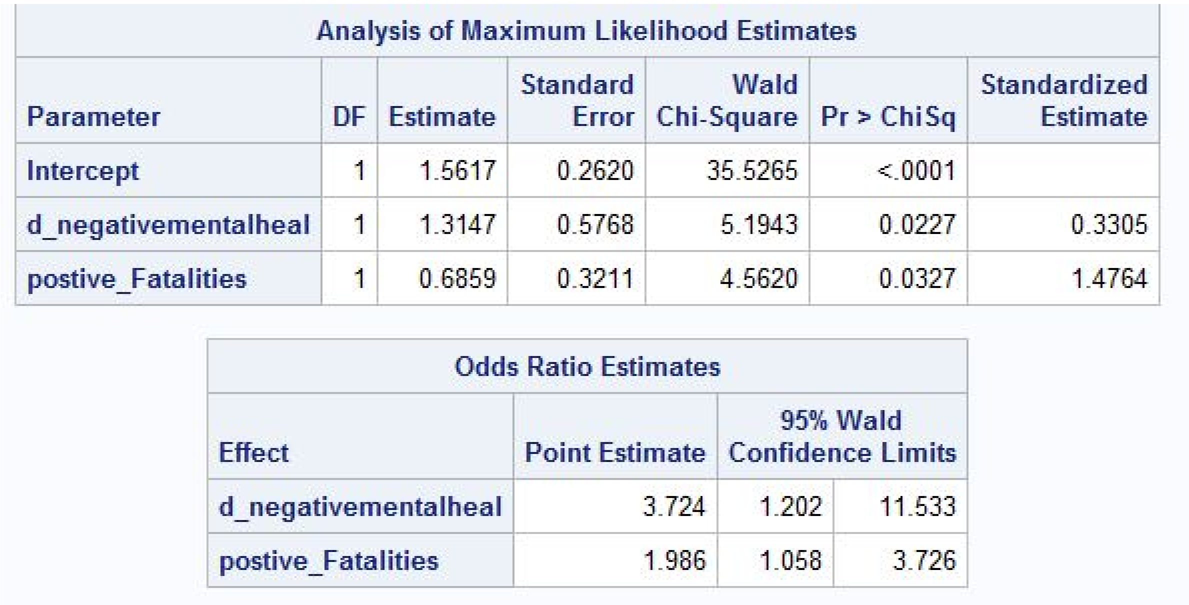


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	156.714	143.530
SC	160.259	154.166
-2 Log L	154.714	137.530

R-Square	0.0649	Max-rescaled R-Square	0.1431

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.1840	2	0.0002
Score	9.0696	2	0.0107
Wald	8.6138	2	0.0135

Image4



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.5617	0.2620	35.5265	<.0001	
d_negativementalheal	1	1.3147	0.5768	5.1943	0.0227	0.3305
postive_Fatalities	1	0.6859	0.3211	4.5620	0.0327	1.4764

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
d_negativementalheal	3.724	1.202	11.533
postive_Fatalities	1.986	1.058	3.726

Model equation per backward selection
LogP(d_males =1)/1-P(d_males = 0) = 1.5617 + 1.3147 d_negativementalhealth+ 0.6859positive_fatalities+e

Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

Selected model :

LogP(d_males =1)/1-P(d_males = 0) = 1.5617 + 1.3147 d_negativementalhealth+ 0.6859positive_fatalities+e

Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

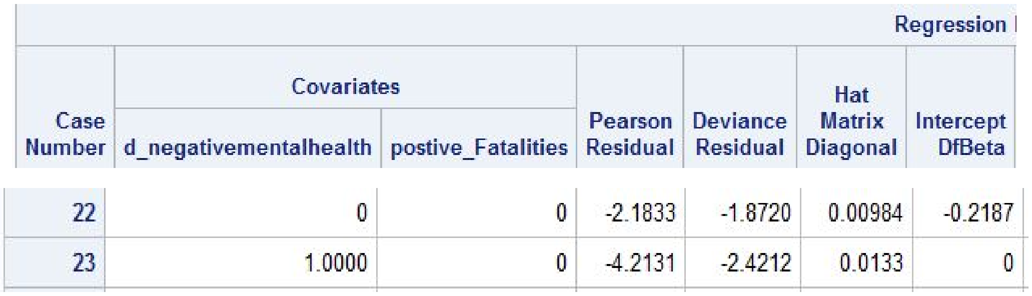
The model fitted by the backward model technique was selected as the model had better diagnostics. The backward model had a better R2=0.0649 compared to forward model which has an R2=0.0552. An R2= 0.0649 indicates that 6.49% of the variability seen in the data is explained by the model. Both the models had the same AIC and SC values at 156.714 and 160.259 respectively . Both AIC and SC are error terms and the model indicates a relatively low error terms. The pr-value for predictors in the selected model is below the 0.05 (Alpha) indicating that the predictors are significant and be included in the model . Finally the model also meets the goodness of fit test

$H_o\beta_j$=0 , the predictors, negative mental health status , and the interaction term :positive mental health status with fatalities has no significant relationship to the response variable , male gender.

$H_a\beta_j\neq0$, at least one of the predictors, negative mental health status ,and the interaction term :positive mental health status with fatalities , has a significant relationship to the response variable, male gender.

The likelihood Ratio for the model =17.1840 with a p-value of 0.002, which is significantly lower than the alpha =0.05 against which it is tested. This indicates that the null hypothesis indicating that the predictors have no significant relationship to the independent variable can be rejected Thus , the alternative hypothesis that at least one of the predictors has significant effect on the response variable can be accepted.

When analyzing the standardized residuals , the predictor interaction term:positive mental health status with number of fatalities. has most influence on the response variable , followed by ,negative mental health status



Regression I						
	Covariates					
Case Number	d_negativementalhealth	postive_Fatalities	Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta
22	0	0	-2.1833	-1.8720	0.00984	-0.2187
23	1.0000	0	-4.2131	-2.4212	0.0133	0

Image 5

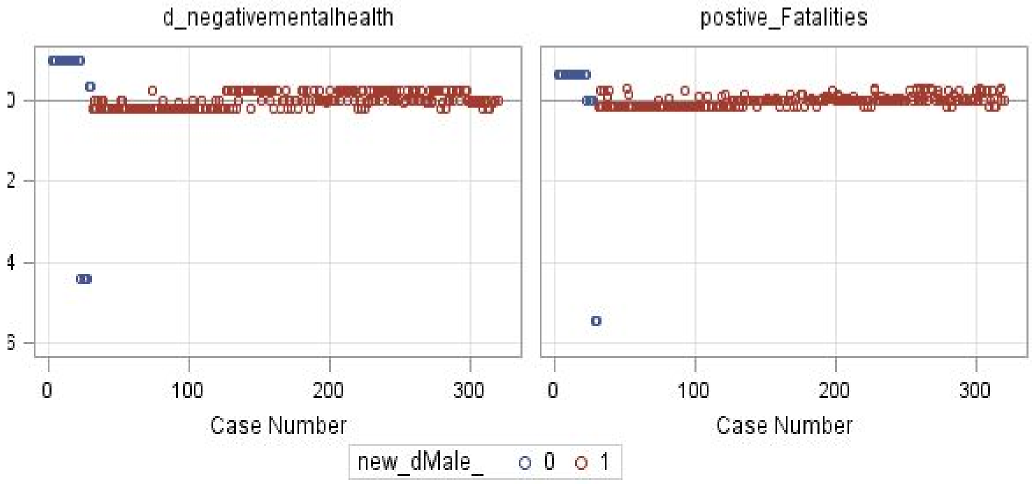
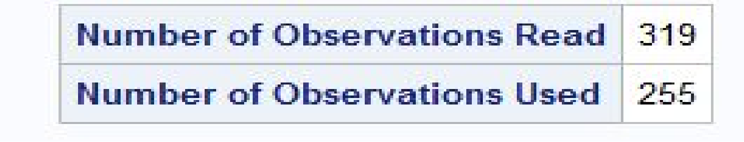


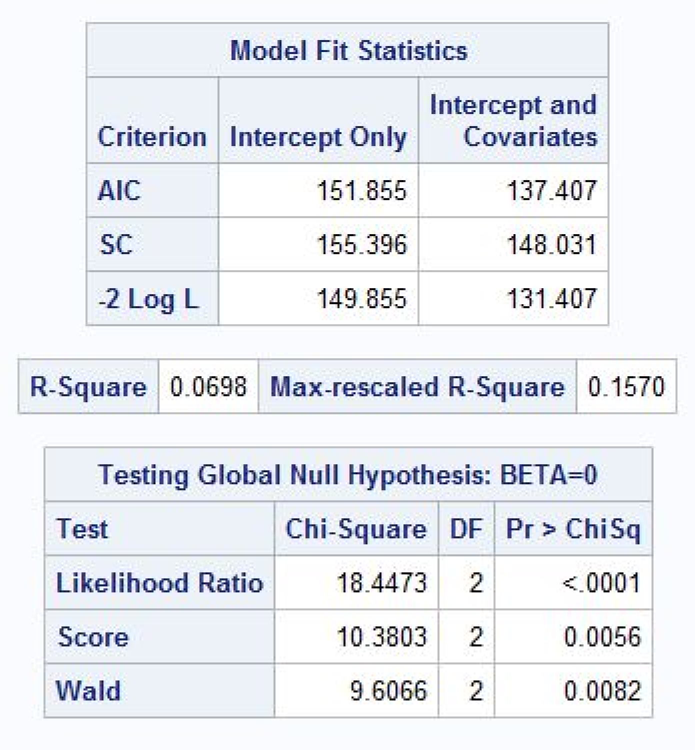
Image 5

The selected model was also checked for outliers and influential points . Values given under pearson residual and deviance residual that were above +3 and -3 were marked as outliers Similarly the Dfbetas graph was analysed for influential points . |Dfbetas| > 2/√n was the criteria used to narrow down the influential points. According to the formula Dfbetas with a value more than 0.11 will be marked as influential point. Though comparing observation against these criterias , observation 23 was removed from the model. The model was then refitted again to check for changes.

Final model



Number of Observations Read	319
Number of Observations Used	255

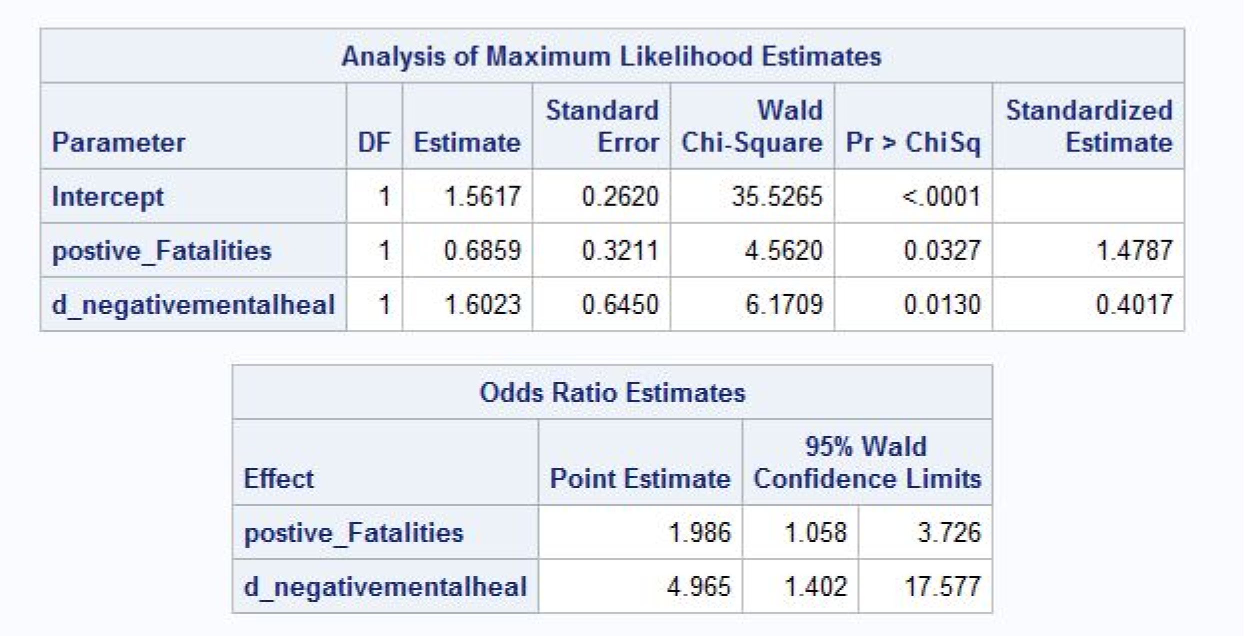


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	151.855	137.407
SC	155.396	148.031
-2 Log L	149.855	131.407

R-Square	0.0698	Max-rescaled R-Square	0.1570

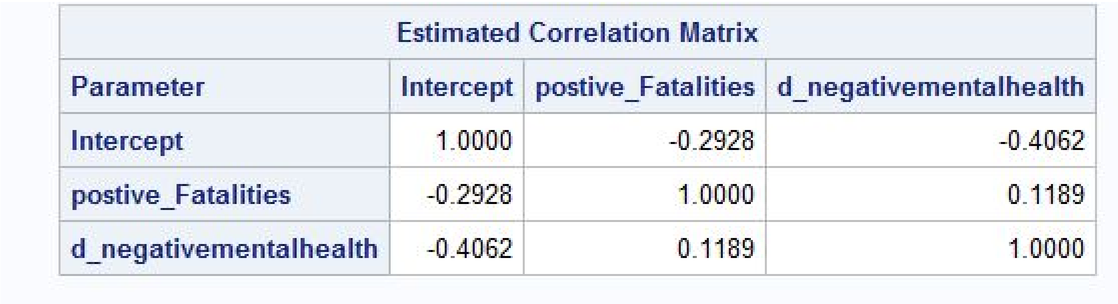
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.4473	2	<.0001
Score	10.3803	2	0.0056
Wald	9.6066	2	0.0082

Image6



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.5617	0.2620	35.5265	<.0001	
postive_Fatalities	1	0.6859	0.3211	4.5620	0.0327	1.4787
d_negativementalheal	1	1.6023	0.6450	6.1709	0.0130	0.4017

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
postive_Fatalities	1.986	1.058	3.726
d_negativementalheal	4.965	1.402	17.577



Estimated Correlation Matrix			
Parameter	Intercept	postive_Fatalities	d_negativementalhealth
Intercept	1.0000	-0.2928	-0.4062
postive_Fatalities	-0.2928	1.0000	0.1189
d_negativementalhealth	-0.4062	0.1189	1.0000

Final Fitted model

LogP(d_males =1)/1-P(d_males = 0) = 1.5617 + 0.6859positive_fatalities+1.6023 d_negativementalhealth++e

Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

The R2 for the final fitted model is 0.0698 , indicating that 6.98% of the variability is explained by the model .No other outliers were removed as the R2 had not increased significantly with the removal of the first outlier All the predictors in the model remain significant as their pr-value is less than alpha 0.05. Finally , the correlation table , indicated that there is no incidence multicollinearity between the variables.

$H_o\beta_j$=0 , the predictors, negative mental health status , and the interaction term :positive mental health status with fatalities has no significant relationship to the response variable , male gender.

$H_a\beta_j\neq$0, at least one of the predictors, negative mental health status ,and the interaction term :positive mental health status with fatalities , has a significant relationship to the response variable, male gender.

The likelihood Ratio for the model =18.4473 with a p-value of less than 0.0001, which is significantly lower than the alpha =0.05 against which it is tested. Hence the null hypothesis can be rejected and the alternative hypothesis can be accepted.

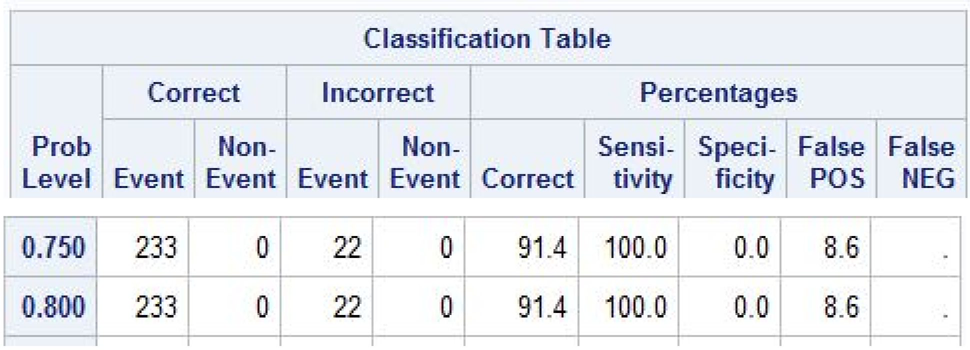
The AIC and SC are relatively high indicating the error in the model is large ,and the R2 is relatively low indicating that a large part of the variability seen is not explained by the model.

Odd Ratio

Positive_fatalities = For shooters with no mental health issues , any new incidence of positive mental health problem and fatalities increases the average odd for the shooter to be male by 629% [(ex(1.986)-1)*100] with a 95% confidence interval that the average increase will be between 188%[(ex(1.058)-1)*100] and 4051%[(ex(3.726)-1)*100] .

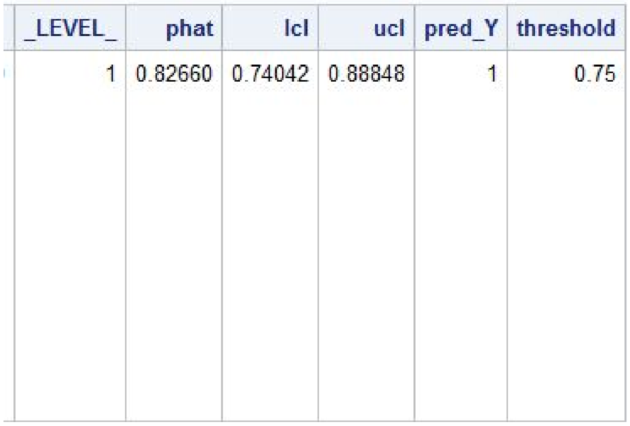
D_negative mental health= for shooter with no incidence of positive mental health problem and fatalities, any new incidence of no mental health problem increases the average odd for the shooter to be a male by 14230% [(ex(4.965)-1)*100] with a 95% confidence interval that the average will be between 306% [(ex(1.402)-1)*100] and 4.401319253*10^9 % [(ex(17.6)-1)*100]

Prediction

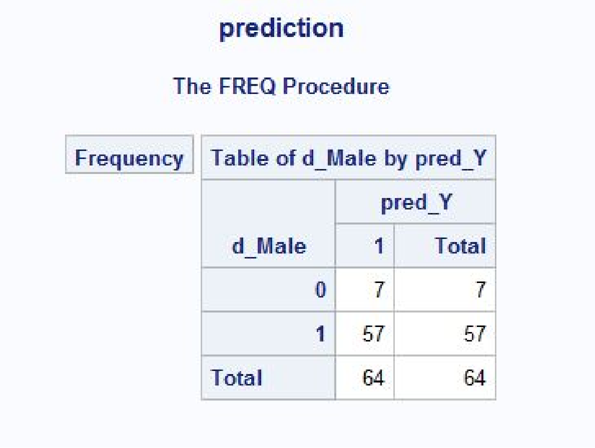


Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.750	233	0	22	0	91.4	100.0	0.0	8.6	.
0.800	233	0	22	0	91.4	100.0	0.0	8.6	.

Image 7



LEVEL	phat	lcl	ucl	pred_Y	threshold
1	0.82660	0.74042	0.88848	1	0.75



prediction

The FREQ Procedure

Frequency	Table of d_Male by pred_Y	
	pred_Y	
d_Male	1	Total
0	7	7
1	57	57
Total	64	64

Using the final fitted model , predicted probability was computed for the testing model . A threshold of 0.75 was used to then computed the predicted Y. The predicted Y would equal to 1 if the predicted probability was greater than 0.75. Similarity , the predicted Y would equal to 0 if the predicted probability was less than or equal to 0.75.Performance matrix was then computed as follow

Precision = TP/(TP+FP)

=57/(64)

=0.8906 *100=89.06%

This indicates that the model is able to predict the incidence of male shooter with 89.06% precision.

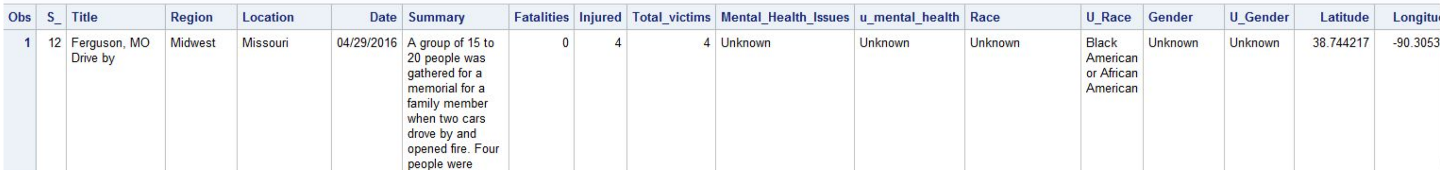
Mode 2 – Kubra Iqbal

Final report analysis.

Introduction:

In Model 1 - “Male” was the dependent variable while “region” was the independent variable. The specific region picked for Model 1 was southeast. Logistic Regression was used to carry out the analysis for Model 1.

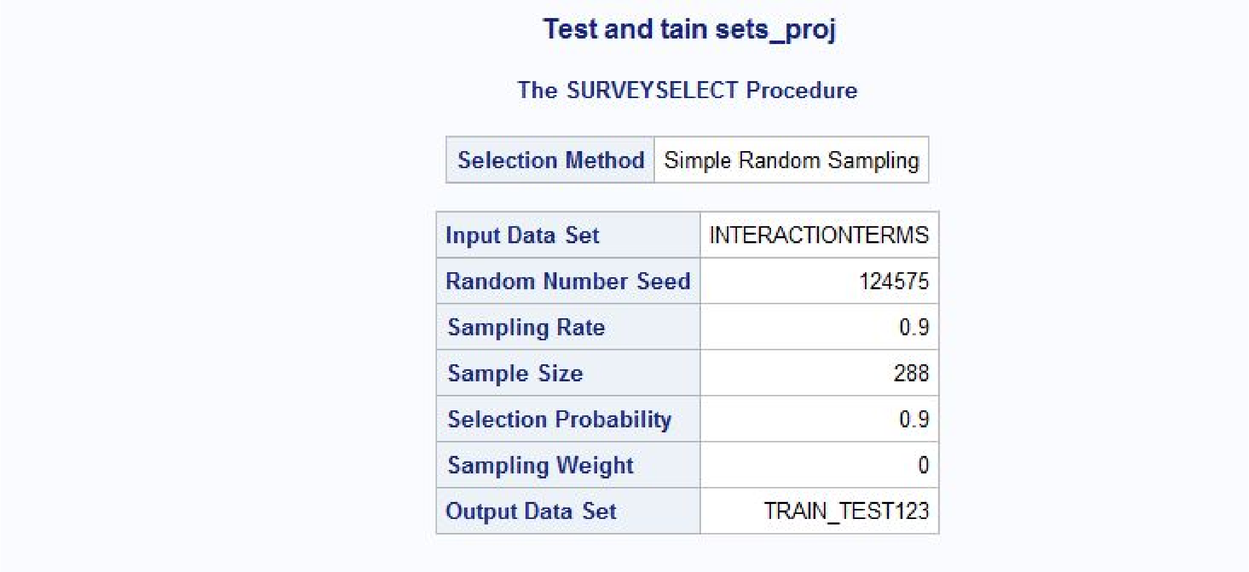
The image below shows the all the variables that were used for Model 1.



Obs	S_	Title	Region	Location	Date	Summary	Fatalities	Injured	Total_victims	Mental_Health_Issues	u_mental_health	Race	U_Race	Gender	U_Gender	Latitude	Longitu
1	12	Ferguson, MO Drive by	Midwest	Missouri	04/29/2016	A group of 15 to 20 people was gathered for a memorial for a family member when two cars drove by and opened fire. Four people were	0	4	4	Unknown	Unknown	Unknown	Black American or African American	Unknown	Unknown	38.744217	-90.3053

Train and Test:

A random sampling of 90-10 was used in Model 1. The training set is made of 90% of data selected through random seed. The final model is determined through this data. The left-over 10% of data is selected for Test Validation purposes and will be further used to check the model accuracy of the data.

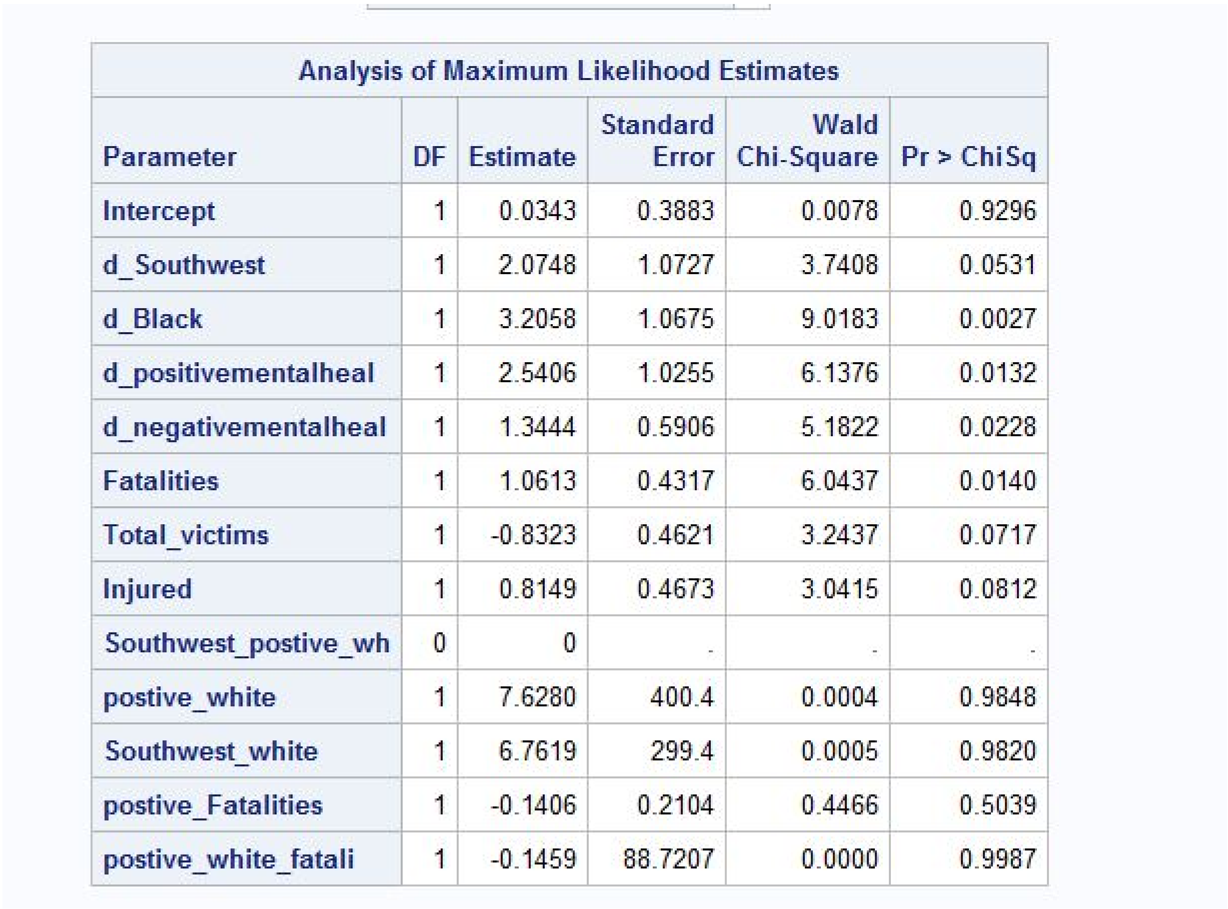


Test and tain sets_proj

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	INTERACTIONTERMS
Random Number Seed	124575
Sampling Rate	0.9
Sample Size	288
Selection Probability	0.9
Sampling Weight	0
Output Data Set	TRAIN_TEST123



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0343	0.3883	0.0078	0.9296
d_Southwest	1	2.0748	1.0727	3.7408	0.0531
d_Black	1	3.2058	1.0675	9.0183	0.0027
d_positivementalheal	1	2.5406	1.0255	6.1376	0.0132
d_negativementalheal	1	1.3444	0.5906	5.1822	0.0228
Fatalities	1	1.0613	0.4317	6.0437	0.0140
Total_victims	1	-0.8323	0.4621	3.2437	0.0717
Injured	1	0.8149	0.4673	3.0415	0.0812
Southwest_postive_wh	0	0	.	.	.
postive_white	1	7.6280	400.4	0.0004	0.9848
Southwest_white	1	6.7619	299.4	0.0005	0.9820
postive_Fatalities	1	-0.1406	0.2104	0.4466	0.5039
postive_white_fatali	1	-0.1459	88.7207	0.0000	0.9987

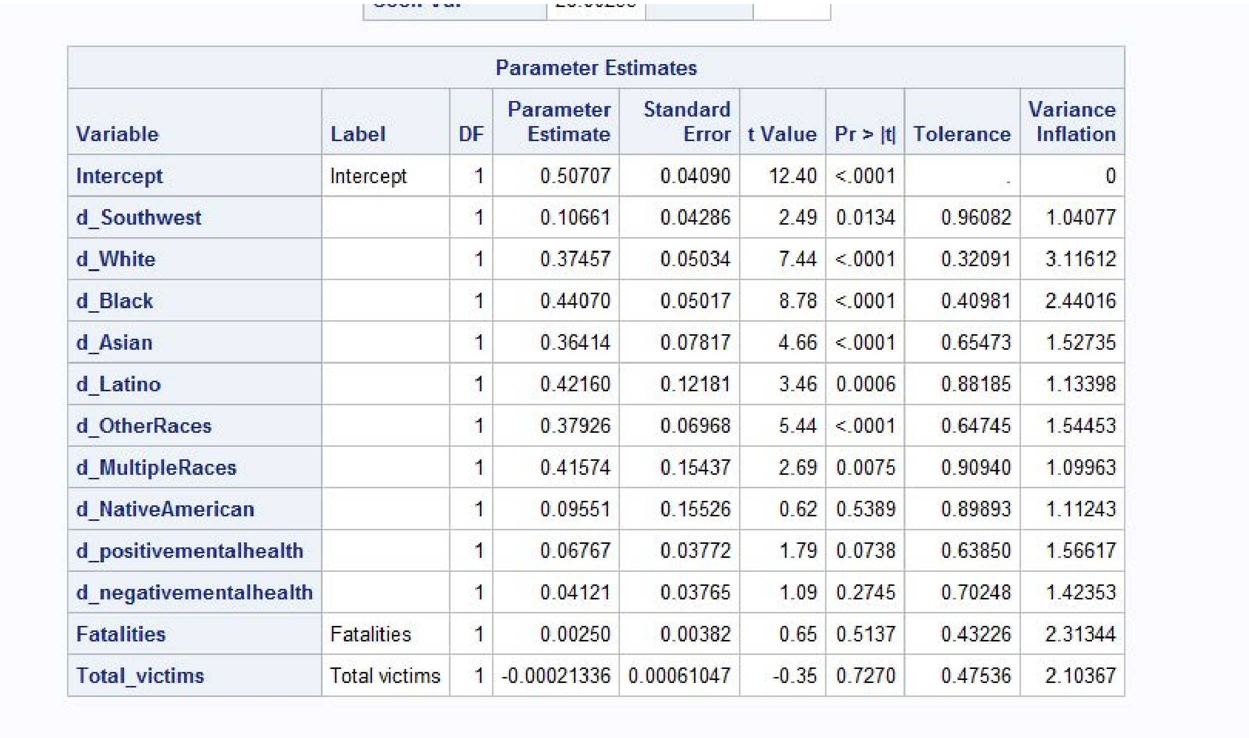
Full model

LogP(d_males =1)/1-P(d_males = 0) = 0.0343 + 2.0748 d_southwest + 3.2058 d_black + 2.5406 d_positivemeantalhealth + 1.344 d_negativemeantalheal + 1.0613 fatalities -.8323 total_victims + 0.8149 injured + 7.6280 positive_White + 6.7619 southwest_white -0.1406 positive_fatalities -0.1459 positive_white_fatalities

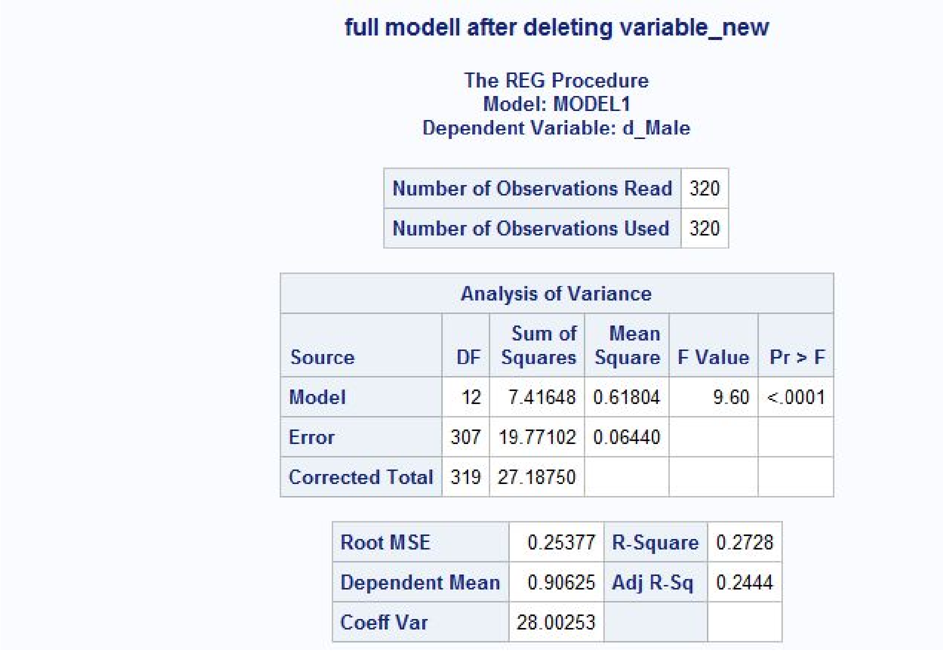
Where d_black = 1 when u_Race = Latino; otherwise = 0
Where d_southwest= 1 when Region = midwest, otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= “Yes”, otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

Multicollinearity:

The output is as below for Multicollinearity:



Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.50707	0.04090	12.40	<.0001	.	0
d_Southwest		1	0.10661	0.04286	2.49	0.0134	0.96082	1.04077
d_White		1	0.37457	0.05034	7.44	<.0001	0.32091	3.11612
d_Black		1	0.44070	0.05017	8.78	<.0001	0.40981	2.44016
d_Asian		1	0.36414	0.07817	4.66	<.0001	0.65473	1.52735
d_Latino		1	0.42160	0.12181	3.46	0.0006	0.88185	1.13398
d_OtherRaces		1	0.37926	0.06968	5.44	<.0001	0.64745	1.54453
d_MultipleRaces		1	0.41574	0.15437	2.69	0.0075	0.90940	1.09963
d_NativeAmerican		1	0.09551	0.15526	0.62	0.5389	0.89893	1.11243
d_positivementalhealth		1	0.06767	0.03772	1.79	0.0738	0.63850	1.56617
d_negativementalhealth		1	0.04121	0.03765	1.09	0.2745	0.70248	1.42353
Fatalities	Fatalities	1	0.00250	0.00382	0.65	0.5137	0.43226	2.31344
Total_victims	Total victims	1	-0.00021336	0.00061047	-0.35	0.7270	0.47536	2.10367



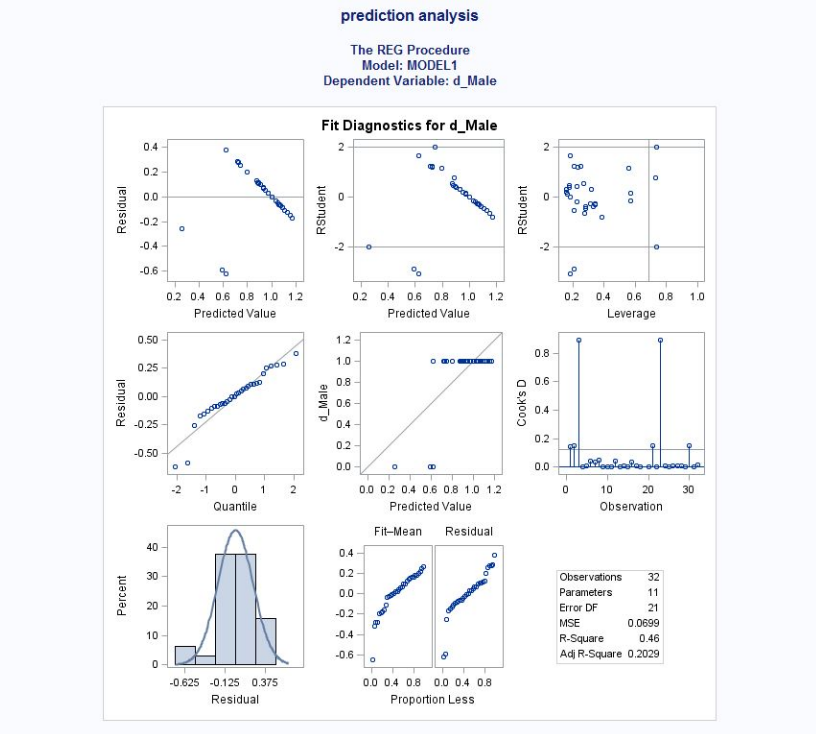
full modell after deleting variable_new

The REG Procedure
Model: MODEL1
Dependent Variable: d_Male

Number of Observations Read	320
Number of Observations Used	320

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	7.41648	0.61804	9.60	<.0001
Error	307	19.77102	0.06440		
Corrected Total	319	27.18750			

Root MSE	0.25377	R-Square	0.2728
Dependent Mean	0.90625	Adj R-Sq	0.2444
Coeff Var	28.00253		



prediction analysis

The REG Procedure
Model: MODEL1
Dependent Variable: d_Male

Fit Diagnostics for d_Male

Observations	32
Parameters	11
Error DF	21
MSE	0.0699
R-Square	0.46
Adj R-Square	0.2029

If the Variation Inflation is greater than 10 – it means multicollinearity exists. According to this method there are no values above than 10 which means multicollinearity doesn't exists.

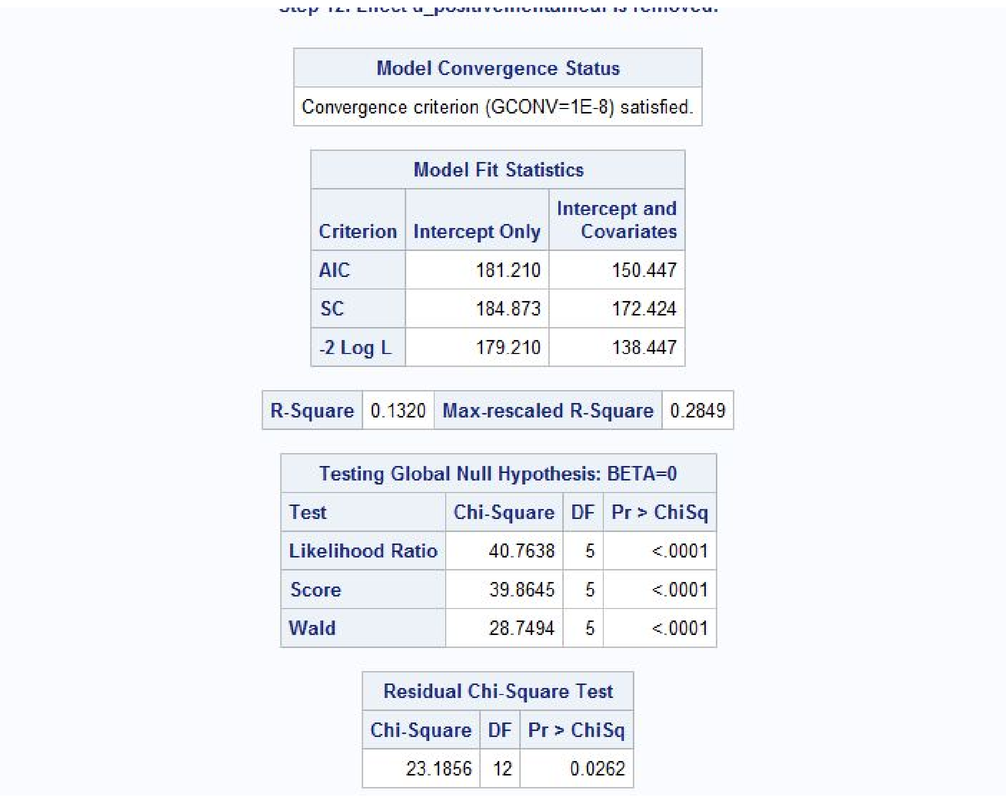
Regression Model Used:

- Logistic Regression

Selection Procedures used in Model:

- Backward method
- Forward Method

<u>Forward Selection Method:</u>



Step 12. Effect d_positivementalheal is removed.

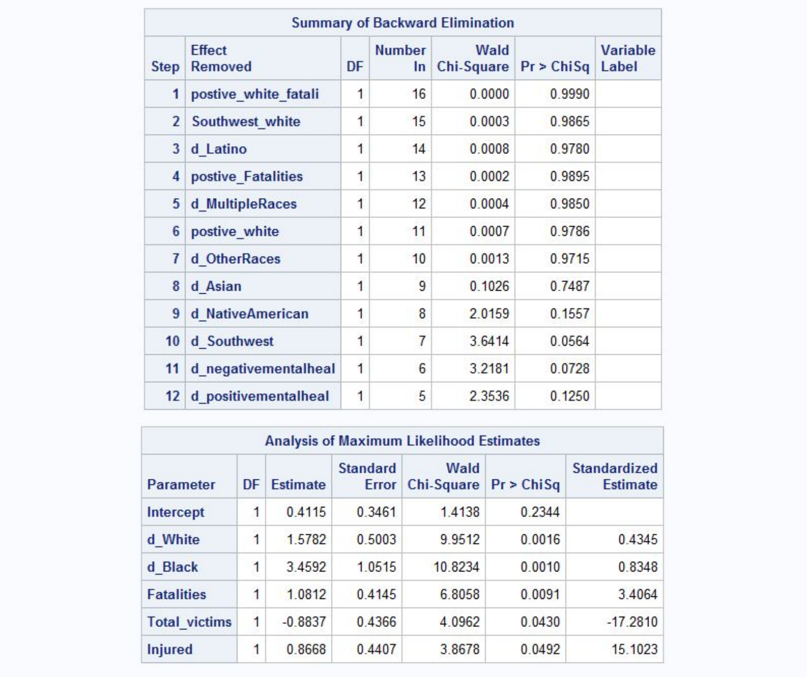
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	181.210	150.447
SC	184.873	172.424
-2 Log L	179.210	138.447

R-Square	0.1320	Max-rescaled R-Square	0.2849

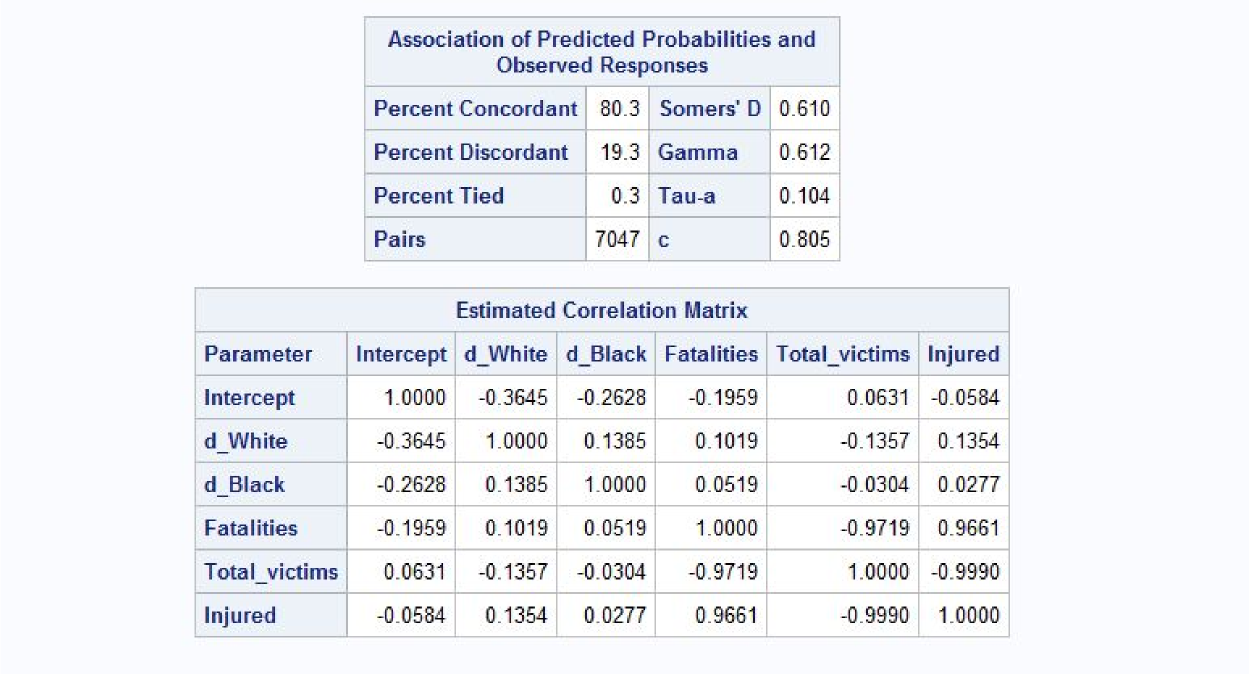
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.7638	5	<.0001
Score	39.8645	5	<.0001
Wald	28.7494	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
23.1856	12	0.0262



Summary of Backward Elimination						
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	postive_white_fatali	1	16	0.0000	0.9990	
2	Southwest_white	1	15	0.0003	0.9865	
3	d_Latino	1	14	0.0008	0.9780	
4	postive_Fatalities	1	13	0.0002	0.9895	
5	d_MultipleRaces	1	12	0.0004	0.9850	
6	postive_white	1	11	0.0007	0.9786	
7	d_OtherRaces	1	10	0.0013	0.9715	
8	d_Asian	1	9	0.1026	0.7487	
9	d_NativeAmerican	1	8	2.0159	0.1557	
10	d_Southwest	1	7	3.6414	0.0564	
11	d_negativementalheal	1	6	3.2181	0.0728	
12	d_positivementalheal	1	5	2.3536	0.1250	

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.4115	0.3461	1.4138	0.2344	
d_White	1	1.5782	0.5003	9.9512	0.0016	0.4345
d_Black	1	3.4592	1.0515	10.8234	0.0010	0.8348
Fatalities	1	1.0812	0.4145	6.8058	0.0091	3.4064
Total_victims	1	-0.8837	0.4366	4.0962	0.0430	-17.2810
Injured	1	0.8668	0.4407	3.8678	0.0492	15.1023



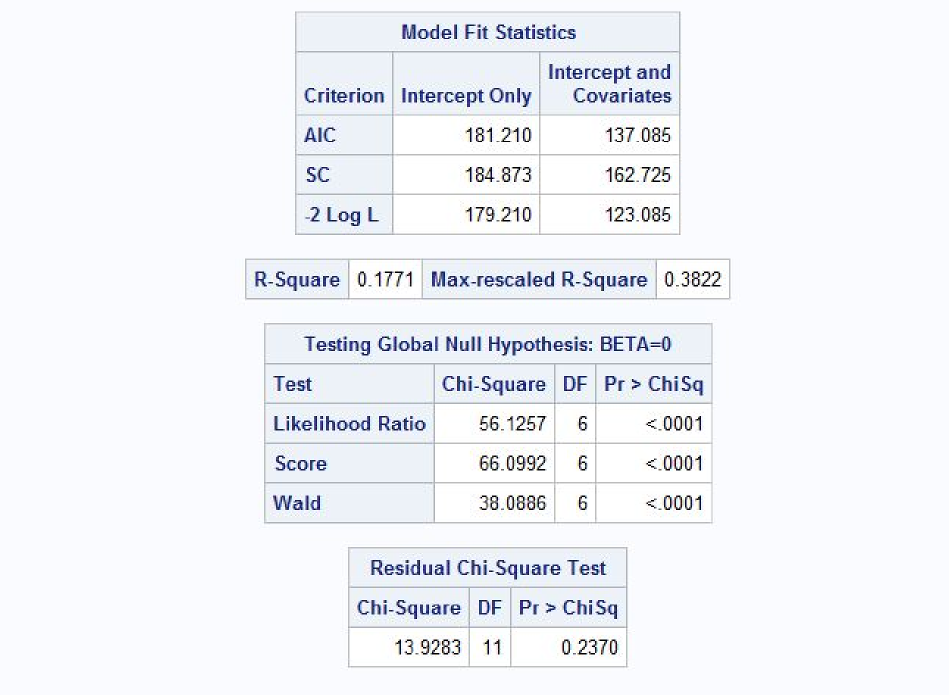
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.3	Somers' D	0.610
Percent Discordant	19.3	Gamma	0.612
Percent Tied	0.3	Tau-a	0.104
Pairs	7047	c	0.805

Estimated Correlation Matrix

Parameter	Intercept	d_White	d_Black	Fatalities	Total_victims	Injured
Intercept	1.0000	-0.3645	-0.2628	-0.1959	0.0631	-0.0584
d_White	-0.3645	1.0000	0.1385	0.1019	-0.1357	0.1354
d_Black	-0.2628	0.1385	1.0000	0.0519	-0.0304	0.0277
Fatalities	-0.1959	0.1019	0.0519	1.0000	-0.9719	0.9661
Total_victims	0.0631	-0.1357	-0.0304	-0.9719	1.0000	-0.9990
Injured	-0.0584	0.1354	0.0277	0.9661	-0.9990	1.0000

The Backward method shows that there are 6 variables selected in this section procedure.
R2= 0.1320 i.e. 13% approximately of data variation will be explained by this model.
Log(P(1-P) = 1 -0.3645 d_white -0.2628 d_black -0.1959 fatalaities + 0.0631 total_victims -0.0584 injured

<u>Backward Selection Method:</u>

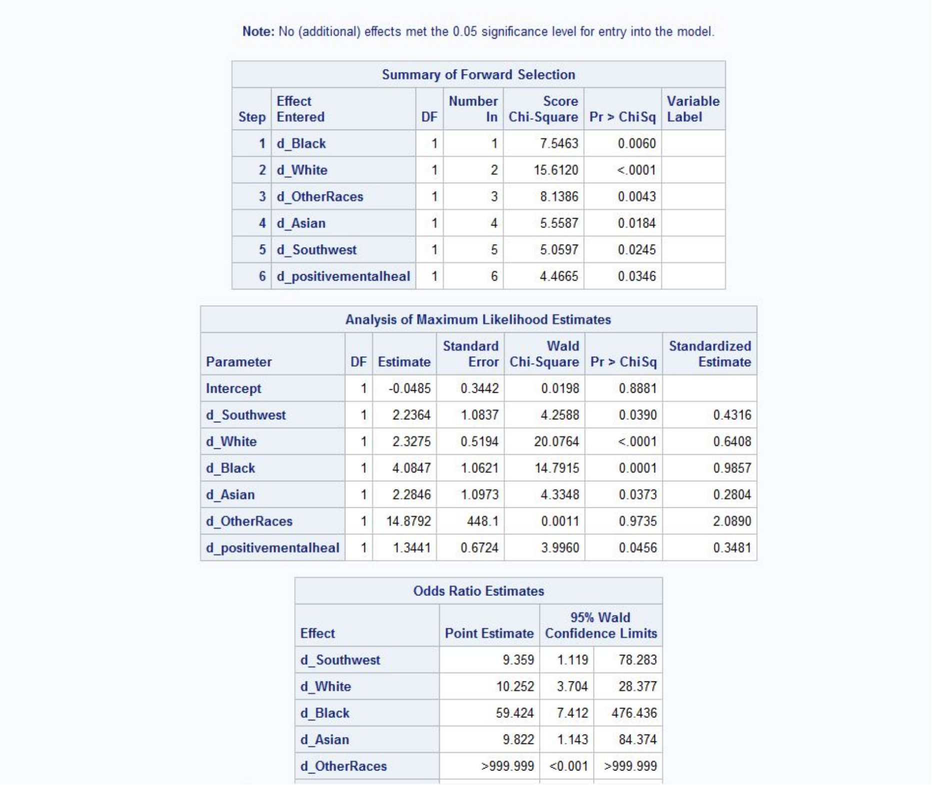


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	181.210	137.085
SC	184.873	162.725
-2 Log L	179.210	123.085

R-Square	0.1771	Max-rescaled R-Square	0.3822

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	56.1257	6	<.0001
Score	66.0992	6	<.0001
Wald	38.0886	6	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
13.9283	11	0.2370

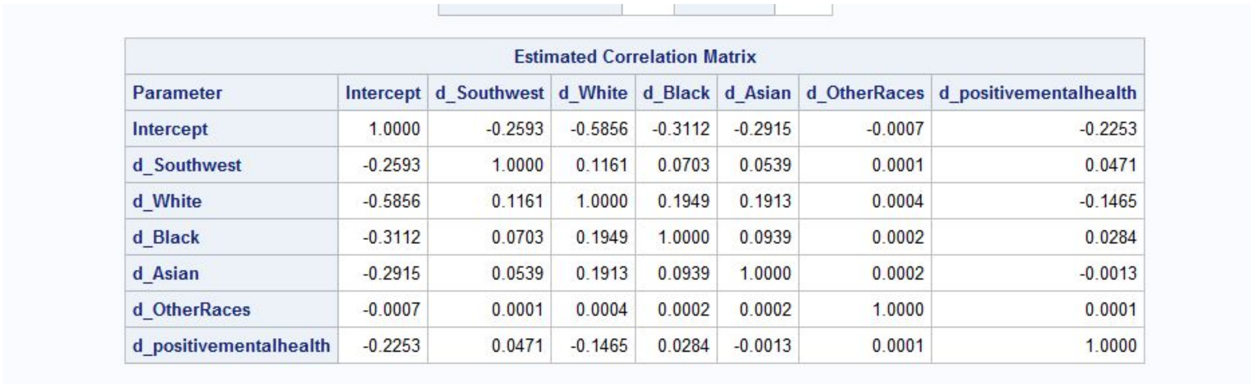


Note: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Forward Selection						
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	d_Black	1	1	7.5463	0.0060	
2	d_White	1	2	15.6120	<.0001	
3	d_OtherRaces	1	3	8.1386	0.0043	
4	d_Asian	1	4	5.5587	0.0184	
5	d_Southwest	1	5	5.0597	0.0245	
6	d_positivementalheal	1	6	4.4665	0.0346	

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.0485	0.3442	0.0198	0.8881	
d_Southwest	1	2.2364	1.0837	4.2588	0.0390	0.4316
d_White	1	2.3275	0.5194	20.0764	<.0001	0.6408
d_Black	1	4.0847	1.0621	14.7915	0.0001	0.9857
d_Asian	1	2.2846	1.0973	4.3348	0.0373	0.2804
d_OtherRaces	1	14.8792	448.1	0.0011	0.9735	2.0890
d_positivementalheal	1	1.3441	0.6724	3.9960	0.0456	0.3481

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
d_Southwest	9.359	1.119	78.283
d_White	10.252	3.704	28.377
d_Black	59.424	7.412	476.436
d_Asian	9.822	1.143	84.374
d_OtherRaces	>999.999	<0.001	>999.999



Estimated Correlation Matrix

Parameter	Intercept	d_Southwest	d_White	d_Black	d_Asian	d_OtherRaces	d_positivementalhealth
Intercept	1.0000	-0.2593	-0.5856	-0.3112	-0.2915	-0.0007	-0.2253
d_Southwest	-0.2593	1.0000	0.1161	0.0703	0.0539	0.0001	0.0471
d_White	-0.5856	0.1161	1.0000	0.1949	0.1913	0.0004	-0.1465
d_Black	-0.3112	0.0703	0.1949	1.0000	0.0939	0.0002	0.0284
d_Asian	-0.2915	0.0539	0.1913	0.0939	1.0000	0.0002	-0.0013
d_OtherRaces	-0.0007	0.0001	0.0004	0.0002	0.0002	1.0000	0.0001
d_positivementalhealth	-0.2253	0.0471	-0.1465	0.0284	-0.0013	0.0001	1.0000

R2= 0.1771 i.e. 17% approximately of data variation will be explained by this model.

Model equation per forward selection
LogP(d_males =1)/1-P(d_males = 0) = 1.000 – 0.2593 d_southwest - 0.5856 d_white -0.3112 d_black -0.2915 d_asian -0.0007 d_otherrace -0.2253 d_postivemeantalheath

Method Selected:
(When compared to Backward and Forward)
Backward Selection Method

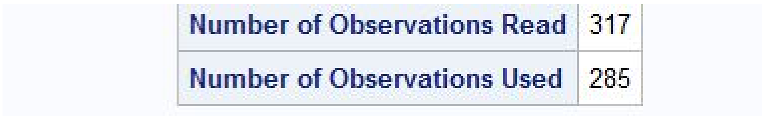
Model equation backward selection
LogP(d_males =1)/1-P(d_males = 0) = 1.000 – 0.2593 d_southwest - 0.5856 d_white -0.3112 d_black -0.2915 d_asian -0.0007 d_otherrace -0.2253 d_postivemeantalheath

R2= 0.177 i.e. 17% approximately of data variation will be explained by this model.

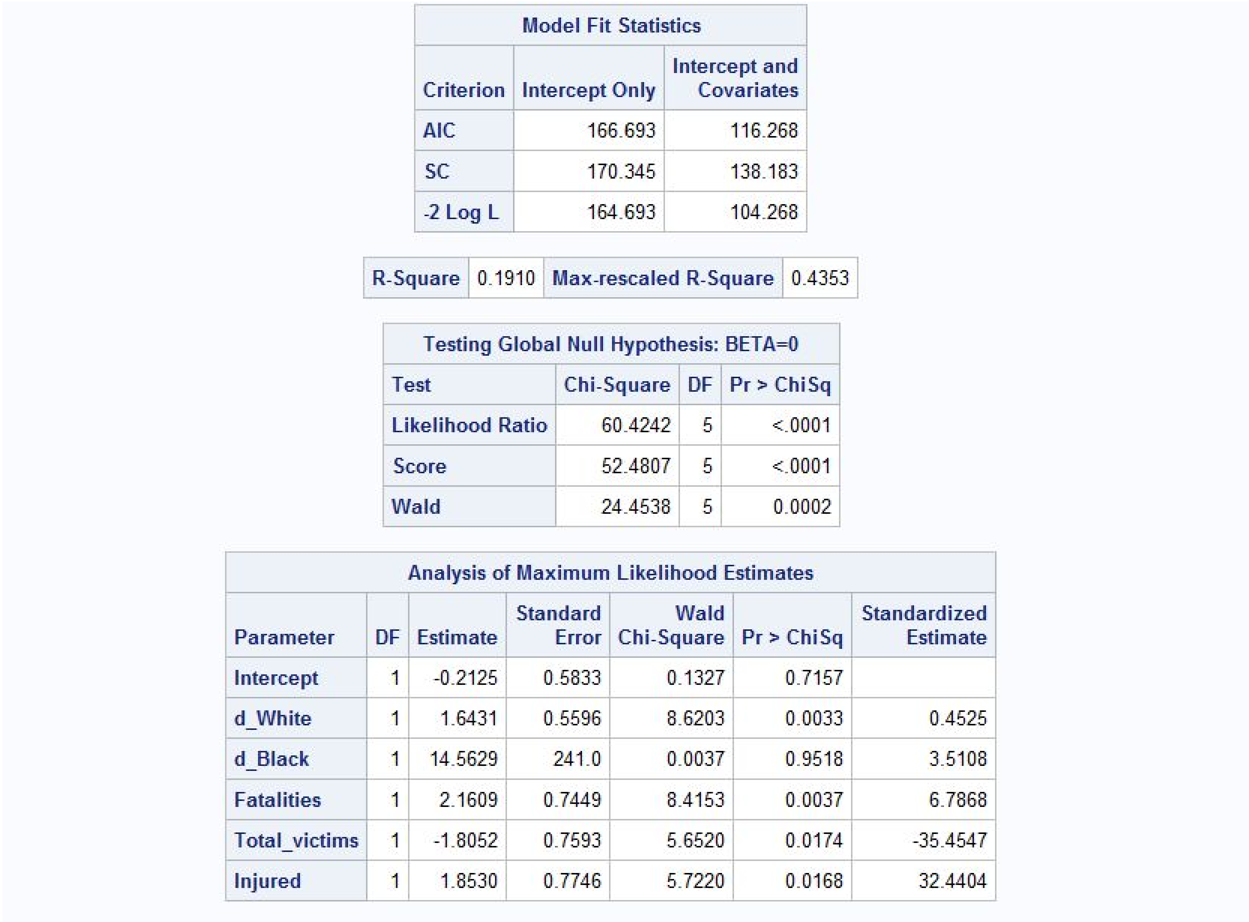
The model fitted by the forward model was selected as it had a better diagnosis. The backward model had a better R2 if compared to the forward model. R2 for the forward model was 13% approximately but R2 for the backward model was 17%. The AIC and SC values were the same for both of the models. The Pr-value in the selected model is below 0.05 which shows that the predictors are significant and can be included in the model and the Backward Selection model also meets the goodness of fit test.

The selected model was also further checked for outliers and influential points. Values given under the pearson residual and deviance residual that were above +3 and -3 were marked as outliers. Similarly the Dfbetas graph was analyzed. After checking the model, observation 1,22 and 24 were removed from the model. The model was then refitted again to check for the changes.

Final Model:



Number of Observations Read	317
Number of Observations Used	285

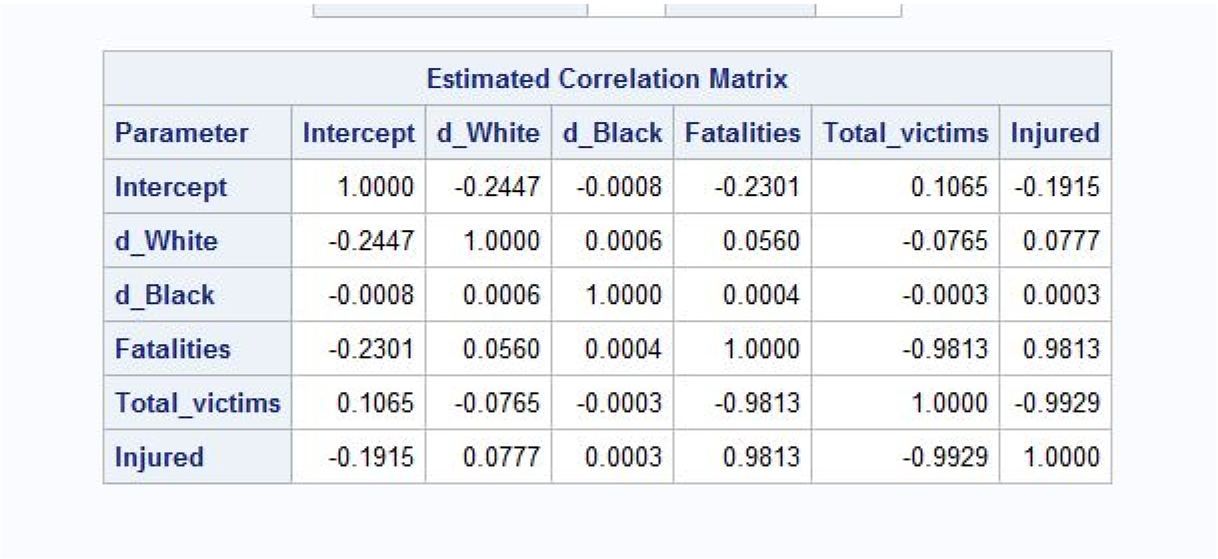


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	166.693	116.268
SC	170.345	138.183
-2 Log L	164.693	104.268

R-Square	0.1910	Max-rescaled R-Square	0.4353

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	60.4242	5	<.0001
Score	52.4807	5	<.0001
Wald	24.4538	5	0.0002

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.2125	0.5833	0.1327	0.7157	
d_White	1	1.6431	0.5596	8.6203	0.0033	0.4525
d_Black	1	14.5629	241.0	0.0037	0.9518	3.5108
Fatalities	1	2.1609	0.7449	8.4153	0.0037	6.7868
Total_victims	1	-1.8052	0.7593	5.6520	0.0174	-35.4547
Injured	1	1.8530	0.7746	5.7220	0.0168	32.4404



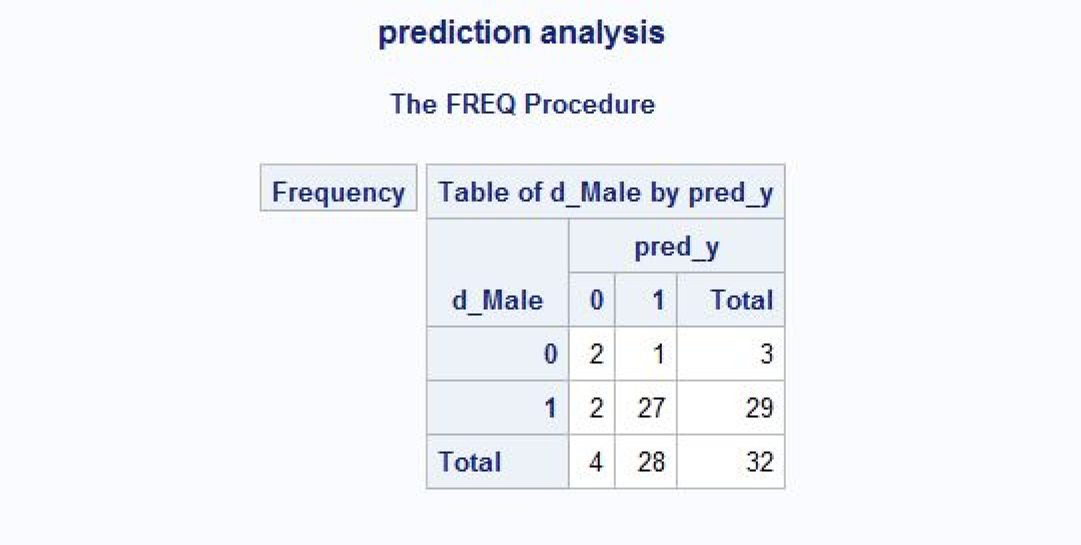
Estimated Correlation Matrix						
Parameter	Intercept	d_White	d_Black	Fatalities	Total_victims	Injured
Intercept	1.0000	-0.2447	-0.0008	-0.2301	0.1065	-0.1915
d_White	-0.2447	1.0000	0.0006	0.0560	-0.0765	0.0777
d_Black	-0.0008	0.0006	1.0000	0.0004	-0.0003	0.0003
Fatalities	-0.2301	0.0560	0.0004	1.0000	-0.9813	0.9813
Total_victims	0.1065	-0.0765	-0.0003	-0.9813	1.0000	-0.9929
Injured	-0.1915	0.0777	0.0003	0.9813	-0.9929	1.0000

Final Fitted Model :
ogP(d_males =1)/1-P(d_males = 0) = 1 -0.2447 d_white -0.0008 d_black -0.2301 fatalities +0.1065 total_victims -0.1915 injured

Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

The R2 for the final fitted model is 0.1910 , indicating that 19% of the variability is explained by the model .No other outliers were removed as the R2 had not increased significantly with the removal of the first outlier All the predictors in the model remain significant as their pr-value is less than alpha 0.05. Finally , the correlation table , indicated that there is no incidence multicollinearity between the variables.

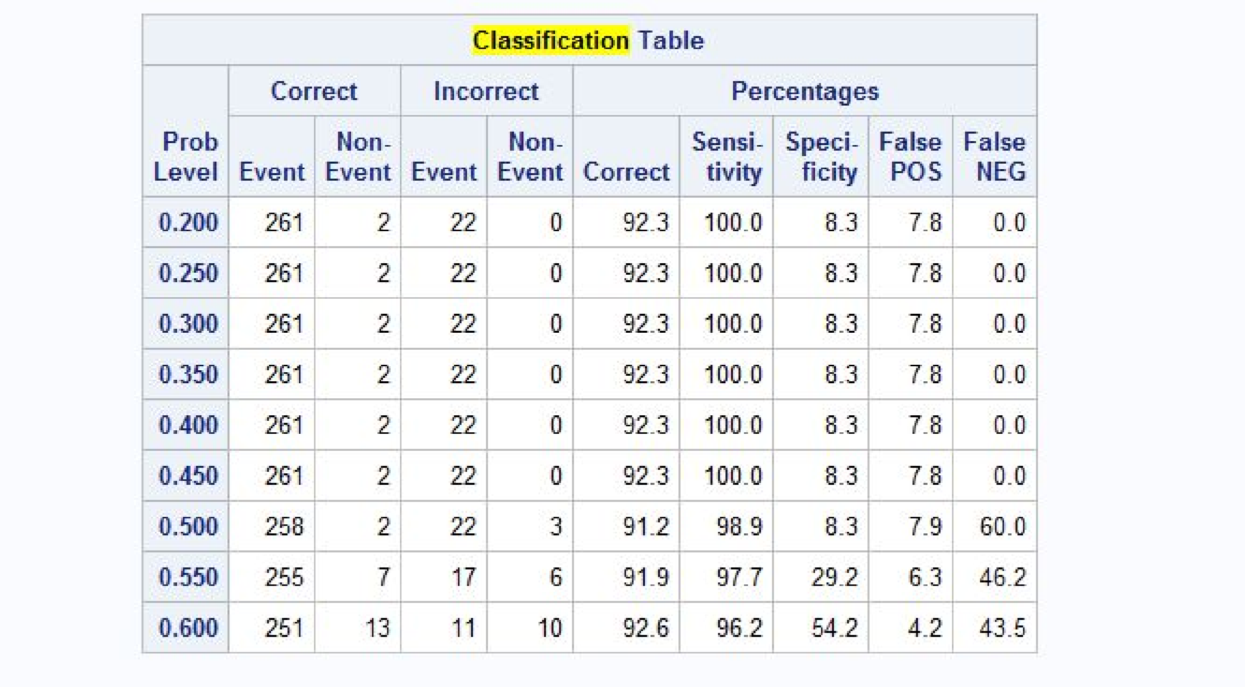
Frequency Table:



prediction analysis

The FREQ Procedure

Frequency	Table of d_Male by pred_y			
		pred_y		
	d_Male	0	1	Total
	0	2	1	3
	1	2	27	29
	Total	4	28	32



Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.200	261	2	22	0	92.3	100.0	8.3	7.8	0.0
0.250	261	2	22	0	92.3	100.0	8.3	7.8	0.0
0.300	261	2	22	0	92.3	100.0	8.3	7.8	0.0
0.350	261	2	22	0	92.3	100.0	8.3	7.8	0.0
0.400	261	2	22	0	92.3	100.0	8.3	7.8	0.0
0.450	261	2	22	0	92.3	100.0	8.3	7.8	0.0
0.500	258	2	22	3	91.2	98.9	8.3	7.9	60.0
0.550	255	7	17	6	91.9	97.7	29.2	6.3	46.2
0.600	251	13	11	10	92.6	96.2	54.2	4.2	43.5

Using the final fitted model, predicited probability was computed for the testing model. A threshold of 0.550 was used to then compute the predicted value of Y. The predicted Y would be equal to 1 if the predicted probability was greater than 0.550. Similarity, the predicted Y would equal to 0

Precision = TP/(TP+FP)
=27/(28)* 100
= 96.42%
This indicates that the model is able to predict the value of the male shooter with the precision of 96.42%

Model 3 – Diksha Joshi

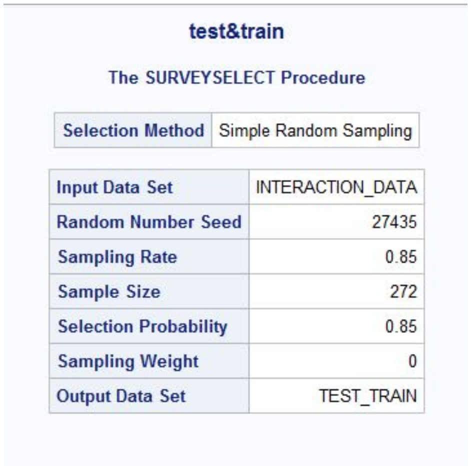
Analysis Report – Diksha Joshi

In this model: "Males" is the dependent variable and "Region" is the independent variable.

In my analysis, I chose to focus on males, in the Northeast region who were "White" with "positive mental health", this was done using logistic regression.

Train and Test

A random sampling of 85-15 is used for this section. For the training set, 85% of the data was used through a random seed number of 27435. The remaining 15% of the data is what is going to be used for the testing aspect of the analysis. The test validation is what will be used to help check the accuracy of the model. The training set had 272 observations, as shown below. Logistic regression was used due to the response variable being binary.



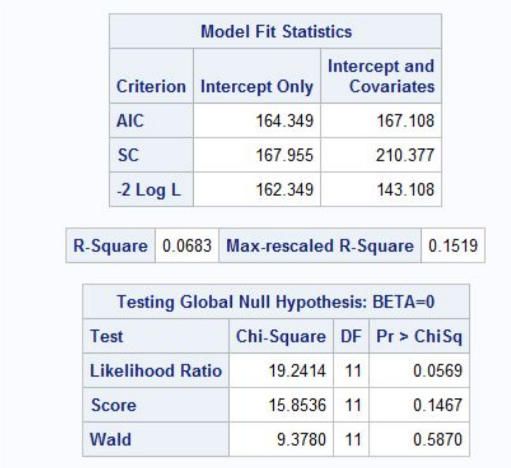
test&train

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	INTERACTION_DATA
Random Number Seed	27435
Sampling Rate	0.85
Sample Size	272
Selection Probability	0.85
Sampling Weight	0
Output Data Set	TEST_TRAIN

A full model logistic regression was ran initially. The results are shown below:



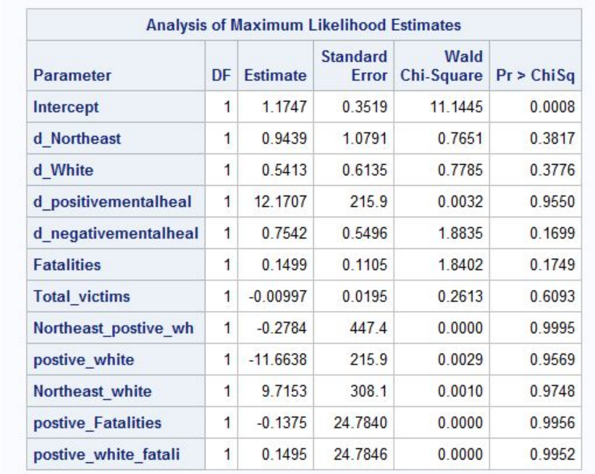
Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	164.349	167.108
SC	167.955	210.377
-2 Log L	162.349	143.108

R-Square	0.0683	Max-rescaled R-Square	0.1519

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.2414	11	0.0569
Score	15.8536	11	0.1467
Wald	9.3780	11	0.5870



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1747	0.3519	11.1445	0.0008
d_Northeast	1	0.9439	1.0791	0.7651	0.3817
d_White	1	0.5413	0.6135	0.7785	0.3776
d_positivementalheal	1	12.1707	215.9	0.0032	0.9550
d_negativementalheal	1	0.7542	0.5496	1.8835	0.1699
Fatalities	1	0.1499	0.1105	1.8402	0.1749
Total_victims	1	-0.00997	0.0195	0.2613	0.6093
Northeast_postive_wh	1	-0.2784	447.4	0.0000	0.9995
postive_white	1	-11.6638	215.9	0.0029	0.9569
Northeast_white	1	9.7153	308.1	0.0010	0.9748
postive_Fatalities	1	-0.1375	24.7840	0.0000	0.9956
postive_white_fatali	1	0.1495	24.7846	0.0000	0.9952

Full model

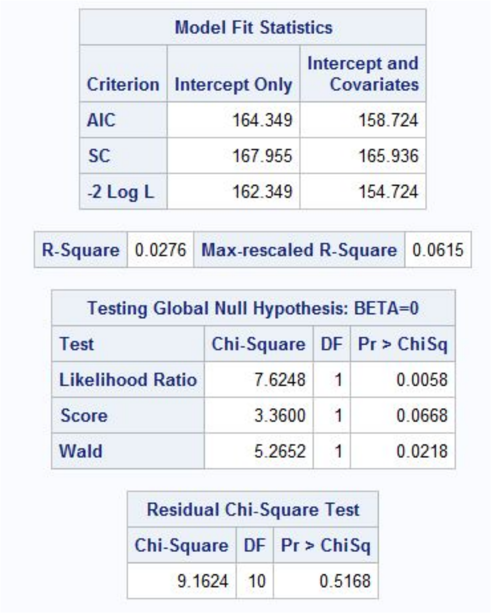
LogP(d_males =1)/1-P(d_males = 0) = 1.14747 + 0.9439 d_Northeast + 0.5413 d_White + 12.1707 positivementalhealth +0.7542 d_negativementalhealth + 0.1499 fatalities -0.00997 total_victims – 0.2784 Northeast_positive_white -11.6638 positive_white + 9.7153 Northeast_white - 0.1375 positive_fatalities+ 0.1495 positive_white_fatalities + e

Where d_White = 1 when u_Race = Latino; otherwise = 0
Where d_Northeast= 1 when Region = midwest, otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= “Yes”, otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

Variable Selection Method

The method used is logistic regression. For the process of variable selection method, backward selection and then forward selection was used.

Backward Selection

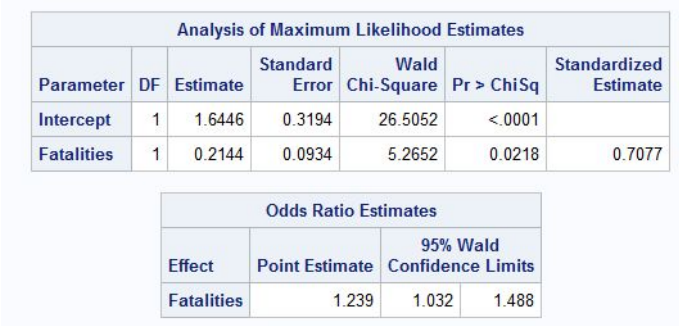


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	164.349	158.724
SC	167.955	165.936
-2 Log L	162.349	154.724

R-Square	0.0276	Max-rescaled R-Square	0.0615

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.6248	1	0.0058
Score	3.3600	1	0.0668
Wald	5.2652	1	0.0218

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.1624	10	0.5168

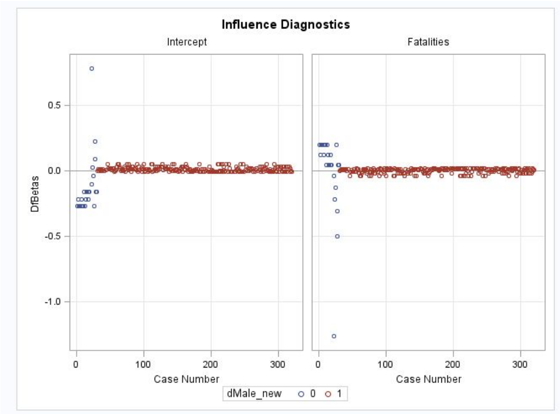


Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.6446	0.3194	26.5052	<.0001	
Fatalities	1	0.2144	0.0934	5.2652	0.0218	0.7077

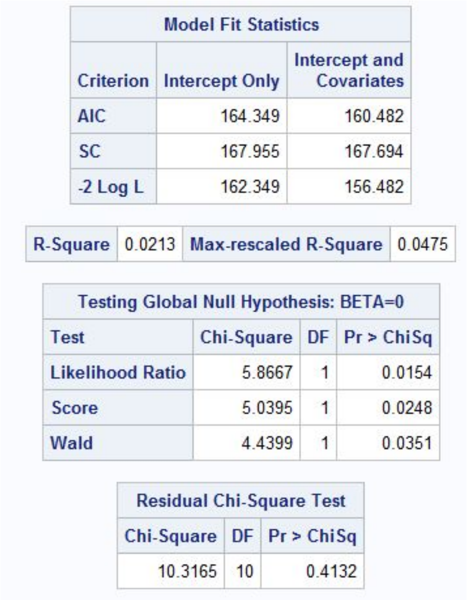
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Fatalities	1.239	1.032	1.488

The Model equation for the backward selection:

LogP(d_males =1)/1-P(d_males = 0) = 1.6446 + 0.2144 fatalities + e



Forward Selection

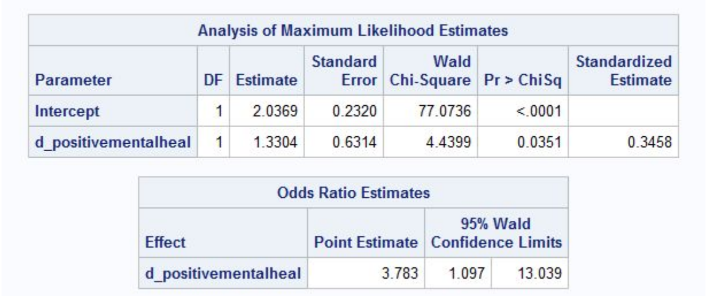


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	164.349	160.482
SC	167.955	167.694
-2 Log L	162.349	156.482

R-Square	0.0213	Max-rescaled R-Square	0.0475

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5.8667	1	0.0154
Score	5.0395	1	0.0248
Wald	4.4399	1	0.0351

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
10.3165	10	0.4132



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	2.0369	0.2320	77.0736	<.0001	
d_positivementalheal	1	1.3304	0.6314	4.4399	0.0351	0.3458

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
d_positivementalheal	3.783	1.097	13.039

The Model equation for the forward selection:

LogP(d_males =1)/1-P(d_males = 0) = 2.0369 + 1.3304 d_positivementalhealth + e

The Selected Model

The Model equation for the backward selection:

LogP(d_males =1)/1-P(d_males = 0) = 1.6446 + 0.2144 fatalities + e

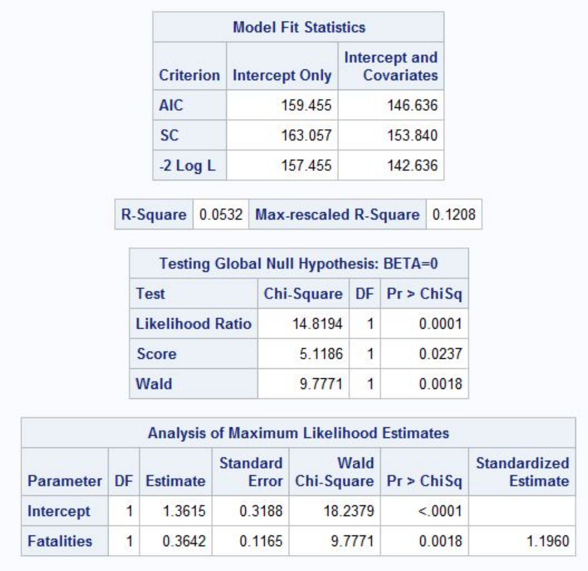
The model that was fitted by the backward selection technique was selected as this model had better diagnostics. In the backward model, R2=0.0276, which is higher than the forward selection model where R2=0.0213. The R2 value of 0.0267 shows that 2.67% of the variability in the data is explained by the model. The AIC and SC values for both forward and backward selection were the same, they had values of 164.349 and 167.955 respectively. These values are error terms and the models show that the error terms are relatively low. The p-value for the predictor shows is less than 0.05 (Alpha), which indicates that the predictor left is significant and should be included in the model. The model also meets the goodness of fit test.
Once the backward selection was run, the only remaining variable left was fatalities. All other variables showed they were insignificant, hence removed from the tables.
The likelihood ratio for the selected model is 7.6248 and has a p value of 0.0058. This shows that it is much lower than alpha=0.05. This shows us that the null hypothesis that indicates that the predictors have no significant relationship to the independent variable and can be rejected. Therefore, the alternative hypothesis that at least one predictor has a significant effect on the response variable can be accepted.

Using backward selection to remove outliers and influential points

The selected model (backward selection) was used to check for possible outliers and influential points. The values that were given under Pearson and deviance residuals were observed and those that were above +3 and -3 were marked as outliers. The Dfbetas graph was also analyzed simultaneously to check for any influential points. |Dfbetas| > 2/√n was the criteria used to

narrow down the influential points. According to the formula Dfbetas with a value more than 0.11 will be marked as influential point.
Observation 22 was removed from the data set as it was the largest outlier value. The model was refitted and then checked again.

The R2 value increased from 0.0276 to 0.0532. The AIC value decreased from 164.349 to 159.455 and the SC value also decreased from 167.955 to 163.057. The likelihood ratio also increased from 7.6248 to 14.8194. Due to the increase in the R2 value and likelihood ratio and the decrease in both AIC and SC values, this shows that the model is closer to a goodness of fit.



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	159.455	146.636
SC	163.057	153.840
-2 Log L	157.455	142.636

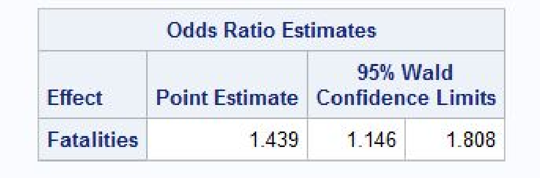
R-Square	0.0532	Max-rescaled R-Square	0.1208

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.8194	1	0.0001
Score	5.1186	1	0.0237
Wald	9.7771	1	0.0018

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.3615	0.3188	18.2379	<.0001	
Fatalities	1	0.3642	0.1165	9.7771	0.0018	1.1960

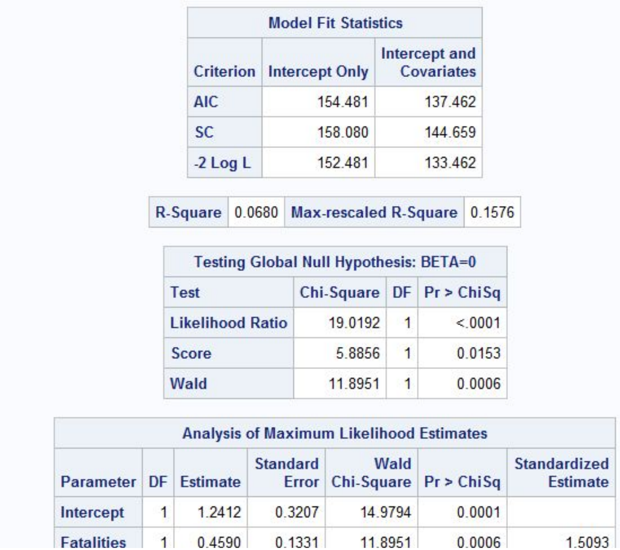


Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Fatalities	1.439	1.146	1.808

After removing 2nd outlier, which was observation 27, the R2 value increased from 0.0532 to 0.0680. The AIC value decreased from 159.455 to 154.481 and the SC value also decreased from 163.057 to 158.080. The likelihood ratio also increased from 14.8194 to 19.0192. Due to the increase in the R2 value and likelihood ratio and the decrease in both AIC and SC values, this shows that the model is even closer to a better goodness of fit.

The Final Model



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	154.481	137.462
SC	158.080	144.659
-2 Log L	152.481	133.462

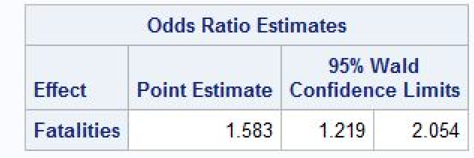
R-Square	0.0680	Max-rescaled R-Square	0.1576

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.0192	1	<.0001
Score	5.8856	1	0.0153
Wald	11.8951	1	0.0006

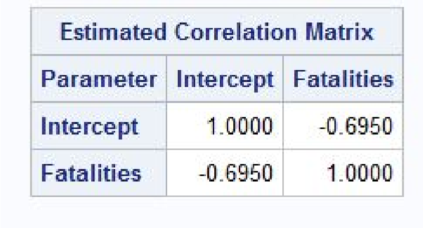
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.2412	0.3207	14.9794	0.0001	
Fatalities	1	0.4590	0.1331	11.8951	0.0006	1.5093



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Fatalities	1.583	1.219	2.054



Estimated Correlation Matrix		
Parameter	Intercept	Fatalities
Intercept	1.0000	-0.6950
Fatalities	-0.6950	1.0000

The final fitted model:

LogP(d_males =1)/1-P(d_males = 0) = 1.2412 + 0.4590 fatalities + e

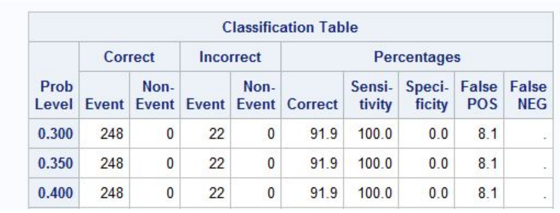
The R2 for the final fitted model is 0.0680 , indicating that 6.80% of the variability is explained by the model. 2 very large outliers were removed and that corresponded to an increase in the overall R2 value. The predictor in the model remains significant as their pr-value is less than alpha 0.05. All other predictors were removed as the pr-value was greater than 0.05, hence being insignificant. The correlation table indicated that there is no incidence multicollinearity between the variables. The null hypothesis can be rejected as the likelihood ratio for the model is 19.0192 and the p-value is less than 0.0001. Because the p value is significantly lower than 0.05, we accept the alternate hypothesis. Due to having large AIC and SC values, this shows that the error in the model is also large. Having such a low R2 value of 0.0680 indicates that a large part of the variability is not being explained by the model.

Odds Ratio

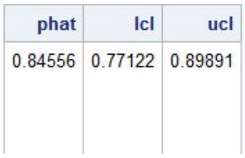
Fatalities: The incidences that resulted in fatalities increases the average odds of the shooter being male by 387% [(ex(1.583)-1)*100], with a 95% confidence interval that he average will be between 238%[(ex(1.219)-1)*100] and 680%[(ex(2.054)-1)*100].

Prediction

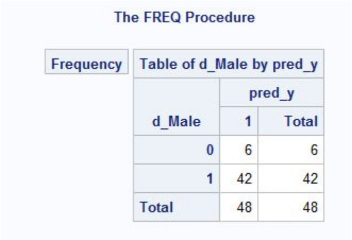
With the classification table shown below, the threshold is 0.30. This was calculated with the sum of each specificity and sensitivity row. The highest sum equals the probability level that shows the threshold.



Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.300	248	0	22	0	91.9	100.0	0.0	8.1	.
0.350	248	0	22	0	91.9	100.0	0.0	8.1	.
0.400	248	0	22	0	91.9	100.0	0.0	8.1	.



phat	lcl	ucl
0.84556	0.77122	0.89891



The FREQ Procedure

Frequency	Table of d_Male by pred_y	
	pred_y	
d_Male	1	Total
0	6	6
1	42	42
Total	48	48

By using the final fitted model, the predicted probability is computed for the testing model. Due to the threshold being 0.30, it was used to compute the predicted Y. If the probability is greater than 0.30, the predicted Y would equal 1. Similarly, the predicted Y would equal 0 if the predicted probability was less than or equal to 0.30.
The performance matrix:

Precision = TP/(TP+FP)
=42/(48)
=0.875 *100=87.50%

This indicates that the model is able to predict the incidence of a male shooter with 87.50% precision.

Model 4 – Preethi Prakash

US mass Shooting - Analysis

Preethi Prakash

The complete dataset is about 320 observations out of which the data is divided into two parts which is training and testing. I have divided my training into 83% of the data and 17% of the data is taken for testing and prediction. The seed value I have for the training set is 76598. Once I split my data into training and testing it has about 266 observations (Fig-1).

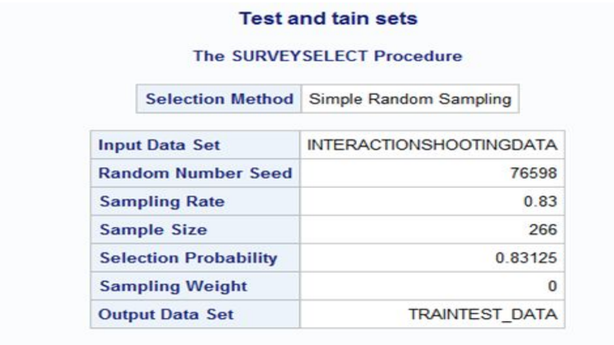
As the response variable is binary, I used **logistic regression model** to fit and train the model. Firstly I have ran the entire model using logistic regression and then used forward and backward model techniques to fit the model accordingly.

The figure below shows the comparative analysis for two models fitted by forward and backward technique.

Code Snippet

```
title"Test and tain sets";
proc surveyselect data=InteractionShootingData out=traintest_data seed=76598
samprate=0.83 outall;
run;
proc print data=traintest_data;
run;
data traintest_data;
set traintest_data;
if selected then d_NewMale=d_Male;
run;
proc print data=traintest_data;
```

```
run;
```



Test and tain sets

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	INTERACTIONSHOOTINGDATA
Random Number Seed	76598
Sampling Rate	0.83
Sample Size	266
Selection Probability	0.83125
Sampling Weight	0
Output Data Set	TRAINTEST_DATA

Fig-1

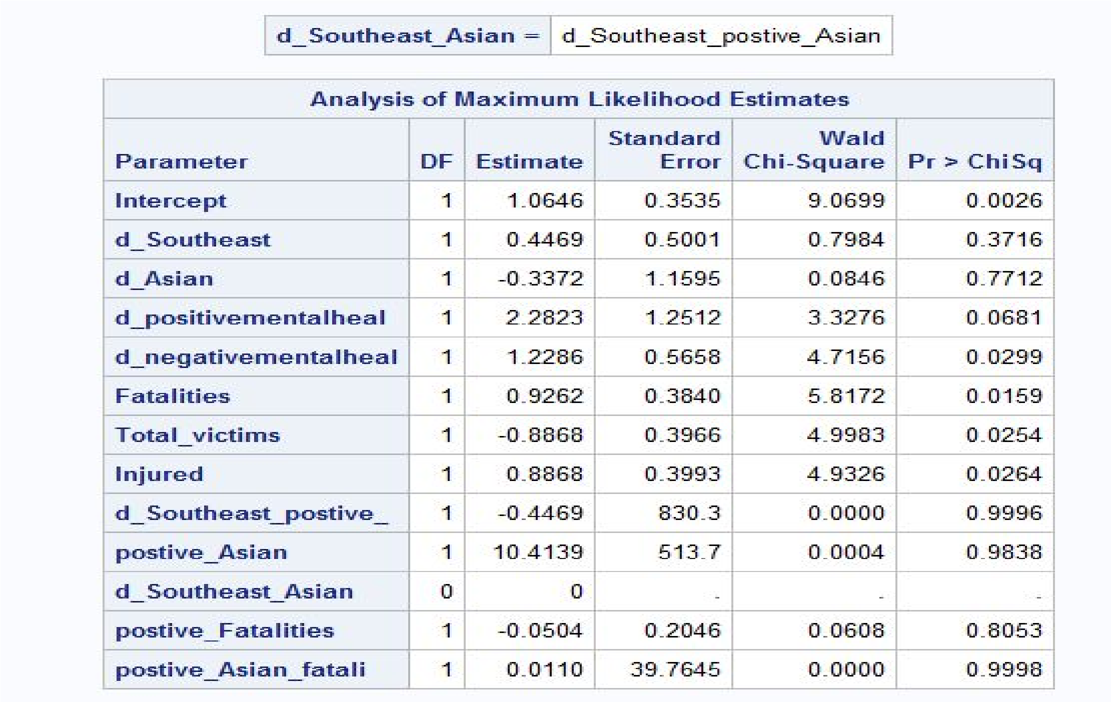
Full model

Code snippet

```
title "fullmodel with my race and region";
proc logistic data=traintest_data;
model d_NewMale(event='1')= d_Southeast d_Asian d_positivementalhealth
d_negativementalhealth Fatalities Total_victims Injured
d_Southeast_postive_Asian postive_Asian d_Southeast_Asian postive_Fatalities
postive_Asian_fatalities / rsquare;
run;
proc logistic data=traintest_data;
```

LogP(d_males =1)/1-P(d_males = 0) = 1.0646+ 0.4469 d_Southeast + -0.3372 d_Asian +2.2823d_positivementalhealth+102286d_negativementalhealth+0.9262Fatalities-0.8868total_victims+0.8868injured 10.4139 positive_Asian+-0.0504 positive_fatalities+e

Where d_Southeast = 1 when u_Race =Asian; otherwise = 0
Where d_Southeast= 1 when Region = Asian otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= "Yes", otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= "No", otherwise=0



d_Southeast_Asian =	d_Southeast_postive_Asian

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0646	0.3535	9.0699	0.0026
d_Southeast	1	0.4469	0.5001	0.7984	0.3716
d_Asian	1	-0.3372	1.1595	0.0846	0.7712
d_positivementalheal	1	2.2823	1.2512	3.3276	0.0681
d_negativementalheal	1	1.2286	0.5658	4.7156	0.0299
Fatalities	1	0.9262	0.3840	5.8172	0.0159
Total_victims	1	-0.8868	0.3966	4.9983	0.0254
Injured	1	0.8868	0.3993	4.9326	0.0264
d_Southeast_postive_	1	-0.4469	830.3	0.0000	0.9996
postive_Asian	1	10.4139	513.7	0.0004	0.9838
d_Southeast_Asian	0	0	.	.	.
postive_Fatalities	1	-0.0504	0.2046	0.0608	0.8053
postive_Asian_fatali	1	0.0110	39.7645	0.0000	0.9998

Fig-2

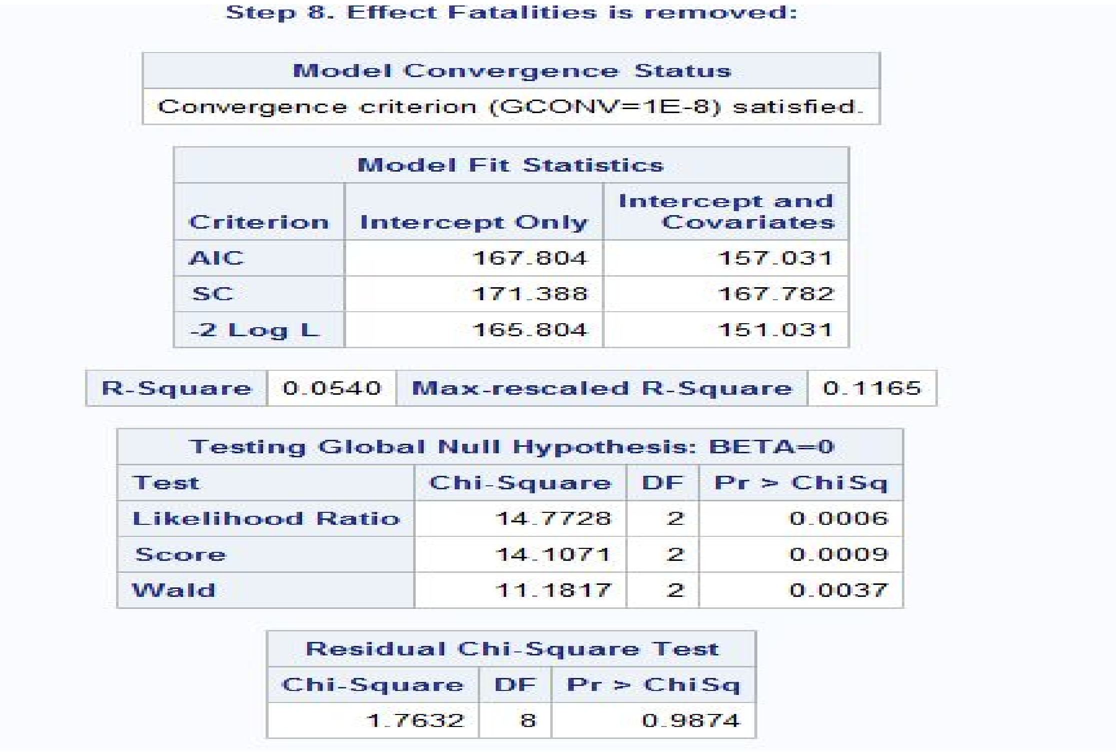
Forward Selection

Code Snippet

```
title "model selection forward";
proc logistic data=traintest_data;
model d_NewMale(event='1')= d_Southeast d_Asian d_positivementalhealth
d_negativementalhealth Fatalities Total_victims
d_Southeast_postive_Asian postive_Asian d_Southeast_Asian postive_Fatalities
postive_Asian_fatalities /selection=forward rsquare influence iplots corrb stb;
run;
```

According to my forward selection model, equation can be written as,

LogP(d_males=1)/1-P(d_males=0) = 1.5404 + 2.2207 d_positivementalhealth + 11.28 d_negativementalhealth+ e



Step 8. Effect Fatalities is removed:

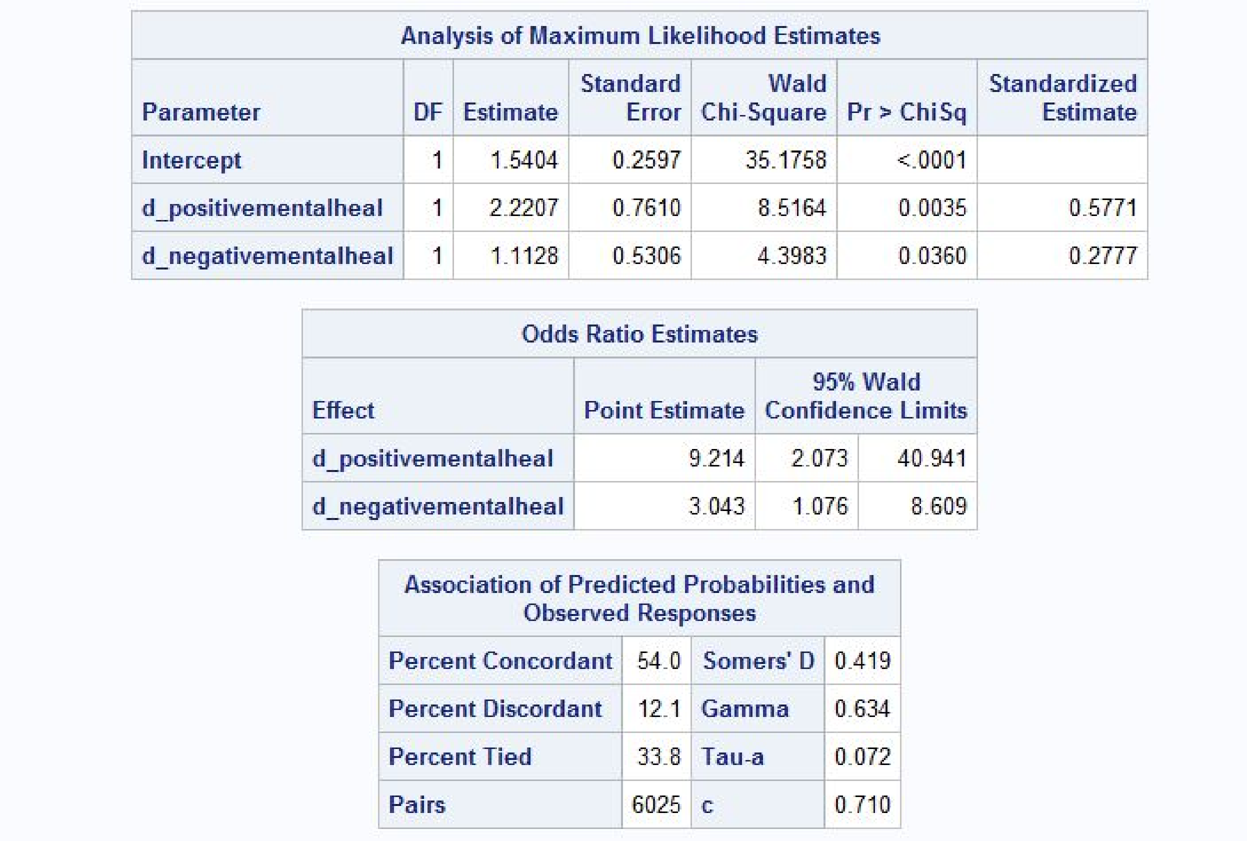
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	167.804	157.031
SC	171.388	167.782
-2 Log L	165.804	151.031

R-Square	0.0540	Max-rescaled R-Square	0.1165

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.7728	2	0.0006
Score	14.1071	2	0.0009
Wald	11.1817	2	0.0037

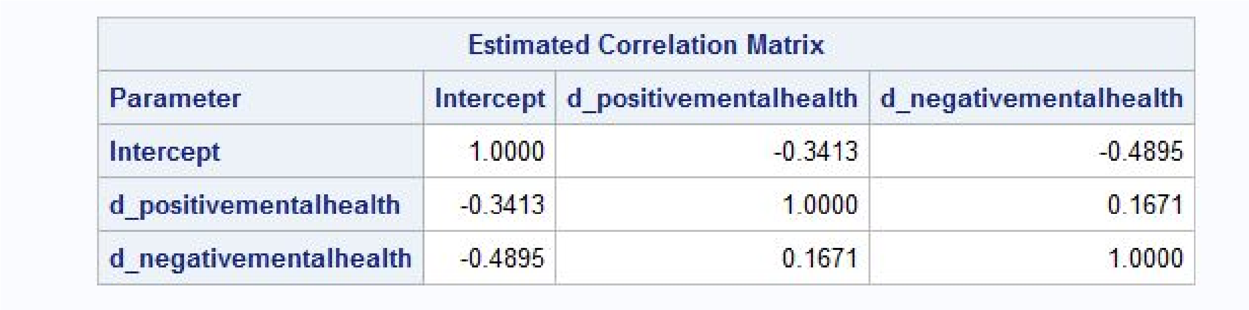
Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
1.7632	8	0.9874



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.5404	0.2597	35.1758	<.0001	
d_positivementalheal	1	2.2207	0.7610	8.5164	0.0035	0.5771
d_negativementalheal	1	1.1128	0.5306	4.3983	0.0360	0.2777

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
d_positivementalheal	9.214	2.073	40.941
d_negativementalheal	3.043	1.076	8.609

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	54.0	Somers' D	0.419
Percent Discordant	12.1	Gamma	0.634
Percent Tied	33.8	Tau-a	0.072
Pairs	6025	c	0.710



Estimated Correlation Matrix			
Parameter	Intercept	d_positivementalhealth	d_negativementalhealth
Intercept	1.0000	-0.3413	-0.4895
d_positivementalhealth	-0.3413	1.0000	0.1671
d_negativementalhealth	-0.4895	0.1671	1.0000

Fig-3

Backward Selection

Code snippet

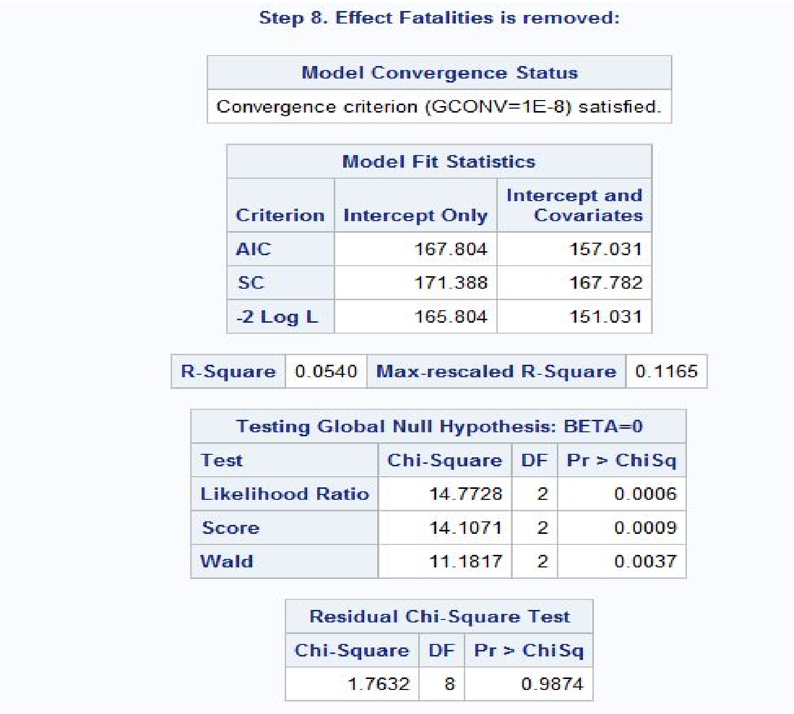
```
title "model selection backward";
proc logistic data=traintest_data;
model d_NewMale(event='1')= d_Southeast d_Asian d_positivementalhealth
d_negativementalhealth Fatalities Total_victims
d_Southeast_postive_Asian postive_Asian d_Southeast_Asian postive_Fatalities
postive_Asian_fatalities /selection=backward rsquare influence iplots corrb stb;
run;
```

According to my backward selection model, equation can be written as,

LogP(d_males=1)/1-P(d_males=0) = 1.5404 + 2.2207 d_positivementalhealth + 11.28 d_negativementalhealth+ e

e

Where we can say that the negative mental health=1 when u_mentalhealth= “No”, otherwise=0



Step 8. Effect Fatalities is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	167.804	157.031
SC	171.388	167.782
-2 Log L	165.804	151.031

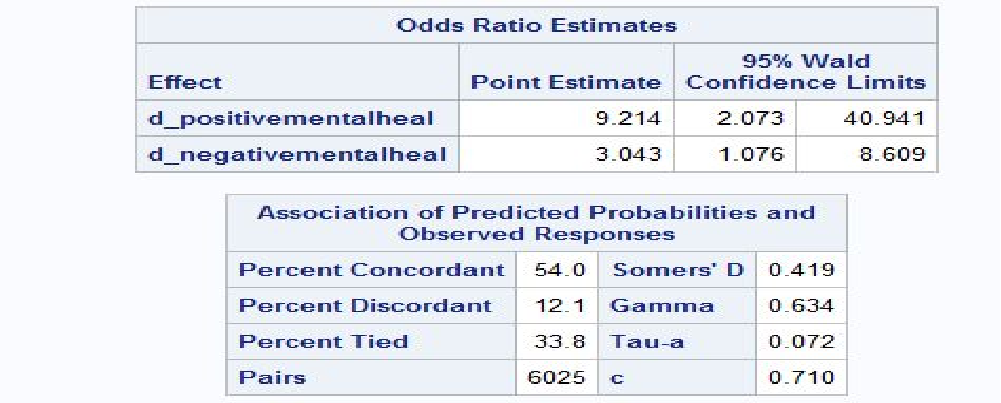
R-Square	0.0540	Max-rescaled R-Square	0.1165

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.7728	2	0.0006
Score	14.1071	2	0.0009
Wald	11.1817	2	0.0037

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
1.7632	8	0.9874

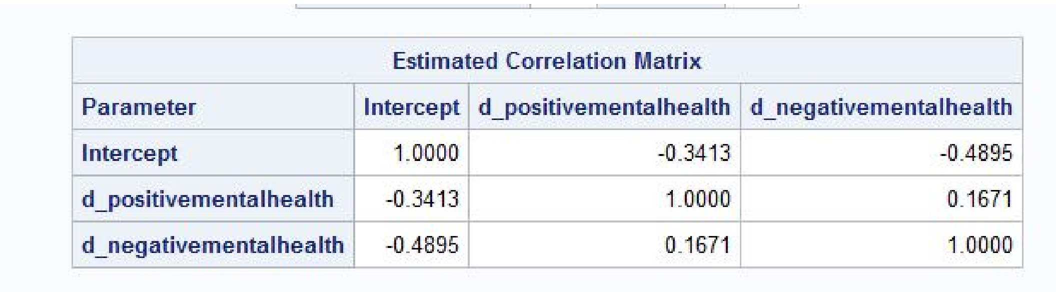


Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
d_positivementalheal	9.214	2.073	40.941
d_negativementalheal	3.043	1.076	8.609

Association of Predicted Probabilities and Observed Responses

Percent Concordant	54.0	Somers' D	0.419
Percent Discordant	12.1	Gamma	0.634
Percent Tied	33.8	Tau-a	0.072
Pairs	6025	c	0.710



Estimated Correlation Matrix

Parameter	Intercept	d_positivementalhealth	d_negativementalhealth
Intercept	1.0000	-0.3413	-0.4895
d_positivementalhealth	-0.3413	1.0000	0.1671
d_negativementalhealth	-0.4895	0.1671	1.0000

Fig-4

Model selected

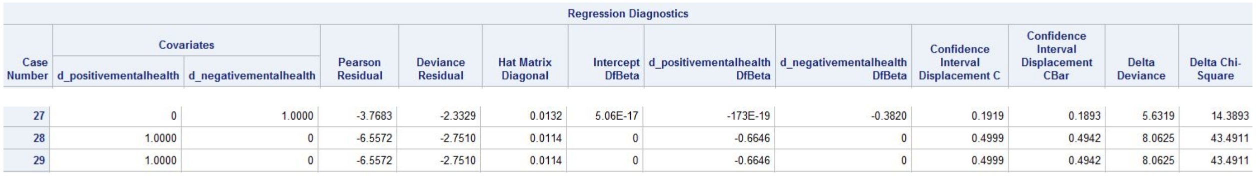
According to both my backward and forward selection model my R square (i.e 0.0540) values and AIC,SC, Variables etc, so there is no different for choosing either one of the model here.

So the model equation for the both my forward and backward model selection can be written as follows,

According to my both my models selection model, equation can be written as,
LogP(d_males=1)/1-P(d_males=0) = 1.5404 + 2.9131 d_positivementalhealth + 1.1128 d_negativementalhealth+ e

Where negative mental health= 1 when U_mental health=”No”,otherwise=0
R2=0.0540
AIC=167.804
SC=171.388 Both AIC and SC are error terms and the model indicates a relatively low error terms.
The pr-value for predictors in the selected model is below the 0.06 (Alpha) indicating that the predictors are significant and be included in the model .Model also meets the goodness of fit test.
$H_o\beta_j=0$, the predictors,positive mental health status and negative mental health status has no significant relationship to the response variable i,e Male.

The likelihood ratio for my model is 14.7728
P- Value of 0.006 , by looking at the above hypothesis we can say that the predictors has no relationships to the independent variable can be rejected. So the alternative hypothesis that at least one of the predictors has significant effect on the response variable can be accepted.



Regression Diagnostics

Case Number	Covariates		Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	d_positivementalhealth DfBeta	d_negativementalhealth DfBeta	Confidence Interval Displacement C	Confidence Interval Displacement CBar	Delta Deviance	Delta Chi-Square
	d_positivementalhealth	d_negativementalhealth										
27	0	1.0000	-3.7683	-2.3329	0.0132	5.06E-17	-173E-19	-0.3820	0.1919	0.1893	5.6319	14.3893
28	1.0000	0	-6.5572	-2.7510	0.0114	0	-0.6646	0	0.4999	0.4942	8.0625	43.4911
29	1.0000	0	-6.5572	-2.7510	0.0114	0	-0.6646	0	0.4999	0.4942	8.0625	43.4911

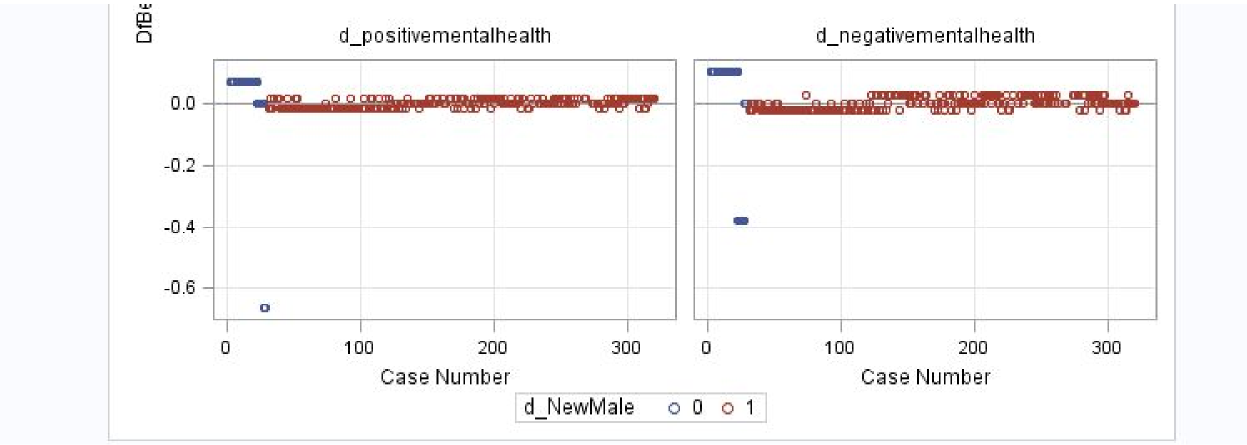


Fig-5

Finally I checked to see if there are any outliers and found one to be at observation 28 and deleted the outliers, after which I noticed that my R square value to be decreasing, The model was then refitted again to check for changes.

Final fitted Model

Code Snippet

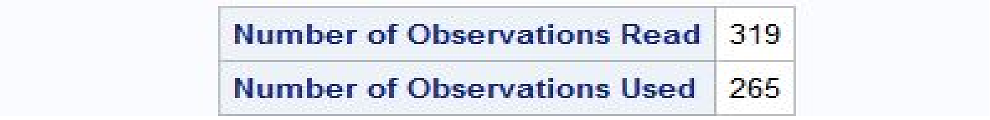
```
proc logistic data=traintest_data;
title "Data set after removing the outlier 29";
model  d_NewMale(event='1')=  d_negativementalhealth  d_positivementalhealth    /  rsquare
influence iplots corrb stb;
run;
```

According to my final model I see increase in my R square value which is 0.0663 and even see a increased likelihood ratio of 18.1774 and two significant values d_Positive mental health and d_negative mental health.

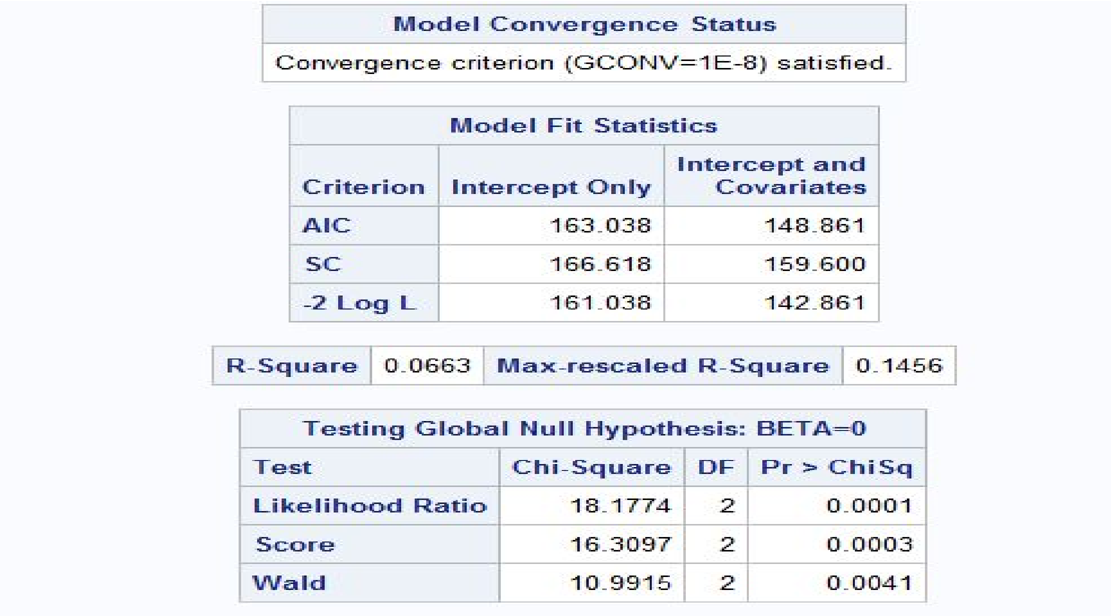
The estimates from the regression for these predictions are:

LogP(d_males =1)/1-P(d_males = 0) = 1.0000 + -0.2500 d_positivementalhealth + -0.4895 d_negativementalhelath +e

So we can say that the person with mental health disorder has more part in the shooting when compared to other factors among male in southeast region who are asian americans.



Number of Observations Read	319
Number of Observations Used	265



Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

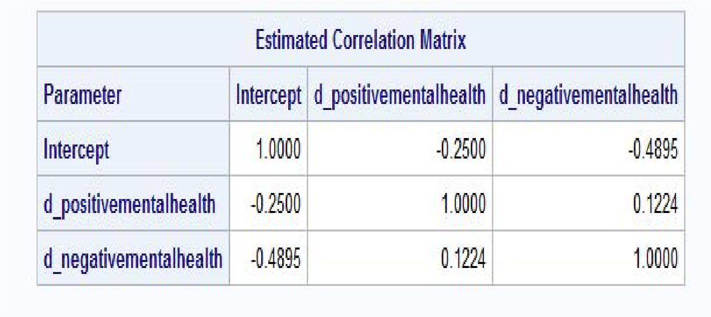
Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	163.038	148.861
SC	166.618	159.600
-2 Log L	161.038	142.861

R-Square	0.0663	Max-rescaled R-Square	0.1456

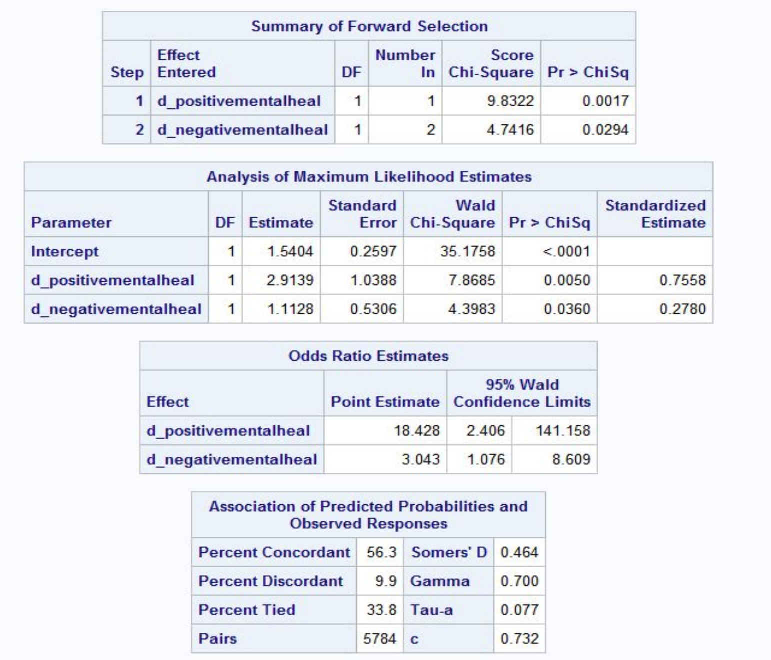
Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.1774	2	0.0001
Score	16.3097	2	0.0003
Wald	10.9915	2	0.0041



Estimated Correlation Matrix

Parameter	Intercept	d_positivementalhealth	d_negativementalhealth
Intercept	1.0000	-0.2500	-0.4895
d_positivementalhealth	-0.2500	1.0000	0.1224
d_negativementalhealth	-0.4895	0.1224	1.0000



Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	d_positivementalheal	1	1	9.8322	0.0017
2	d_negativementalheal	1	2	4.7416	0.0294

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.5404	0.2597	35.1758	<.0001	
d_positivementalheal	1	2.9139	1.0388	7.8685	0.0050	0.7558
d_negativementalheal	1	1.1128	0.5306	4.3983	0.0360	0.2780

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
d_positivementalheal	18.428	2.406	141.158
d_negativementalheal	3.043	1.076	8.609

Association of Predicted Probabilities and Observed Responses

Percent Concordant	56.3	Somers' D	0.464
Percent Discordant	9.9	Gamma	0.700
Percent Tied	33.8	Tau-a	0.077
Pairs	5784	c	0.732

Fig- 6

Threshold Value

Code Snippet

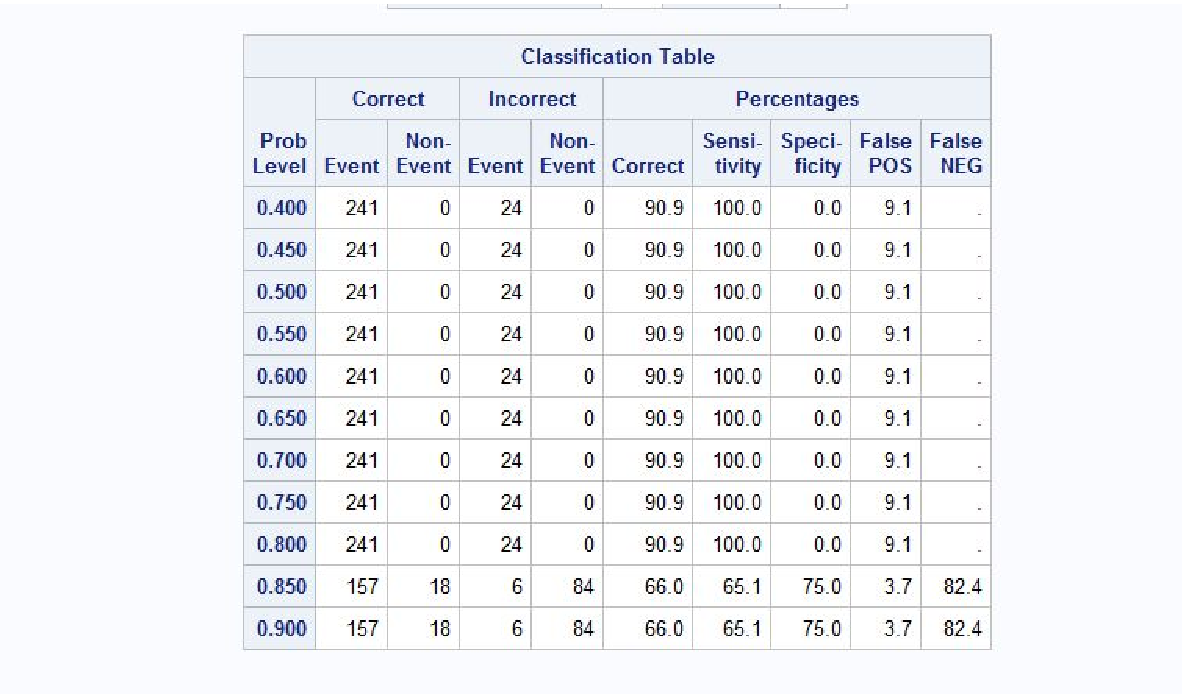
```
Title "Threshold value";
proc logistic data=traintest_data;
model d_NewMale(event='1')= d_positivementalhealth d_negativementalhealth
/selection=forward rsquare influence iplots corrb stb;
run;
Proc Print;
Run;

proc logistic data=traintest_data;
model d_NewMale (event='1') = d_positivementalhealth d_negativementalhealth / ctable
pprob= (0.4 to 0.9 by 0.05);
run;

proc logistic data=traintest_data;
model d_NewMale (event='1') = d_positivementalhealth d_negativementalhealth ;
output out =outpred(where=(d_NewMale=.)) p=phat lower=lcl upper=ucl
predprobs=(individual);
Run;
```

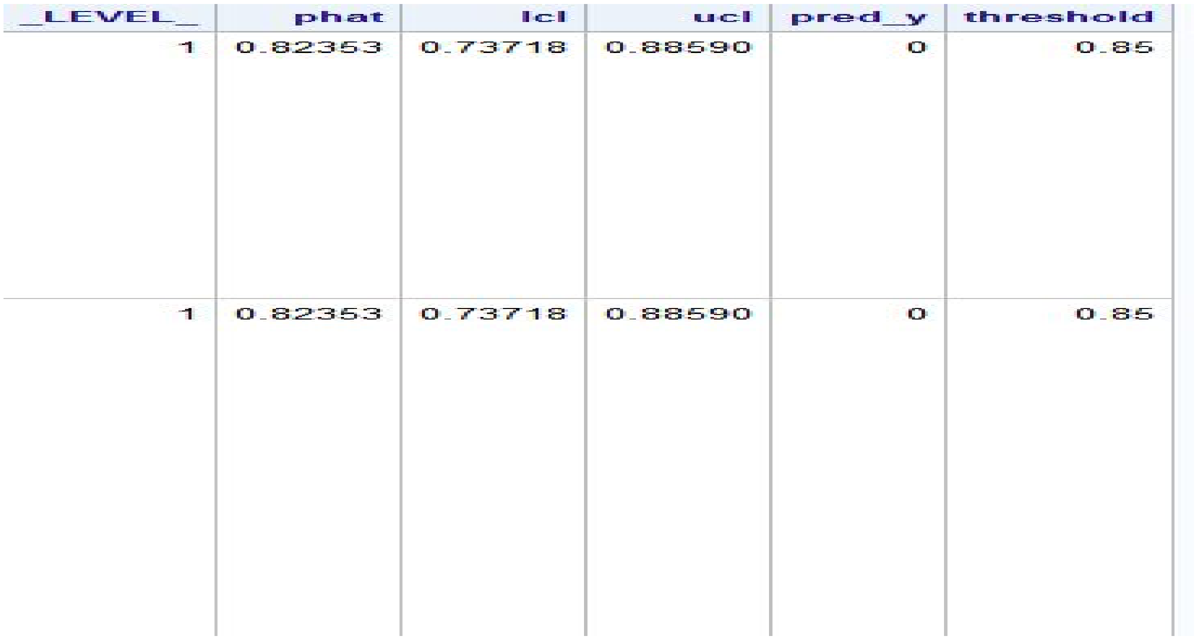
```
proc print;
run;
```

Probability range I choose here is between 0.4 to 0.9, where increment is by 0.05 it helps in classifying the data to get the best threshold.



Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.400	241	0	24	0	90.9	100.0	0.0	9.1	.
0.450	241	0	24	0	90.9	100.0	0.0	9.1	.
0.500	241	0	24	0	90.9	100.0	0.0	9.1	.
0.550	241	0	24	0	90.9	100.0	0.0	9.1	.
0.600	241	0	24	0	90.9	100.0	0.0	9.1	.
0.650	241	0	24	0	90.9	100.0	0.0	9.1	.
0.700	241	0	24	0	90.9	100.0	0.0	9.1	.
0.750	241	0	24	0	90.9	100.0	0.0	9.1	.
0.800	241	0	24	0	90.9	100.0	0.0	9.1	.
0.850	157	18	6	84	66.0	65.1	75.0	3.7	82.4
0.900	157	18	6	84	66.0	65.1	75.0	3.7	82.4

Fig-7



LEVEL	phat	lcl	ucl	pred_y	threshold
1	0.82353	0.73718	0.88590	0	0.85
1	0.82353	0.73718	0.88590	0	0.85

Fig- 8

Fig -8 table shows the lower level (lcl),upper level (ucl) and selected threshold value 0.85

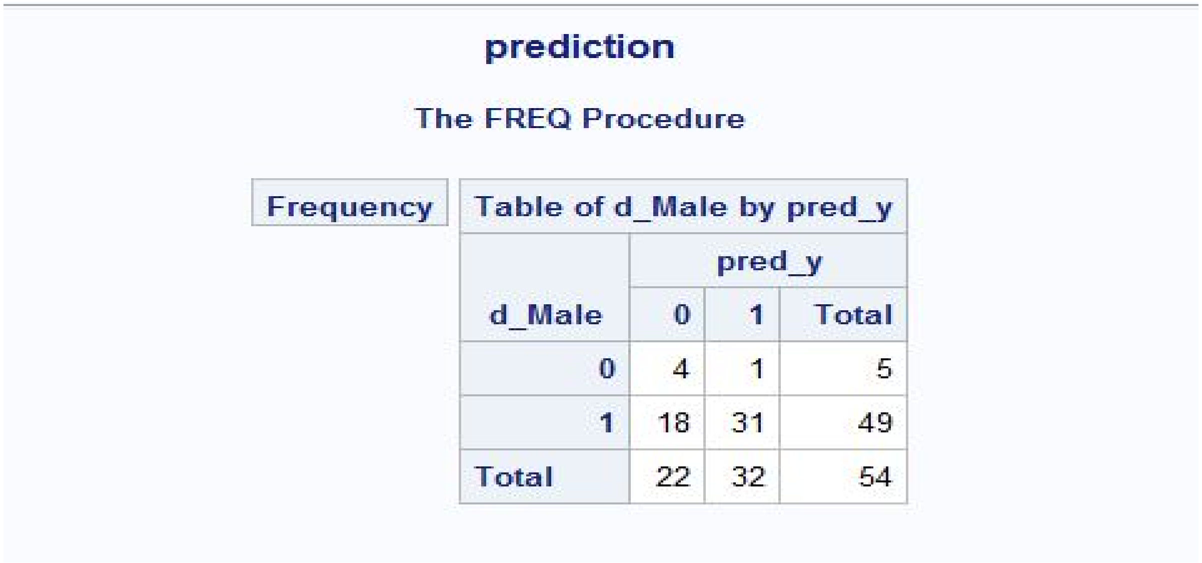
Prediction

Code Snippet

```
data final;
set outpred;
pred_y=0;
threshold=0.85;
if phat>threshold then pred_y=1;
run;
proc print;
run;

title"prediction";
proc freq data=final;
tables d_Male*pred_Y/norow nocol nopercent;
run;
```

Predicted y can be calculated based on the threshold value, I have taken my cut off threshold value to be 0.85.



prediction

The FREQ Procedure

Frequency

Table of d_Male by pred_y			
	pred_y		
d_Male	0	1	Total
0	4	1	5
1	18	31	49
Total	22	32	54

Fig-8

Using the final fitted model,predicted probability was computed for the testing model.
we had 54 in total records of male asian shooters who were from South-East region in our dataset. The model predicted 35 records out of 54 records correctly i.e, 35 records were male shooters from Southeast region who are asian americans and Predicted 19 out of 54 records correctly that male shooters from Southeast region who are not asian americans

A threshold of 0.85 was used to then computed the predicted Y. The predicted Y would equal to 1 if the predicted probability was greater than 0.85. Similarity , the predicted Y would equal to 0 if the predicted probability was less than or equal to 0.85.Performance matrix was then computed as follow

Recall(Sensitivity)= TP/(TP+FN)
=4 / 5
=0.8
=0.8*100
=80%

Precision= TP/ (TP+FP)
=4 / 5
=0.8
=0.8*100
=80%

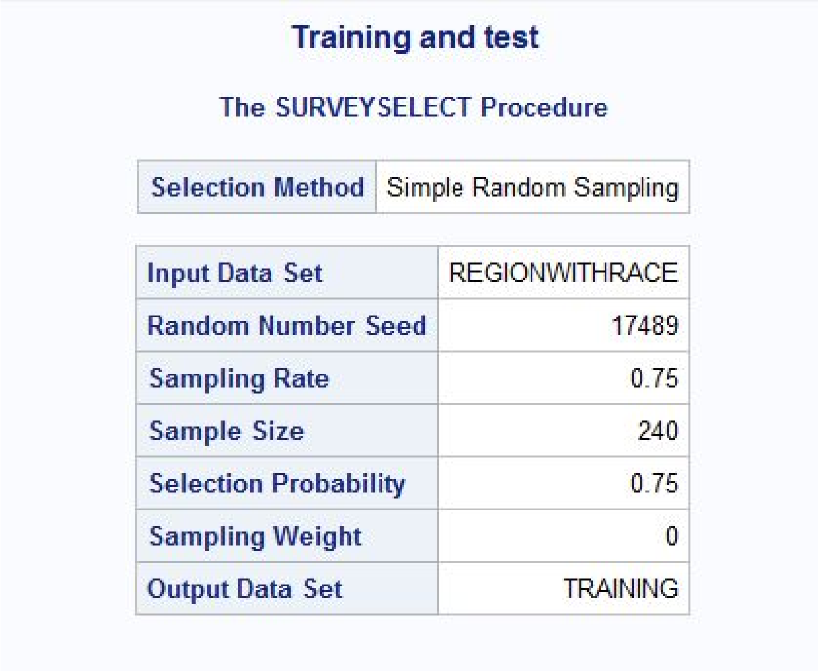
F-Matrix= 2 (0.8 * 0.8) / 0.8 + 0.8
=80%

The model is 80% sensitive and precise for the predicted data that the male from the southeast region are the shooters.

Model 5 – Khizra Masood

- **Training and Testing**

The entire data set was divided into two set for training and testing. 75% of the data was used for model training and the remaining 25% was used model testing and prediction. As you can see below, after the split, the training set has 240 observations. Since the response variable is binary, logistic regression was utilized to both fit and train the model. Initially a full model logistic regression was ran and then both forward and backward model selection techniques were utilized to fit the model accurately. The comparative analysis for two models fitted by forward and backward technique is given below.



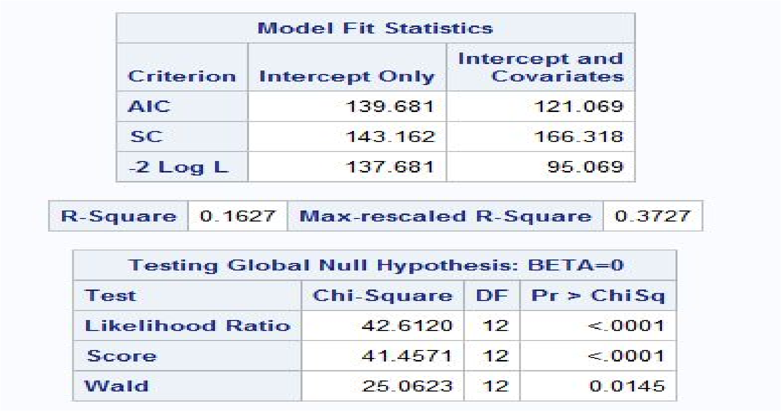
Training and test

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	REGIONWITHRACE
Random Number Seed	17489
Sampling Rate	0.75
Sample Size	240
Selection Probability	0.75
Sampling Weight	0
Output Data Set	TRAINING

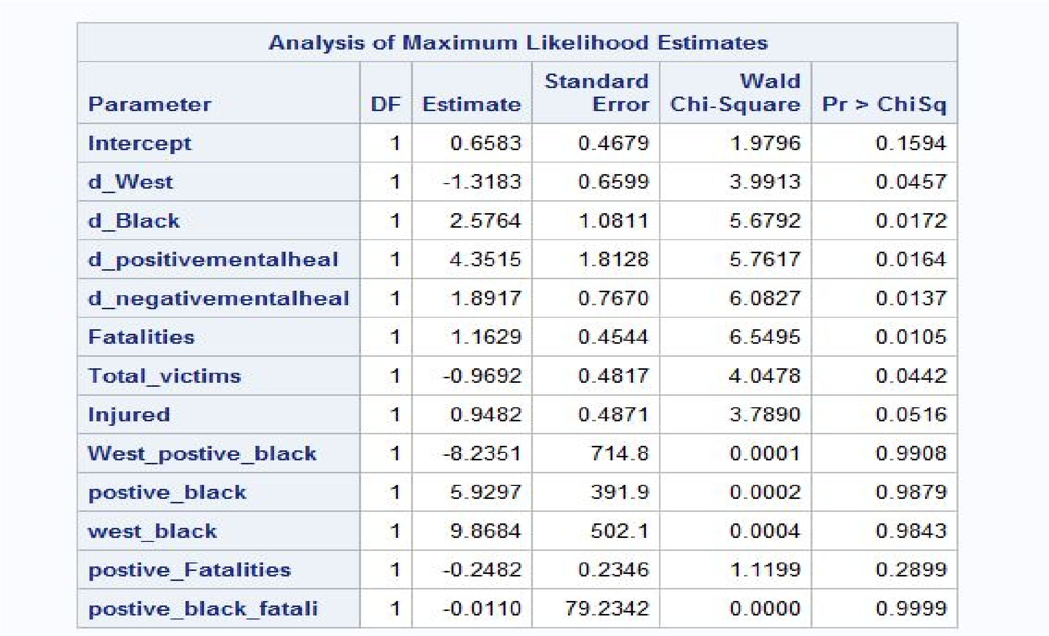
- **Full Model Analysis- Specific West region and African American Race**



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	139.681	121.069
SC	143.162	166.318
-2 Log L	137.681	95.069

R-Square	0.1627	Max-rescaled R-Square	0.3727

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	42.6120	12	<.0001
Score	41.4571	12	<.0001
Wald	25.0623	12	0.0145



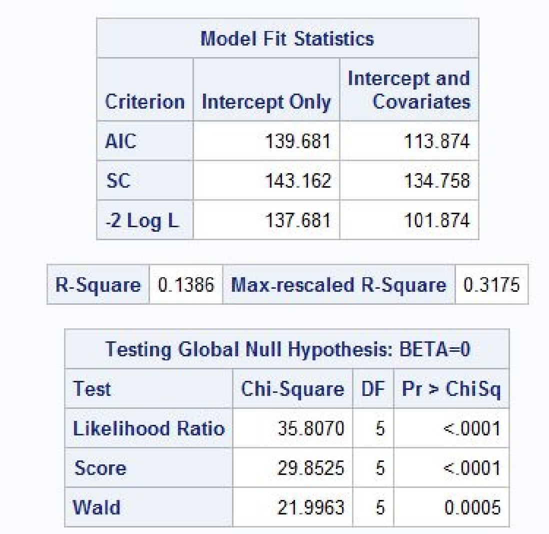
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6583	0.4679	1.9796	0.1594
d_West	1	-1.3183	0.6599	3.9913	0.0457
d_Black	1	2.5764	1.0811	5.6792	0.0172
d_positivementalheal	1	4.3515	1.8128	5.7617	0.0164
d_negativementalheal	1	1.8917	0.7670	6.0827	0.0137
Fatalities	1	1.1629	0.4544	6.5495	0.0105
Total_victims	1	-0.9692	0.4817	4.0478	0.0442
Injured	1	0.9482	0.4871	3.7890	0.0516
West_postive_black	1	-8.2351	714.8	0.0001	0.9908
postive_black	1	5.9297	391.9	0.0002	0.9879
west_black	1	9.8684	502.1	0.0004	0.9843
postive_Fatalities	1	-0.2482	0.2346	1.1199	0.2899
postive_black_fatali	1	-0.0110	79.2342	0.0000	0.9999

Full Model Equation

LogP(d_males =1)/1-P(d_males = 0) = 0.6583 -1.3183 d_west + 2.5764 d_black + 4.3515 positivementalhealth + 1.8917 d_negativementalhealth + 1.1629 fatalities -0.9692 total_victims + 0.9482 injured – 8.2351 west_positive_black + 5.9297 positive_black + 9.8684 west_black – 0.2482 positive_fatilites – 0.0110 positive_black_fatalities + e

Where d_black = 1 when u_Race = black; otherwise = 0
Where d_west= 1 when Region = West, otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= “Yes”, otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

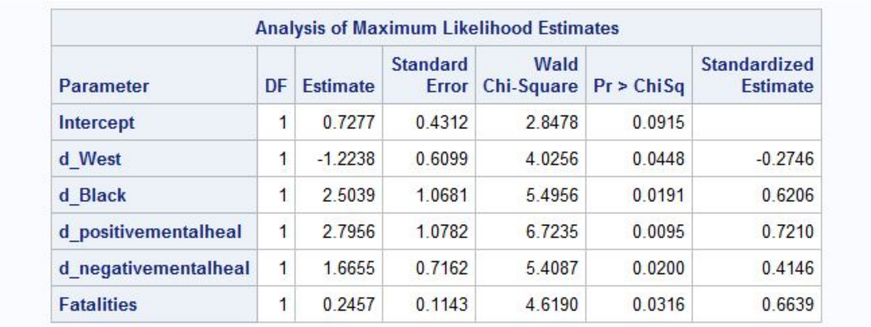
- **Backwards testing selection**



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	139.681	113.874
SC	143.162	134.758
-2 Log L	137.681	101.874

R-Square	0.1386	Max-rescaled R-Square	0.3175

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	35.8070	5	<.0001
Score	29.8525	5	<.0001
Wald	21.9963	5	0.0005



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.7277	0.4312	2.8478	0.0915	
d_West	1	-1.2238	0.6099	4.0256	0.0448	-0.2746
d_Black	1	2.5039	1.0681	5.4956	0.0191	0.6206
d_positivementalheal	1	2.7956	1.0782	6.7235	0.0095	0.7210
d_negativementalheal	1	1.6655	0.7162	5.4087	0.0200	0.4146
Fatalities	1	0.2457	0.1143	4.6190	0.0316	0.6639

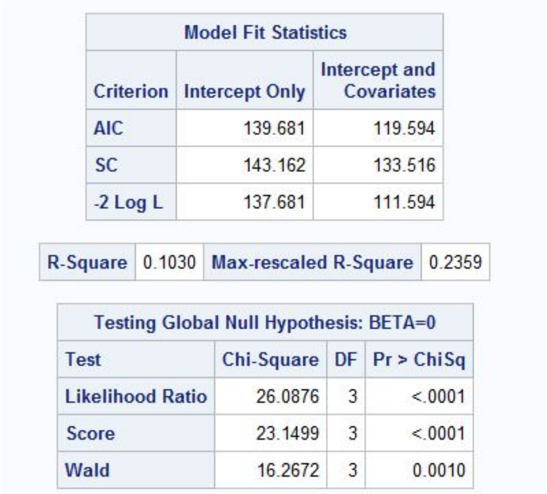
R^2 = 0.1386
AIC and SC = 139.681 and 143.162
LR= 35.8070 and PR > Chisqr = <0.001
5 significant variables: d_black, d_black, d_positivementalhealth, and d_negativementalhealth, fatalities, and total_victims

Model Equation – Backwards selection

LogP(d_males =1)/1-P(d_males = 0) = 0.7277 -1.2238 d_west + 2.5039 d_black + 2.7956 d_positivementalhealth + 1.6655 d_negativementalhealth + 0.2457 fatalities + e

Where d_west = 1 when u_Region=West; otherwise 0
Where d_black = 1 when u_Race = black; otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= "Yes", otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= "No", otherwise=0

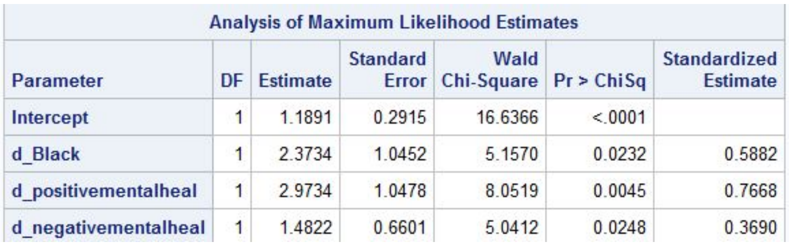
- **Forwards testing selection**



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	139.681	119.594
SC	143.162	133.516
-2 Log L	137.681	111.594

R-Square	0.1030	Max-rescaled R-Square	0.2359

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	26.0876	3	<.0001
Score	23.1499	3	<.0001
Wald	16.2672	3	0.0010



Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.1891	0.2915	16.6366	<.0001	
d_Black	1	2.3734	1.0452	5.1570	0.0232	0.5882
d_positivementalheal	1	2.9734	1.0478	8.0519	0.0045	0.7668
d_negativementalheal	1	1.4822	0.6601	5.0412	0.0248	0.3690

$R^2 = 0.1030$
AIC and SC = 139.681 and 143.162
LR= 26.0876 and PR > Chisqr = <0.001
3 significant variables: d_black, d_positivementalhealth, and d_negativementalhealth

Model Equation – Forward selection
LogP(d_males =1)/1-P(d_males = 0) = 1.1891 + 2.3734 d_black + 2.9734 d_positivementalhealth + 1.4822 d_negativementalhealth + e

Where d_black = 1 when u_Race = black; otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= “Yes”, otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

- **Model Selection = Backwards**

LogP(d_males =1)/1-P(d_males = 0) = 0.7277 -1.2238 d_west + 2.5039 d_black + 2.7956 d_positivementalhealth + 1.6655 d_negativementalhealth + 0.2457 fatalities + e

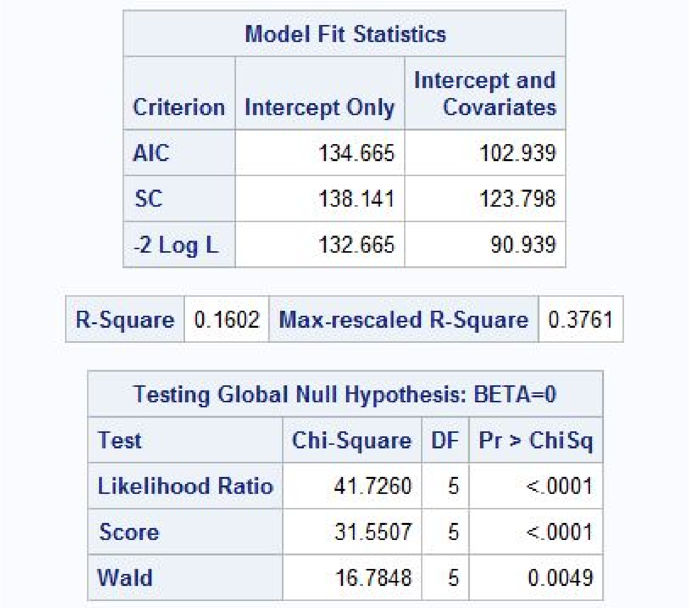
Where d_west = 1 when u_Region=West; otherwise 0
Where d_black = 1 when u_Race = black; otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= “Yes”, otherwise=0

Where negative mentalhealth=1 when u_mentalhealth= “No”, otherwise=0

A good model consists of a maximized R^2 value and minimized AIC and SC. Comparing the two methods of selection, even though backwards selection had more variables, it had a higher R^2 value and likelihood ratio value which mean it is a better to use. The AIC and SC show errors terms and both values remained the same for both.

- **Removing Outliers-**

Using backwards testing model - found observation 28 to be an outlier



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	134.665	102.939
SC	138.141	123.798
-2 Log L	132.665	90.939

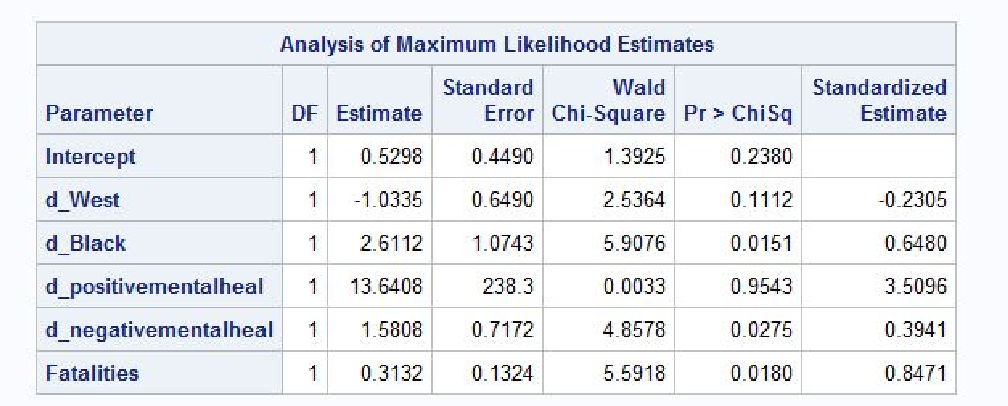
R-Square	0.1602	Max-rescaled R-Square	0.3761

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.7260	5	<.0001
Score	31.5507	5	<.0001
Wald	16.7848	5	0.0049

After removing observation 28,
R^2 = 0.1602 (increased)
AIC and SC = 134.665 and 138.141 (decreased)
LR= 41.7260 and PR > Chisqr = <0.001

Comparing this to the backwards selection model, the r^2 increased and the AIC/SC values decreased. The Liklihood ratio value also increased which shows the model is getting closer to goodness of fit

New Final Model

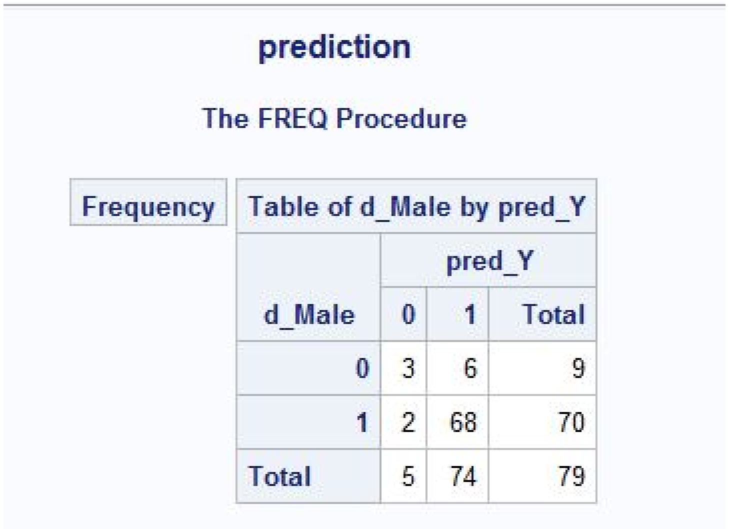


Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.5298	0.4490	1.3925	0.2380	
d_West	1	-1.0335	0.6490	2.5364	0.1112	-0.2305
d_Black	1	2.6112	1.0743	5.9076	0.0151	0.6480
d_positivementalheal	1	13.6408	238.3	0.0033	0.9543	3.5096
d_negativementalheal	1	1.5808	0.7172	4.8578	0.0275	0.3941
Fatalities	1	0.3132	0.1324	5.5918	0.0180	0.8471

LogP(d_males =1)/1-P(d_males = 0) = 0.5298 – 1.0335 d_west + 2.6112 d_black + 13.6408 d_positivementalhealth + 1.5808 d_negativementalhealth + 0.3132 fatalities + e

Where d_west = 1 when u_Region=West; otherwise 0
Where d_black = 1 when u_Race = black; otherwise = 0
Where positivementalhealth=1 when u_mentalhealth= "Yes", otherwise=0
Where negative mentalhealth=1 when u_mentalhealth= "No", otherwise=0

- **Prediction**



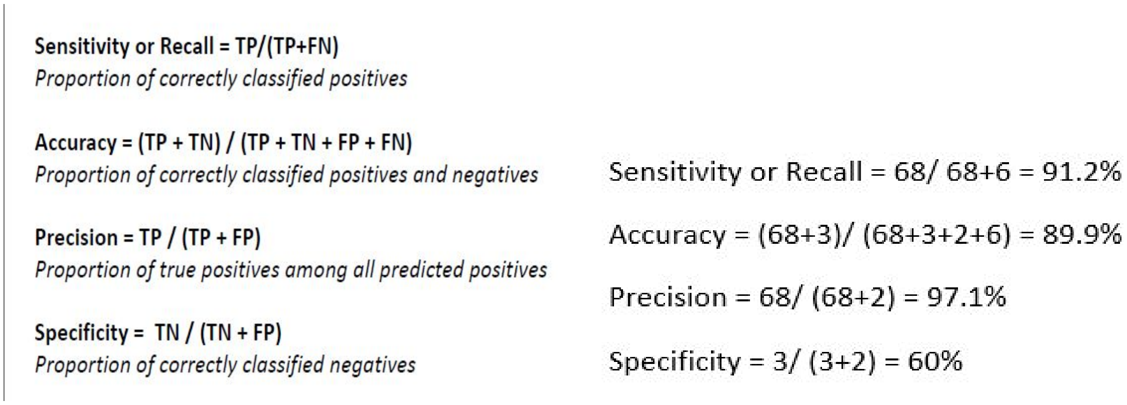
prediction

The FREQ Procedure

Frequency	Table of d_Male by pred_Y		
	pred_Y		
d_Male	0	1	Total
0	3	6	9
1	2	68	70
Total	5	74	79

In the chart above,
0→0 Correctly rejected (TN) True negative = 3
1→0 Mistakenly rejected (FN) False negative = 6

0→1 Mistakenly selected (FP) False positive = 2
1→1 Correctly Selected (TP) = 68



Sensitivity or Recall = TP/(TP+FN)
Proportion of correctly classified positives

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Proportion of correctly classified positives and negatives

Precision = TP / (TP + FP)
Proportion of true positives among all predicted positives

Specificity = TN / (TN + FP)
Proportion of correctly classified negatives

Sensitivity or Recall = 68/ 68+6 = 91.2%

Accuracy = (68+3)/ (68+3+2+6) = 89.9%

Precision = 68/ (68+2) = 97.1%

Specificity = 3/ (3+2) = 60%

F-Metric = 2(Precision*Recall)/(Precision + Recall)
= 2(0.971*0.912)/(0.971+0.912)
= 1.771104/1.883 = **94.1%**

The F-Score above is the measure of the test's accuracy. It considers all samples that are positive (Rcall and precision). The best value is at 1 and worst value is at 0. With 94%, this test is very accurate because it is close to 1.

Final Model :

We picked the Southwest region as our Final Model because it had the highest R^2 value compared to all other values. The model was compared with Southwest region to Caucasian American Males. Compared to other models, the AIC and SC values were lower which showed that there were less lower error terms compared to the others. Lastly, the selected model showed less significant predictors when compared to every other model showing that it's the best fit model for the analysis. In order to get a better prediction for male shooters, more indicators are indicated such as : presence of guns, gun laws within regions, what part of regions the shooters live in and we also had many unknown values within our variables.

Conclusion/Future Work:

For future work: within our model we will be comparing either positive or negative mental health status rather than combining both positive and negative.

The Philadelphia Tribune talked about how most of the shootings are carried out by Caucasian American male shooters – this is because they are more in number in the US. Which concludes that we can not specify a race that conducts more shootings.

The CNN talked about the shootings that have happened in 2018 in the United States (Saeed Ahmed and Christina Walker, 2018). The map showed the most shootings had occurred in the East/South East region. The article only talks about shootings happened in 2018 but the data that we picked shows how many shootings have happened over the years and gives a wider time frame about these particular shootings.

According to – Mass Shootings of America, had also come to a similar conclusion that no significant evidence could be drawn from limited variables. In order to have a better understanding, other phenomenon needs to be explored. (Jiang, 2016)

. Even though shootings are largely happening in the US – gun control laws are trying to be implemented. This is one of the biggest debates that is occurring in the media these days and to stop it or partially make it a less of an issue in the country – gun laws should be supported. Thus, this analysis gives an overview of all the shootings that have occurred in the past few years with all the details. (DeLator 2014)

Lastly, the article – Contagion in Mass Killings and School shootings talks about how it is very important to have health care facilities for every citizen of the US(Towers, Gomez-Lievano, Khan, Mubayi & Castillo-Chavez, 2015) . According to the results of multiple surveys conducted – it has shown a significant difference in the incidents of shooting within areas that have mental health facilities.

References

Fox, J., & DeLateur, M. (2013). Mass Shootings in America. Homicide Studies, 18(1), 125-145. doi: 10.1177/1088767913510297

Jiang, P. (2016). Analysis of Mass Shootings in America. Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/190570_10e739afeb21496f94d38acf0138a2b7.html

Mitchell, J. (2018). Retrieved from http://www.phillytrib.com/news/majority-of-mass-shootings-carried-out-by-white-men/article_8b8b0145-c512-525a-8a7d-256bfb3a959f.html

Saeed Ahmed and Christina Walker, C. (2018). There has been, on average, 1 school shooting every week this year. Retrieved from https://www.cnn.com/2018/03/02/us/school-shootings-2018-list-trnd/index.html

Towers, S., Gomez-Lievano, A., Khan, M., Mubayi, A., & Castillo-Chavez, C. (2015). Contagion in Mass Killings and School Shootings. PLOS ONE, 10(7), e0117259. doi: 10.1371/journal.pone.0117259