PM 511bL - FINAL PROJECT

NAME : AKSHAYMANI SORNALINGAM
ROLL NUMBER : 8467-1951-23

Data Cleaning

In any Data Analysis project, The primary factor to be checked is whether the data is clean or not & if all the required information is there to make the analysis easier. Using the command **misstable summarize** we can check if missing data is there in the table. In our case it does not give us any results (which indicates there are no missing values). We can also use the command **codebook** to check if all the variables used are of the correct data type and if they are labelled (i.e if all the variables have a description). Few of the variables do not have a label, therefore we use the command **. label variable var_name "variable description"** to label them. (The screenshot is just a part of .codebook response which has few unlabeled variables).

```
. log using pm511bproject_smcl_Sornalingam_Akshaymani_inspection.smcl
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\sornalin\Downloads\pm511bproject_smcl_Sornalingam_Akshayma
> ni_inspection.smcl
  log type:  smcl
 opened on:   2 Dec 2024, 01:44:25

. cd "C:\Users\sornalin\Downloads"
C:\Users\sornalin\Downloads

. use project_fall2023.dta, clear

. summarize

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         sid |      3,464      1832.5    1000.115       101       3564
        died |      3,464    .0623557    .2418353         0          1
         age |      3,464    41.70439    17.70749        11         96
        male |      3,464     .767321    .4226003         0          1
        race |      3,464    3.077945    .9124156         1          4
-------------+---------------------------------------------------------
         sbp |      3,464    133.4789    26.74831         7        242
          rr |      3,464    20.20006    6.546318         1         99
         gcs |      3,464    13.83372    2.999577         3         15
       asaps |      3,464    2.543591    1.153961         1          5
         val |      3,464    .1948614    .3961512         0          1

. misstable summarize
(variables nonmissing or string)
```

```
gcs                                                             (unlabeled)

                  Type: Numeric (double)

                 Range: [3,15]                        Units: 1
           Unique values: 12                          Missing .: 0/3,464

                  Mean: 13.8337
             Std. dev.: 2.99958

           Percentiles:     10%       25%       50%       75%       90%
                             11        15        15        15        15

asaps                                                           (unlabeled)

                  Type: Numeric (double)

                 Range: [1,5]                         Units: 1
           Unique values: 5                           Missing .: 0/3,464

              Tabulation: Freq.  Value
                            621  1
                          1,352  2
                            720  3
                            529  4
                            242  5

val                                                             (unlabeled)

                  Type: Numeric (byte)

                 Range: [0,1]                         Units: 1
           Unique values: 2                           Missing .: 0/3,464

              Tabulation: Freq.  Value
                          2,789  0
                            675  1
```

```
. label variable val "Validation"

. label variable asaps "Measure of physical fitness"

. label variable sbp "Systolic blood pressure"

. label variable rr "Respiratory rate"

. label variable gcs "Glasgow Coma Scale"

. codebook
```

Data Analysis

Data Analysis in this case is done by answering the 10 questions given below.

Q1) Provide a publication-quality descriptive table of the sample (with val=0) on the variables above. Report your summary statistics (means, frequencies, etc.) separately by mortality (i.e., your table will have 2 columns – died and survived).
a.Write a short paragraph describing the sample and possible differences by mortality.

```
. keep if val == 0
(675 observations deleted)

.
. table died, statistic(freq) statistic(percent)
```

	Frequency	Percent
In-Hospital Death		
0	2,615	93.76
1	174	6.24
Total	2,789	100.00

```
. tabstat age sbp rr gcs, by(died) stat(mean sd min max)

Summary statistics: Mean, SD, Min, Max
Group variable: died (In-Hospital Death)
```

died	age	sbp	rr	gcs
0	41.08413	133.9728	20.17897	14.13499
	17.40073	25.28805	6.337916	2.516613
	11	17	4	3
	96	242	99	15
1	50.45402	127.7414	21.3046	8.942529
	20.78977	41.72102	9.463302	5.139345
	18	7	1	3
	96	241	90	15
Total	41.6687	133.5841	20.24919	13.81104
	17.77205	26.64314	6.579767	3.02568
	11	7	1	3
	96	242	99	15

```
. table died (male race), statistic(freq) statistic(percent)
```

	male													
	0					1					Total			
	Race					Race					Race			
	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4
In-Hospital Death														
0														
Frequency	56	95	238	228	617	109	387	680	822	1,998	165	482	918	1,050
Percent	2.01	3.41	8.53	8.17	22.12	3.91	13.88	24.38	29.47	71.64	5.92	17.28	32.92	37.65
1														
Frequency	6	4	13	14	37	14	27	52	44	137	20	31	65	58
Percent	0.22	0.14	0.47	0.50	1.33	0.50	0.97	1.86	1.58	4.91	0.72	1.11	2.33	2.08
Total														
Frequency	62	99	251	242	654	123	414	732	866	2,135	185	513	983	1,108
Percent	2.22	3.55	9.00	8.68	23.45	4.41	14.84	26.25	31.05	76.55	6.63	18.39	35.25	39.73

Stata command **table outcome variable (Categorical variable - if needed)** is used to collect information about the frequency and the percentage representation of each level predictor variable in the 2 classes of the binary output variable and to determine the general proportion of the val==0 data in the two classes of died outcome variable. The command **tabstat (continuous predictor variables) by(outcome variable) stat(summary stats parameters)**
Helps us to determine the statistical parameter values (such as mean, standard deviation, minimum value, maximum value of all continuous variables). The values from these tables will be used to create the publication quality table. For our dataset since there are 2700 data points minimum and maximum values are not included in the final table as the minimum and maximum values represent outliers and may not reflect typical variability
The publication quality table should have a clear and Bold heading in Arial font with well defined rows & columns (along with abbreviations / units if any) and a footnote in italics to indicate what format is followed to include values for each variable.

Table 1: Descriptive Summary Statistics table for sample (Val=0)

Variable	Survived (n=2615)	Died (n=174)
Continuous Variables Age	41.08 ± 17.40	50.45 ± 20.79
Systolic Blood Pressure (mm/Hg) [sbp]	133.97 ± 25.29	127.74 ± 41.72
Respiratory Rate (breaths/minute) [rr]	20.18 ± 6.33	21.90 ± 9.46
Glasgow Coma Scale (GCS) [gcs]	14.13 ± 2.51	8.94 ± 5.14
Categorical Variables Gender [male] *Female* *male*	 617 (22.12 %) 37 (0.97 %)	 1998 (71.64 %) 137 (4.91 %)
Ethnicity [race] *Asian* *African American* *Non Hispanic White* *Hispanic White*	 165 (5.92 %) 492 (17.28 %) 910 (32.92 %) 1050 (37.65 %)	 20 (0.72 %) 31 (1.11 %) 65 (2.33 %) 58 (2.08 %)

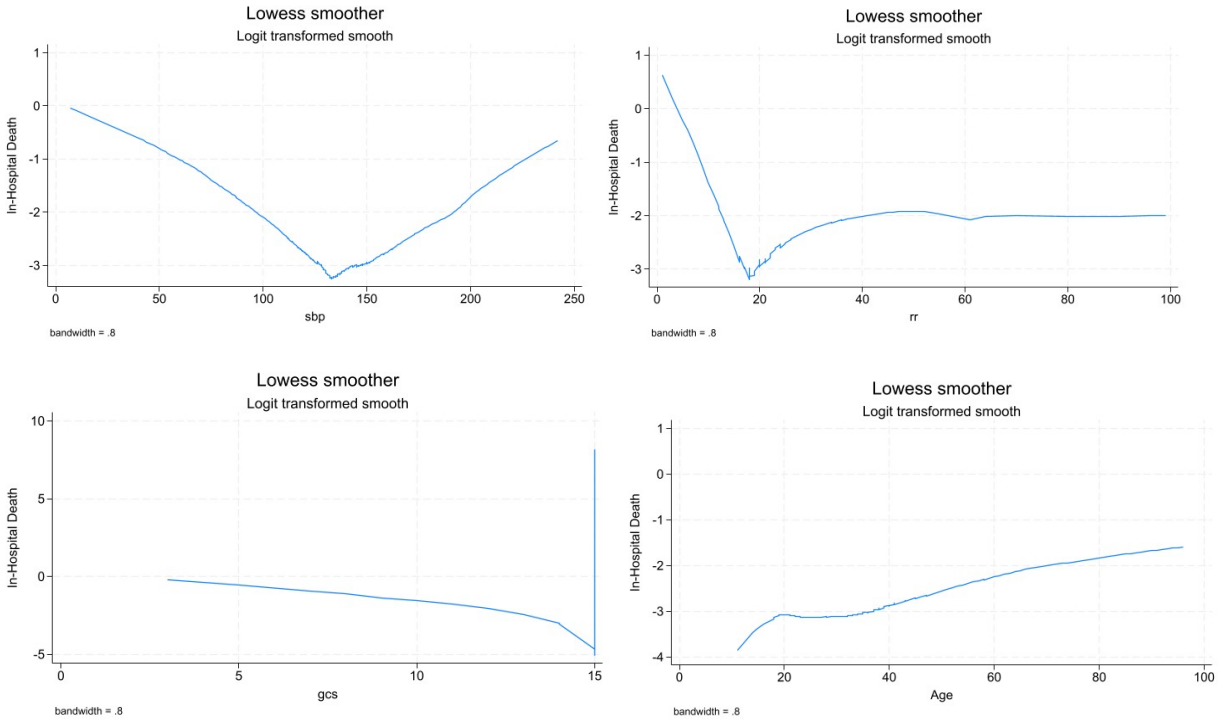
Note : Values for continuous variables are presented as mean ± standard deviation. Categorical variables are presented as frequencies (percentages)

The following table describes the summary statistics of 2789 patients with trauma (with Val=0). Among these 2615 (93.76 %) of them survived (died=0) and 174 (6.24 %) of them died (died=1). When we look at the clinical parameters (i.e the continuous variables) the mean values are completely different from each other. The Average age of patient who died (50.45 ± 20.79) is much more than the age of patient who survived (41.08 ± 17.40) and this aligns with the general medical notion too. While the differences in Systolic blood pressure and Glasgow coma scale were very predominant between the two groups , the difference in Respiratory rate was not that predominant. The Respiratory rate was slightly higher in class died=1 (21.90 + 9.46 breaths/minute) than class died=0 i.e survived (20.18 + 6.33 breaths/minute). On observing the trends in categorical variables we can notice that the percentage of women in both survived and died is higher than men (22.12 % of Women Vs 0.97 % Men with respect to class died=0 and 71.64 % of Women Vs 4.91 % Men with respect to class died=1). When Race is considered we can conclude that Hispanic white patients form the highest proportion in both classes mortality (i.e 37.65 % in class survived and 2.08 % in class died) and Asians forms the lowest proportion in the two classes (i.e 5.92 % in class survived and 0.73 % in class died)

Q2) Using the listed variables above (age through gcs), develop a predictive model (among subjects with val=0), considering main effects and 2-way interactions as possible model terms. Pay attention to how you model continuous variables. Show all of your steps in developing your predictive model. At each step, provide a rationale for the modeling choices you have made.

Logistic regression assumes that the relationship between the predictor and the log-odds of the outcome variable is linear. If the continuous predictors do not have a linear relationship with the output variable then the interpretation of the beta coefficients of the predictors which helps us understand the relationship between the predictor and the output variable will become misleading. Moreover transformation of a predictor variable will improve the fit of the model and help us make better predictions. It will help us understand the underlying complex trends in data. Therefore we first need to assess the linearity of the predictor variables

To do so we can either plot Logit transformed Lowess smooth curve or a two way plot (which has both line graph and scatter plot). Logit transformed Lowess smooth curve is preferred because this transformation is reflective of the general logistic model assumptions and interpretations which makes linearity assessment very effective and accurate. The syntax **lowess outcome variable predictor variable, logit** was used to plot the logit transformed lowess curve. First the curve was plotted for the for continuous variables age, rr, gcs, sbp.



The curves for sbp, rr, gcs clearly suggest that they are not linear with the outcome variable died, therefore they must be transformed. The curve for age is linear but not completely/perfectly linear. So a fractional polynomial test is needed to check what power can this variable be raised to for transformation. The syntax **used is mfp: logit outcome variable predictor variable** which searches and tests for different fractional powers and gives the results of the best power (best model) alone. Following are the frac. Poly. Results for all 4 variables. The results will show the possible powers, the final chosen power and a statistical test with the transformed variable (using the chosen power) [Only 2 sample snapshots are attached]

```
. mfp: logit died rr

Deviance for model with all terms untransformed = 1298.418, 2789 observations

Variable     Model (vs.)   Deviance  Dev diff.   P      Powers   (vs.)

rr           Lin.   FP2    1298.418    39.690  0.000+  1        -.5 -.5
             FP1           1280.751    22.023  0.000+  -2
             Final         1258.728                    -.5 -.5


Transformations of covariates:

-> gen double Irr__1 = X^-.5-.7027423674 if e(sample)
-> gen double Irr__2 = X^-.5*ln(X)-.4958057246 if e(sample)
   (where: X = rr/10)

Final multivariable fractional polynomial model for died

Variable |      -----Initial-----        -----Final-----
         |  df    Select   Alpha    Status   df    Powers
---------+-------------------------------------------------
      rr |   4    1.0000   0.0500      in     4    -.5 -.5

Logistic regression                               Number of obs =  2,789
                                                  LR chi2(2)    =  43.67
                                                  Prob > chi2   = 0.0000
Log likelihood = -629.36398                       Pseudo R2     = 0.0335

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    Irr__1 |  -11.86179   2.080324    -5.70   0.000    -15.93915   -7.784433
    Irr__2 |  -7.918804   1.310974    -6.04   0.000    -10.48827   -5.349342
       _cons |  -2.842558    .084418   -33.67   0.000    -3.008014   -2.677102
------------------------------------------------------------------------------
Note: 0 failures and 2 successes completely determined.
Deviance = 1258.728.
```

```
. mfp: logit died age

Deviance for model with all terms untransformed = 1260.839, 2789 observations

Variable     Model (vs.)   Deviance  Dev diff.   P      Powers   (vs.)

age          Lin.   FP2    1260.839     2.797  0.424   1        0 0
             Final         1260.839                    1


Transformations of covariates:

-> gen double Iage__1 = age-41.66869846 if e(sample)

Final multivariable fractional polynomial model for died

Variable |      -----Initial-----        -----Final-----
         |  df    Select   Alpha    Status   df    Powers
---------+-------------------------------------------------
     age |   4    1.0000   0.0500      in     1    1

Logistic regression                               Number of obs =  2,789
                                                  LR chi2(1)    =  41.56
                                                  Prob > chi2   = 0.0000
Log likelihood = -630.41939                       Pseudo R2     = 0.0319

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    Iage__1 |   .0263506   .0040063     6.58   0.000     .0184983    .0342029
       _cons |  -2.812548   .0851536   -33.03   0.000    -2.979446   -2.64565
------------------------------------------------------------------------------
Deviance = 1260.839.
```

The final transformed variables for the individual main predictors are

Variable Name	Power Chosen	Transformed Variable
age	1	(age)^1
gcs	3	(gcs/10)^3
sbp	2, 2	(sbp/100)^2, (sbp/100)^2 * ln(sbp/100)
rr	-0.5, -0.5	(rr/10)^-0.5, (rr/10)^-0.5 * ln(rr/10)

Square root transformation is chosen by FP for the variable rr to address the issues caused by skewed distribution. For sbp and gcs the transformation of squaring and cubing is chosen as a result of its existing curvilinear relationship between the outcome and untransformed predictor variable. FP did not transform the age variable therefore we can assume its relationship with died as linear.

We first develop a full logit model with all the main effects alone (simple logit model)

```
. logit died age race male mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2

Iteration 0:  Log likelihood = -651.19837
Iteration 1:  Log likelihood =  -491.0746
Iteration 2:  Log likelihood = -459.65924
Iteration 3:  Log likelihood = -459.34148
Iteration 4:  Log likelihood = -459.34106
Iteration 5:  Log likelihood = -459.34106

Logistic regression                               Number of obs =  2,789
                                                  LR chi2(8)    =  383.71
                                                  Prob > chi2   = 0.0000
Log likelihood = -459.34106                       Pseudo R2     = 0.2946

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0337462   .0050015     6.75   0.000     .0239435    .0435489
        race |  -.0922615   .0934456    -0.99   0.323    -.2754115    .0908885
        male |   .4518276   .2235025     2.02   0.043     .0137708    .8898843
    mod_rr_1 |  -8.716768   2.446277    -3.56   0.000    -13.51138   -3.922153
    mod_rr_2 |  -4.542091   1.496472    -3.04   0.002    -7.475123   -1.60906
      mod_gcs |  -.9736907   .0655283   -14.86   0.000    -1.102124   -.8452575
   mod_sbp_1 |  -2.198556   .4362334    -5.04   0.000    -3.053558   -1.343555
   mod_sbp_2 |   2.32786   .4999664     4.66   0.000     1.347944    3.307776
       _cons |   8.790985   2.424988     3.63   0.000     4.038096    13.54387
------------------------------------------------------------------------------
```

```
. logit died age male mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2

Iteration 0:  Log likelihood = -651.19837
Iteration 1:  Log likelihood = -491.36584
Iteration 2:  Log likelihood = -460.13871
Iteration 3:  Log likelihood = -459.82455
Iteration 4:  Log likelihood = -459.82415
Iteration 5:  Log likelihood = -459.82415

Logistic regression                               Number of obs =  2,789
                                                  LR chi2(7)    =  382.75
                                                  Prob > chi2   = 0.0000
Log likelihood = -459.82415                       Pseudo R2     = 0.2939

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0339946   .0049938     6.81   0.000     .0242071    .0437822
        male |   .4456788   .2233792     2.00   0.046     .0078637    .8834939
    mod_rr_1 |  -8.678546   2.449997    -3.54   0.000    -13.48045   -3.876639
    mod_rr_2 |  -4.537358   1.500553    -3.02   0.002    -7.478387   -1.596329
      mod_gcs |  -.9752049   .0654507   -14.90   0.000    -1.103486   -.8469239
   mod_sbp_1 |  -2.213462   .4356198    -5.08   0.000    -3.067261   -1.359663
   mod_sbp_2 |   2.344444   .4983853     4.70   0.000     1.367626    3.321261
       _cons |   8.495946   2.409992     3.53   0.000     3.77245    13.21944
------------------------------------------------------------------------------

. lrtest full_no_interaction

Likelihood-ratio test
Assumption: . nested within full_no_inte~n

 LR chi2(1) =    0.97
Prob > chi2 = 0.3256
```

The following model results (simple model) tell us that the categorical variable race is not statistically significant. A model was run without race was run and an LR test was carried out. Lr test gave a p value of 0.33 which shows that the test was not significant , therefore the model without race performs betters. Moreover if we look at the distribution of races in class died =0 and died=1 since the number of data points for died=1 is much lower there seems to be an imbalance

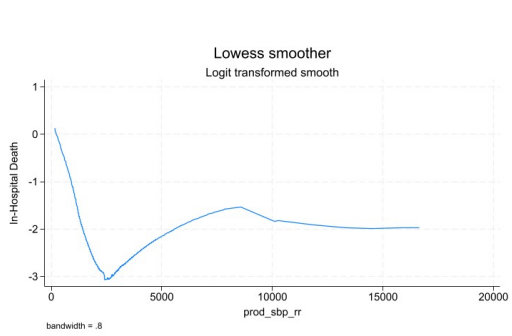
for race in the data across the levels of outcome variable. Along with this the LR test and the p value are reasons for omitting race from our prediction model.

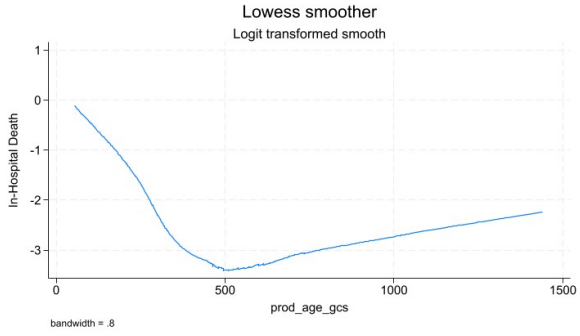
We then introduce interaction terms between the continuous variables and the categorical variables. In this model race is not included.

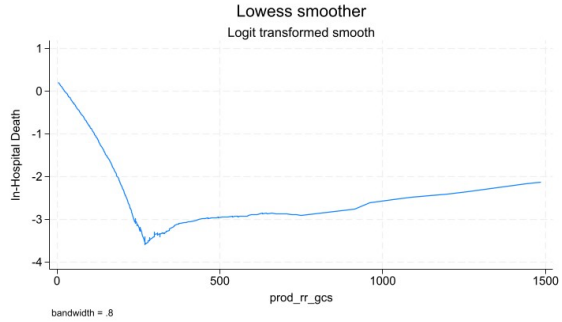
```
. logit died c.age##i.male c.mod_rr_1##i.male c.mod_rr_2##i.male c.mod_gcs##i.male c.mod_sbp_1##i.male c.mod_sbp_2##i.male

Iteration 0:  Log likelihood = -651.19837
Iteration 1:  Log likelihood = -489.43797
Iteration 2:  Log likelihood = -456.34347
Iteration 3:  Log likelihood = -455.57182
Iteration 4:  Log likelihood = -455.56841
Iteration 5:  Log likelihood = -455.56841

Logistic regression                                     Number of obs =  2,789
                                                        LR chi2(13)   = 391.26
                                                        Prob > chi2   = 0.0000
Log likelihood = -455.56841                             Pseudo R2     = 0.3004
```

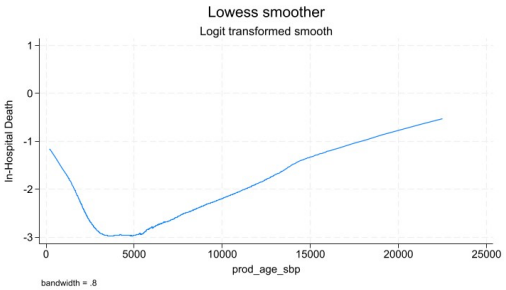
died	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.0459144	.0107271	4.28	0.000	.0248896	.0669391
1.male	-1.179047	5.666103	-0.21	0.835	-12.28441	9.926311
male#c.age						
1	-.0144455	.0121906	-1.18	0.236	-.0383386	.0094476
mod_rr_1	-13.35922	5.247906	-2.55	0.011	-23.64493	-3.073517
male#c.mod_rr_1						
1	5.140468	5.97834	0.86	0.390	-6.576863	16.8578
mod_rr_2	-6.524441	2.981596	-2.19	0.029	-12.36826	-.6806197
male#c.mod_rr_2						
1	2.098667	3.489803	0.60	0.548	-4.741222	8.938555
mod_gcs	-1.201169	.1570449	-7.65	0.000	-1.508972	-.8933671
male#c.mod_gcs						
1	.2767758	.1732458	1.60	0.110	-.0627798	.6163313
mod_sbp_1	-.3197179	.949703	-0.34	0.736	-2.181102	1.541666
male#c.mod_sbp_1						
1	-2.434577	1.077844	-2.26	0.024	-4.547112	-.3220421

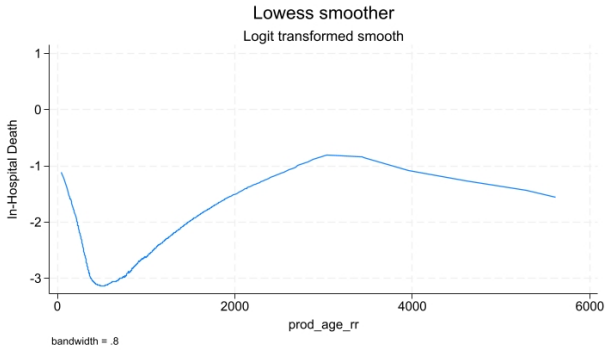
To experiment more with interaction I developed a model with interaction within continuous variables and ran a logit model. We have to check if the interaction terms in this case are linear with the output variable. The same procedure that was previously followed is applied here (i.e - lowess curve for linearity assessment followed by Fractional polynomial test to get the transformed interaction term)

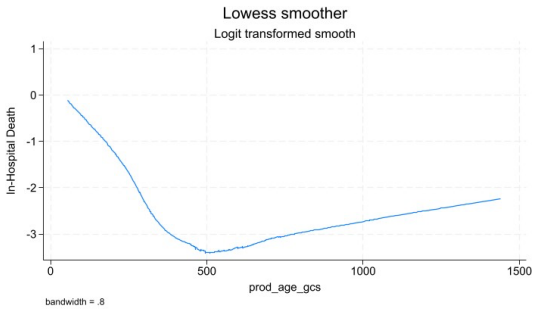












The Fractional Polynomial results were as follows: (for sample only 2 snapshots are attached)

```
. mfp: logit died prod_sbp_rr

Deviance for model with all terms untransformed = 1302.310, 2789 observations

Variable     Model (vs.)   Deviance  Dev diff.   P      Powers   (vs.)
prod_sbp_rr  Lin.   FP2    1302.310    67.206  0.000+  1         -.5 0
             FP1           1252.596    17.491  0.000+  -2
             Final         1235.104                    -.5 0


Transformations of covariates:

-> gen double Iprod__1 = X^-.5-1.924043084 if e(sample)
-> gen double Iprod__2 = ln(X)+1.308857489 if e(sample)
   (where: X = prod_sbp_rr/10000)

Final multivariable fractional polynomial model for died

    Variable |      Initial       |        Final
             |  df    Select Alpha | Status  df  Powers
-------------+--------------------+--------------------
 prod_sbp_rr |   4    1.0000 0.0500 |   in     4  -.5 0

Logistic regression                               Number of obs =  2,789
                                                  LR chi2(2)    =  67.29
                                                  Prob > chi2   =  0.0000
Log likelihood = -617.5522                        Pseudo R2     =  0.0517

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    Iprod__1 |   4.321812   .7014679     6.16   0.000     2.946961    5.696664
    Iprod__2 |   4.379607     .76199     5.75   0.000     2.886134    5.87308
       _cons |  -2.907981   .0872317   -33.34   0.000    -3.078952   -2.73701
------------------------------------------------------------------------------

Deviance = 1235.104.
```

```
. mfp: logit died prod_sbp_gcs

Deviance for model with all terms untransformed = 1043.115, 2789 observations

Variable     Model (vs.)   Deviance  Dev diff.   P      Powers   (vs.)
prod_sbp_gcs Lin.   FP2    1043.115     5.324  0.150   1         -2 1
             Final         1043.115                    1

Transformations of covariates:

-> gen double Iprod__1 = prod_sbp_gcs-1849.477949 if e(sample)

Final multivariable fractional polynomial model for died

    Variable |      Initial       |        Final
             |  df    Select Alpha | Status  df  Powers
-------------+--------------------+--------------------
 prod_sbp_... |   4    1.0000 0.0500 |   in     1  1

Logistic regression                               Number of obs =  2,789
                                                  LR chi2(1)    =  259.28
                                                  Prob > chi2   =  0.0000
Log likelihood = -521.55765                       Pseudo R2     =  0.1991

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    Iprod__1 |  -.0019799   .0001266   -15.64   0.000    -.002228   -.0017318
       _cons |  -3.291057   .1123697   -29.29   0.000    -3.511298   -3.070817
------------------------------------------------------------------------------

Deviance = 1043.115.
```

The final transformed variables for the interaction parameters are

Variable Name	Power Chosen	Transformed Variable
age * sbp	1, 0.5	((age * sbp)/10000)^1 , ((age * sbp)/10000)^0.5
sbp * gcs	1	(sbp * gcs)
rr * gcs	0.5, 0.5	(rr &* gcs)^0.5 , (rr &* gcs)^0.5 * ln((rr*gcs)^0.5)
age * gcs	0.5, 0.5	(age * gcs/1000)^0.5, (age * gcs/1000)^-0.5 * ln(gcs/10)
sbp * rr	0, -0.5	(sbp * rr)^-0.5 * ln(sbp * rr)
age * rr	-0.5, -0.5	((age*rr)/1000)^-0.5, ((age*rr)/1000)^-0.5 * ln((age*rr)/1000)

All interaction terms were added to the normal logit model and in that case we noticed that few of the interaction that predictor gcs has with other variables has very high p values (>0.7) which is clearly indicative of the fact that , they do not have any impact on the outcome. Also a correlation matrix between every predictor variable was done . In that the correlation coefficients of gcs interaction terms and other predictor variables was much more than 0.8 which is again indicative

of collinearity between interaction term of gcs and other predictors. Therefore the gcs interaction terms were removed. As an additional proof LR test for the models with and without gcs interactions terms was caried out which gave us a statistically insignificant result (p=0.419 >>> 0.05). This tells us that the model without gcs interaction terms perform better.

```
. logit died age male mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2 mod_age_rr_1 mod_age_rr_2 mod_age_sbp_1 mod_age_sbp_2 mod_r
> r_gcs_1  mod_rr_gcs_2  mod_age_gcs_1  mod_age_gcs_2 mod_sbp_rr_1

Iteration 0:  Log likelihood = -651.19837
Iteration 1:  Log likelihood = -494.23834
Iteration 2:  Log likelihood = -457.14678
Iteration 3:  Log likelihood = -451.95751
Iteration 4:  Log likelihood = -451.72771
Iteration 5:  Log likelihood =  -451.7259
Iteration 6:  Log likelihood =  -451.7259

Logistic regression                                     Number of obs =  2,789
                                                        LR chi2(16)   =  398.94
                                                        Prob > chi2   =  0.0000
Log likelihood = -451.7259                              Pseudo R2     =  0.3063

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |     .0505039   .0333451     1.51   0.130    -.0148513    .115859
        male |     .4956358   .2262237     2.19   0.028     .0522454    .9390261
      mod_rr_1 |  -55.79139   29.26994    -1.91   0.057    -113.1594    1.576645
      mod_rr_2 |  -16.09535   8.904759    -1.81   0.071    -33.54835    1.357661
      mod_gcs |   -.8919257   .1995207    -4.47   0.000    -1.282979   -.5008723
    mod_sbp_1 |   -4.734444   1.397632    -3.39   0.001    -7.473752   -1.995136
    mod_sbp_2 |    4.166488   1.191322     3.50   0.000     1.83154    6.501437
  mod_age_rr_1 |   29.06126   14.20389     2.05   0.041     1.222148   56.90038
  mod_age_rr_2 |   6.639845   3.378913     1.97   0.049     .0172962   13.26239
 mod_age_sbp_1 |  -11.78114   4.948892    -2.38   0.017    -21.48079   -2.081495
 mod_age_sbp_2 |   37.49722   14.26648     2.63   0.009     9.535439   65.45901
 mod_rr_gcs_1 |   -.2097227   1.678259    -0.12   0.901    -3.499049   3.079604
 mod_rr_gcs_2 |    .0283555   .1787728     0.16   0.874    -.3220327   .3787438
mod_age_gcs_1 |   -2.396893   12.8828     -0.19   0.852    -27.64671   22.85293
mod_age_gcs_2 |   1.751749   4.498665     0.39   0.697    -7.065473   10.56897
 mod_sbp_rr_1 |   5776.348   2282.275     2.53   0.011     1303.172   10249.52
        _cons |  -7.544545   12.70353    -0.59   0.553      -32.443   17.35391
------------------------------------------------------------------------------
Note: 0 failures and 1 success completely determined.

. est store full
```

```
. logit died age male mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2 mod_age_rr_1 mod_age_rr_2 mod_age_sbp_1 mod_age_sbp_2 mod_s
> bp_rr_1

Iteration 0:  Log likelihood = -651.19837
Iteration 1:  Log likelihood =  -486.8899
Iteration 2:  Log likelihood = -456.72774
Iteration 3:  Log likelihood = -454.27526
Iteration 4:  Log likelihood = -453.69943
Iteration 5:  Log likelihood = -453.67585
Iteration 6:  Log likelihood =  -453.6758
Iteration 7:  Log likelihood =  -453.6758

Logistic regression                                     Number of obs =  2,789
                                                        LR chi2(12)   =  395.05
                                                        Prob > chi2   =  0.0000
Log likelihood = -453.6758                              Pseudo R2     =  0.3033

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |     .067761   .029701     2.28   0.023     .0095481    .1259739
        male |     .4844439   .2263116     2.14   0.032     .0408813    .9280066
      mod_rr_1 |  -59.78076   17.50549    -3.41   0.001    -94.09089   -25.47063
      mod_rr_2 |  -17.62064   5.070028    -3.48   0.001    -27.55771   -7.683563
      mod_gcs |   -.9805284   .066445    -14.76   0.000    -1.110758   -.8502985
    mod_sbp_1 |   -4.722141   1.400139    -3.37   0.001    -7.466362   -1.97792
    mod_sbp_2 |    4.173876   1.194574     3.49   0.000     1.832553    6.515199
  mod_age_rr_1 |   31.48973   10.91359     2.89   0.004     10.09948    52.87998
  mod_age_rr_2 |    7.13405   2.669254     2.67   0.008     1.902407    12.36569
 mod_age_sbp_1 |  -11.86057   4.972864    -2.39   0.017    -21.6072   -2.113931
 mod_age_sbp_2 |   37.3381   14.33833     2.60   0.009     9.235492    65.4407
 mod_sbp_rr_1 |    5717.3   2287.586     2.50   0.012     1233.713    10200.89
        _cons |  -10.10144   6.981454    -1.45   0.148    -23.78484    3.581954
------------------------------------------------------------------------------
Note: 0 failures and 1 success completely determined.

. lrtest full

Likelihood-ratio test
Assumption: . nested within full

 LR chi2(4) =   3.90
Prob > chi2 = 0.4197
```

Having a model with both types of interactions (i.e continuous- continuous interactions and continuous – categorical interactions) will not be feasible as its will increase the number of model parameters therefore the model complexity will increase and it breaks the requirements of a parsimonious model. The BIC value for such a model will be very high which will eventually affect the quality of the prediction model. We have three models of interest : A simple model with no interaction terms, A model with only continuous – categorical interactions and a model with continuous- continuous interactions. It is always better to have interactions terms in a model as they help us to model complex relationships and improve the fit of the model. So we will have to choose between the two models with interactions terms. Since both the models have the same degrees of freedom we cannot carry on an LR test. So in this scenario we will have to conduct AIC-BIC comparative model test and choose the model with lower AIC and BIC. Our aim is to define an apt ”prediction model”. AIC-BIC is the best way to choose the best model as they provide a way to compare different models based on goodness of fit and the number of parameters which prevents overfitting (by penalizing for extra complexity and increasing BIC) and select the model that generalizes well to new data. The command **estat ic** was used to get the AIC-BIC values. The snap in the left is for the categorical-continuous variable interaction and on the right is for Continuous- continuous variable interaction. Since the values on the right are lesser (though by a smaller amount) we therefore choose the **logit model with transformed main terms and transformed continuous-continuous interaction terms.**

```
. estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+---------------------------------------------------------------
           . |      2,789  -651.1984  -455.5684      14   939.1368   1022.205
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] IC note.

.
```

```
Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+---------------------------------------------------------------
           . |      2,789  -651.1984  -453.1258      14   934.2515   1017.32
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] IC note.

.
```

Q3) Provide an appropriate test of goodness of fit. Interpret the result.

To choose the best model among multiple model options comparative mode fit tests like Likelihood ratio tests and AIC-BIC tests were done and the best model was chosen

For the best model, first, a normal Pearson chi-sq goodness of fit test is conducted using the command **estat gof**. The results might have a p value much greater than 0.05 but the number of

covariate patterns and number of observations are almost equal. In such cases Pearsons goodness of fit test is not appropriate. (snapshot in the left)

So the most appropriate test is Hosmer Lemeshow goodness of fit. We would now have to group the variables , into groups of 20 (in general 10 is chosen) and carry out a goodness of fit test using the command **estat gof, group (n)**.This test is called Hosmer Lemeshow goodness of fit test. We chose 20 in this case as it fairly improved the p value and the dataset that we are dealing with is large (approximately 2800 data points). To improve the sensitivity of the model and provide more granular information about the models fit in cases of large data sets more groups are required (but preferably within 20 groups). The p value in this case is 0.077 (>0.05) which fails to reject the null hypothesis of the goodness of fit test. Therefore the model chosen does not depart from the state of good fit. However, P value can be much better than what it is, for which the model can be improved and refit by carrying out few model diagnostic tests. (snapshot in the right side)

```
. estat gof

Goodness-of-fit test after logistic model
Variable: died

      Number of observations =    2,789
Number of covariate patterns =    2,751
          Pearson chi2(2738) = 2330.56
                Prob > chi2 =  1.0000
```

```
. estat gof, group(20)
note: obs collapsed on 20 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
Variable: died

  Number of observations =  2,789
        Number of groups =     20
Hosmer-Lemeshow chi2(18) =  27.08
                Prob > chi2 = 0.0775
```

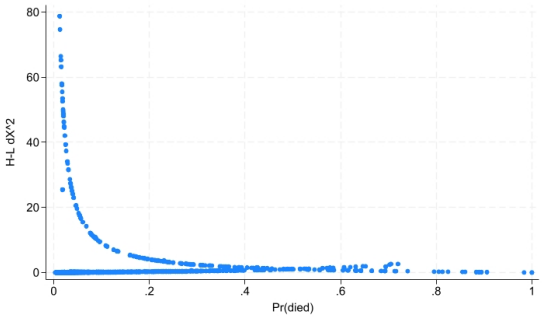
Q4) Complete model diagnostics and re-fit your model if needed. Explain any issues you might find in this process and how you will deal with it in your model

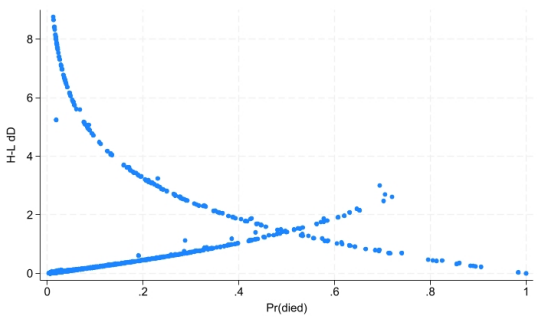
The Pictures included for model diagnostics are as follows (1. Pearson chi square vs Prob , 2. Change is deviance vs Prob, 3.Cooks distance vs Probability 4.statements to filter outliers 5. Re-fit model)

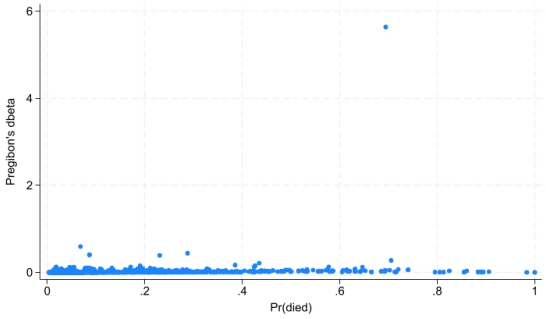
Model diagnostics is the process of examining the fitted model for outliers, influential points using parameters such as Cooks distance , deviance and pearson chi square value. Following are the problems that were noticed when the diagnostics graphs were plotted

- The problem that I observed in pearson chi square value vs Probability graph is that the graphs is like an exponential-like curve which is indicative of the fact that there are some data points that have a very large influence on the overall model fit and the number of points with difference in chi square >10 is high (i.e the part the exponential curve starts)
- With respect to the Graph of Difference in deviance and predicted probability , the problem is similar. The graph here is again exponential which shows a lot of data points have a difference of deviance greater than 4 which indicates that the values are highly varying therefore influncential.
- When we look at the graph of Dbeta (cooks distance) Vs probability we can see a bunch of points having debeta >1 with one in the right corner of the graph which surely means the value is highly influential. A set of filters (in terms of statements) should be defined to omit such values.

For omission of data points the rule says that if the difference in Pearson chi square is greater than 10 , difference in cooks distance is greater than 1 and change if deviance is greater than 4 then the point is considered to be highly influential. Following code was run the model was re-fit. The p value of Hosmer Lemeshow GOF test for the re-fit model was much higher (0.70) than the original model (0.07) which shows that, the process of omitting influential points is right.








```
gen exclude_dx2 = (delta_x2 > 10)

gen exclude_dbeta = (dbeta > 1)

gen exclude_ddev = (delta_dev > 4)
```

```
. logit died age male mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2 mod_age_rr_1 mod_age_rr_2 mod_age_sbp_1 mod_age_sbp_2 mod_s
> bp_rr_1 if exclude_dx2 == 0 & exclude_dbeta == 0 & exclude_ddev == 0

Iteration 0:  Log likelihood =   -479.882
Iteration 1:  Log likelihood = -308.71629
Iteration 2:  Log likelihood = -230.93735
Iteration 3:  Log likelihood = -216.87459
Iteration 4:  Log likelihood = -215.77907
Iteration 5:  Log likelihood = -215.56197
Iteration 6:  Log likelihood = -215.55923
Iteration 7:  Log likelihood = -215.55923

Logistic regression                                     Number of obs =  2,730
                                                        LR chi2(12)   = 528.65
                                                        Prob > chi2   = 0.0000
Log likelihood = -215.55923                             Pseudo R2     = 0.5508

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .1806412   .0420815     4.29   0.000      .098163    .2631194
        male |    .525162   .3073328     1.71   0.087    -.0771992    1.127523
    mod_rr_1 |  -110.0164   24.69617    -4.45   0.000      -158.42   -61.61281
    mod_rr_2 |   -30.5259     7.2049    -4.24   0.000    -44.64724   -16.40455
     mod_gcs |   -1.79411   .1429652   -12.55   0.000    -2.074317   -1.513903
   mod_sbp_1 |  -7.984502   2.705824    -2.95   0.003    -13.28782   -2.681183
   mod_sbp_2 |   6.861753   2.186058     3.14   0.002     2.577158    11.14635
mod_age_rr_1 |   67.28842   16.27755     4.13   0.000     35.38501    99.19182
mod_age_rr_2 |    14.8669   3.897121     3.81   0.000     7.228684    22.50512
mod_age_sbp_1 |  -23.3832   7.435368    -3.14   0.002    -37.95625   -8.810143
mod_age_sbp_2 |   70.87315   22.25727     3.18   0.001     27.24971    114.4966
 mod_sbp_rr_1 |   8500.786   2988.509     2.84   0.004     2643.416    14358.16
        _cons |  -29.35276   10.23022    -2.87   0.004    -49.40363   -9.301892
------------------------------------------------------------------------------

Note: 0 failures and 2 successes completely determined.
```

Q5) Provide a publication-quality table reporting your resulting model (variables, beta (SE), p-value).

Following table included the beta(SE) and p values of the model that was re-fit after model diagnostics was done.

Table 2: Beta(SE), P value of the Re-fit logit model

Variable	Beta (SE)	P Value
age	0.18 (0.04)	0.000 ***
male	0.52 (0.31)	0.087
rr (1)	-110.02 (24.70)	0.000 ***
rr (2)	-30.53 (7.20)	0.000 ***
gcs	-1.79 (0.14)	0.000 ***
sbp (1)	-7.98 (2.71)	0.003 **
sbp (2)	6.86 (2.19)	0.002 **
age * rr (1)	67.29 (16.28)	0.000 ***
age * rr (2)	14.87 (3.90)	0.000 ***
age * sbp (1)	-23.38 (7.43)	0.002 **
age * sbp (2)	70.87 (22.26)	0.001 **
sbp * rr (1)	8500.79 (2988.51)	0.004 **
_cons	-29.35 (10.23)	0.005 **

Note :
1.sbp = Systolic Blood Pressure (mmHg), rr = Respiratory Rate (breaths per minute), gcs = Glasgow Coma Scale
2. Interaction terms between continuous terms is indicated using "" as a multiplicative operator*
3. All predictors here are variables that are transformed using powers provided by Frac poly
4. (1) indicates first transformation and (2) indicates second transformation
*5. star convention: * : p <0.05 , ** : p <0.01, *** : p <0.001*

Q6) Provide the model formula for predicting the probability of dying in the hospital.

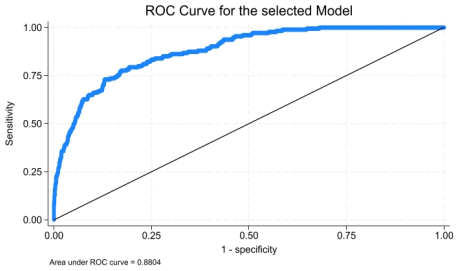
Probability of dying : P(dying) : π_{dying}

Logit (π_{dying}) = ln (π_{dying} / 1- π_{dying})
= -29.35 + 0.18 * X_{age} + 0.52 * X_{male} - 110.02 * $X_{rr(1)}$ - 30.53 * $X_{rr(2)}$ - 1.79 * X_{gcs} - 7.98 * $X_{sbp(1)}$ + 6.86 * $X_{sbp(2)}$ + 67.29 * $X_{age*rr(1)}$ + 14.87 * $X_{age*rr(2)}$ - 23.38 * $X_{age*sbp(1)}$ + 70.87 * $X_{age*sbp(2)}$ + 8500.79 * $X_{sbp*rr(1)}$

Where the transformed variables are:
X_{age} = (age)^1, $X_{rr(1)}$ = (rr/10)^-0.5, $X_{rr(2)}$ = (rr/10)^-0.5 * ln(rr/10), X_{gcs} = (gcs/10)^3,
$X_{sbp(1)}$ = (sbp/100)^2, $X_{sbp(2)}$ = (sbp/100)^2 * ln(sbp/100), $X_{age*rr(1)}$ = ((age*rr)/1000)^-0.5,
$X_{age*rr(2)}$ = ((age*rr)/1000)^-0.5 * ln((age*rr)/1000), $X_{age*sbp(1)}$ = ((age * sbp)/10000)^1 ,
$X_{age*sbp(2)}$ = ((age * sbp)/10000)^0.5, $X_{sbp*rr(1)}$ = (sbp * rr)^-0.5 * ln(sbp * rr)

Q7) Provide an ROC curve, an estimate of the area under the ROC curve (AROC, with 95% CI) and a classification table for your model.

ROC Curve



AROC Value

```
. roctab died prob_p
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
2,789	0.8675	0.0144	0.83930	0.89567

Classification report

```
. estat classification

Logistic model for died
```

Classified	True D	True ~D	Total
+	43	22	65
-	73	2592	2665
Total	116	2614	2730

```
Classified + if predicted Pr(D) >= .5
True D defined as died != 0
```

Sensitivity	Pr(+| D)	37.07%
Specificity	Pr(-|~D)	99.16%
Positive predictive value	Pr(D| +)	66.15%
Negative predictive value	Pr(~D| -)	97.26%
False + rate for true ~D	Pr(+|~D)	0.84%
False - rate for true D	Pr(-| D)	62.93%
False + rate for classified +	Pr(~D| +)	33.85%
False - rate for classified -	Pr(D| -)	2.74%
Correctly classified		96.52%

Q8) Evaluate the usefulness of your predictive model in the validation sample (val=1). Write a short conclusion.

```
. logit died age male mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2 mod_age_rr_1
>  mod_age_rr_2 mod_age_sbp_1 mod_age_sbp_2 mod_sbp_rr_1

Iteration 0:  Log likelihood = -157.30117
Iteration 1:  Log likelihood = -117.81565
Iteration 2:  Log likelihood = -109.82015
Iteration 3:  Log likelihood = -109.55427
Iteration 4:  Log likelihood = -109.55398
Iteration 5:  Log likelihood = -109.55398

Logistic regression                             Number of obs =    675
                                                LR chi2(12)   =  95.49
                                                Prob > chi2   = 0.0000
Log likelihood = -109.55398                     Pseudo R2     = 0.3035
```

died	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	-.0037303	.0643913	-0.06	0.954	-.1299349	.1224743
male	-.0098523	.4385611	-0.02	0.982	-.8694162	.8497116
mod_rr_1	26.34743	41.84456	0.63	0.529	-55.6664	108.3612
mod_rr_2	4.271622	10.86037	0.39	0.694	-17.01431	25.55756
mod_gcs	-1.099734	.135616	-8.11	0.000	-1.365536	-.8339314
mod_sbp_1	2.075123	4.895316	0.42	0.672	-7.519519	11.66977
mod_sbp_2	-1.70598	4.025287	-0.42	0.672	-9.595397	6.183437
mod_age_rr_1	-17.30456	28.74834	-0.60	0.547	-73.65027	39.04114
mod_age_rr_2	-3.016136	6.520199	-0.46	0.644	-15.79549	9.763219
mod_age_sbp_1	13.11412	14.16494	0.93	0.355	-14.64865	40.87688
mod_age_sbp_2	-32.16633	41.28225	-0.78	0.436	-113.078	48.74538
mod_sbp_rr_1	-2602.405	3705.732	-0.70	0.483	-9865.507	4660.697
_cons	15.41016	18.42509	0.84	0.403	-20.70235	51.52268

```
. predict prob_p_val, pr

. roctab died prob_p_val
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
675	0.8532	0.0364	0.78175	0.92456

```
. estat classification

Logistic model for died
```

Classified	True D	True ~D	Total
+	13	7	20
-	29	626	655
Total	42	633	675

```
Classified + if predicted Pr(D) >= .5
True D defined as died != 0
```

Sensitivity	Pr(+| D)	30.95%
Specificity	Pr(-|~D)	98.89%
Positive predictive value	Pr(D| +)	65.00%
Negative predictive value	Pr(~D| -)	95.57%
False + rate for true ~D	Pr(+|~D)	1.11%
False - rate for true D	Pr(-| D)	69.05%
False + rate for classified +	Pr(~D| +)	35.00%
False - rate for classified -	Pr(D| -)	4.43%
Correctly classified		94.67%

```
. estat gof, group(20)
note: obs collapsed on 20 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
Variable: died

  Number of observations =    675
        Number of groups =     20
Hosmer-Lemeshow chi2(18) =  12.96
            Prob > chi2 = 0.7937
```

The usefulness of the model can be evaluated by parameters such as prob of the entire model along with LR chi sq, Hosmer Lemeshow goodness of fit test, AROC value, classification report with values of Sensitivity, Specificity, Accuracy and pesudo R square value.

- The probability of the entire model is 0.00 with an LR chi sq value of 95.49 for 12 degrees of freedom. The critical chi sq value for 12 degrees of freedom and 0.05 level of significance is 21.026 which is much lesser than the test statistics value of 95.49. This indicates that the chi square test is statistically significant and the null hypothesis is rejected. Therefore at least one of the predictors add significance to the model's fit.
- Hosmer Lemeshow goodness of fit test converts the covariates to 20 groups in this case and a statistical test is carried out. The p value in our case is 0.79 which is much greater than 0.05. This fails to reject the null hypothesis. Therefore we come to a conclusion that the model does not depart from the state of good fit and there exists a proper model fit.
- For this model the Area under the ROC is 0.853 which is greater than 0.5. This model shows great ability to discriminate and therefore classify patients who died and those who survived. More evaluation terms from the classification report is needed to evaluate the veracity of the model.
- The sensitivity of the model is 30.95% which is low for a model in a healthcare setting as it does not classify a significant portion of data as "died" even if it actually considered to be a part of class died=1. On the other hand the sensitivity of the data is 98.9% which means the model is completely perfect in predicting people who will survive (died =0). This model will never incorrectly predict someone dead when they actually survive. The accuracy in predicting the right results is 94.67% which is appreciable and considerable. The positive prediction value is around 65% which means the whenever model predicts death there is 65% chance that the person actually died and the negative prediction value which is round 98% means whenever model predicts survival the chances of a patient actually surviving is 98%. Therefore this model is much better to predict survival than death.

Q9) Write a paragraph on your statistical methods, including model development, for the investigator to use in the manuscript. This should be publication quality text.

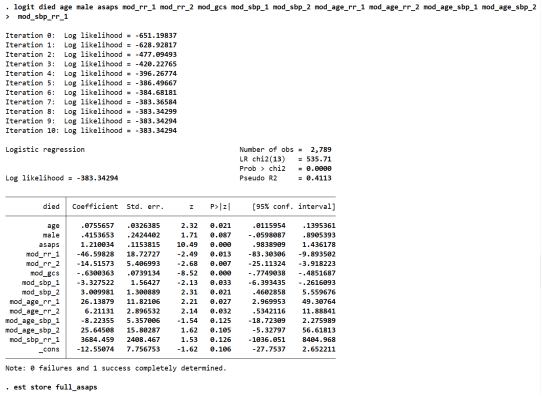
The following case study utilizes a multi-covariate binary logistic regression model to evaluate the association between continuous, categorical predictor variables and binary outcome variables. Before the model was designed, an extensive linearity assessment was done for the continuous variables using a logit transformed lowess curve. A frac poly test was carried out to handle the case of non-linearity. This was used to determine the powers to which the predictors can be raised for transformation. Interaction terms were also introduced into the simple model to model the underlying complexity. Since there are multiple models for consideration, to choose the best model, comparative model tests like the LR test and AIC-BIC tests were conducted, and the best model was chosen based on significant p values in the LR test and low AIC and BIC values. Model diagnostics were also done for the final model to remove outliers and influential data points using values of parameters such as $\Delta\chi^2$, Cook's distance, and difference in deviance. Based on a threshold value of all parameters, we neglected those data points that were creating a negative impact, and the model was then refitted. Hosmer Lemeshow's goodness of fit test was chosen for

the final GOF test because the number of observations and covariate groups/patterns were almost equal. The discriminatory ability of the final model was evaluated using the area under the receiver operating characteristic (ROC) curve. A classification table and sensitivity-specificity analysis were finally done to validate further and elucidate the chosen model.

Q10) Investigators hypothesized they could better predict in-hospital mortality in this populati**on by adding a measure of fitness for surgery assessed in the ER prior to surgery. This measure is "asaps" in the dataset. Again using just, the val=0 sample:**

a. Decide if the new asaps measure would be appropriate to add to your predictive model that you developed above. Do NOT modify your predictive model above (other than possibly adding the new asaps measure. Attend to proper modeling of the variable.

Modeling of variable (screenshots not attached): The Lowess curve and Frac Poly test was carried out for the variable asaps. The curve was almost linear but it was not a good choice to assume full linearity. A frac poly test was carried out which suggested m=1 model with power "1" to be the best transformation. Therefore, this variable need not be transformed / modeled. Deciding if asaps would be appropriate or not: Run a model with asaps as a predictor along with the other existing transformed predictors and a model without asaps. LR test was then carried out with both the models and the p value for the test was found. Since p =0.000<0.05, We can reject the null hypothesis and come to a conclusion that adding asaps has significantly improved the fit and prediction of the model. Therefore, it is appropriate to add asaps measure to existing model



```
. logit died age male asaps mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2 mod_age_rr_1 mod_age_rr_2 mod_age_sbp_1 mod_age_sbp_2
>  mod_sbp_rr_1

Iteration 0:  Log likelihood = -651.19837
Iteration 1:  Log likelihood = -628.92817
Iteration 2:  Log likelihood = -477.09493
Iteration 3:  Log likelihood = -420.22765
Iteration 4:  Log likelihood = -396.26774
Iteration 5:  Log likelihood = -386.49667
Iteration 6:  Log likelihood = -384.68181
Iteration 7:  Log likelihood = -383.36584
Iteration 8:  Log likelihood = -383.34299
Iteration 9:  Log likelihood = -383.34294
Iteration 10: Log likelihood = -383.34294

Logistic regression                                     Number of obs =  2,789
                                                        LR chi2(13)   =  535.71
                                                        Prob > chi2   = 0.0000
Log likelihood = -383.34294                             Pseudo R2     = 0.4113

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0755657   .0326385     2.32   0.021     .0115954    .1395361
        male |   .4153653   .2424402     1.71   0.087    -.0598087    .8905393
       asaps |   1.210034   .1153815    10.49   0.000     .9838909    1.436178
      mod_rr_1 |  -46.59828   18.72727    -2.49   0.013    -83.30306   -9.893502
      mod_rr_2 |  -14.51573   5.406993    -2.68   0.007    -25.11324   -3.918223
     mod_gcs |  -.6300363   .0739134    -8.52   0.000    -.7749038   -.4851687
    mod_sbp_1 |  -3.327522    1.56427    -2.13   0.033    -6.393435   -.2616093
    mod_sbp_2 |   3.009981   1.300889     2.31   0.021     .4602858    5.559676
mod_age_rr_1 |   26.13879   11.82106     2.21   0.027     2.969953    49.30764
mod_age_rr_2 |    6.21131   2.896532     2.14   0.032     .5342116    11.88841
mod_age_sbp_1 |  -8.22355   5.357006    -1.54   0.125    -18.72309    2.275989
mod_age_sbp_2 |  25.64508   15.80287     1.62   0.105    -5.32797    56.61813
 mod_sbp_rr_1 |   3684.459   2408.467     1.53   0.126    -1036.051    8404.968
        _cons |  -12.55074   7.756753    -1.62   0.106    -27.7537    2.652211
------------------------------------------------------------------------------
Note: 0 failures and 1 success completely determined.

. est store full_asaps
```

```
. logit died age male mod_rr_1 mod_rr_2 mod_gcs mod_sbp_1 mod_sbp_2 mod_age_rr_1 mod_age_rr_2 mod_age_sbp_1 mod_age_sbp_2 mod_s
> bp_rr_1

Iteration 0:  Log likelihood = -651.19837
Iteration 1:  Log likelihood =  -486.8899
Iteration 2:  Log likelihood = -456.72774
Iteration 3:  Log likelihood = -454.27526
Iteration 4:  Log likelihood = -453.69943
Iteration 5:  Log likelihood = -453.67585
Iteration 6:  Log likelihood =  -453.6758
Iteration 7:  Log likelihood =  -453.6758

Logistic regression                                     Number of obs =  2,789
                                                        LR chi2(12)   =  395.05
                                                        Prob > chi2   = 0.0000
Log likelihood = -453.6758                              Pseudo R2     = 0.3033

------------------------------------------------------------------------------
        died | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |    .067761   .029701     2.28   0.023     .0095481    .1259739
        male |   .4844439   .2263116     2.14   0.032     .0408813    .9280066
      mod_rr_1 |  -59.78076   17.58549    -3.41   0.001    -94.89089   -25.47063
      mod_rr_2 |  -17.62064   5.070028    -3.48   0.001    -27.55771   -7.683563
     mod_gcs |  -.9805284   .066445   -14.76   0.000    -1.110758   -.8502985
    mod_sbp_1 |  -4.722141   1.400139    -3.37   0.001    -7.466362   -1.97792
    mod_sbp_2 |   4.173876   1.194574     3.49   0.000     1.832553    6.515199
mod_age_rr_1 |   31.48973   10.91359     2.89   0.004     10.09948    52.87998
mod_age_rr_2 |    7.13405   2.669254     2.67   0.008     1.902407    12.36569
mod_age_sbp_1 |  -11.86057   4.972864    -2.39   0.017    -21.6072   -2.113931
mod_age_sbp_2 |   37.3381   14.33833     2.60   0.009     9.235492    65.4407
 mod_sbp_rr_1 |    5717.3   2287.586     2.50   0.012     1233.713    10200.89
        _cons |  -10.10144   6.981454    -1.45   0.148    -23.78484    3.581954
------------------------------------------------------------------------------
Note: 0 failures and 1 success completely determined.

. lrtest full_asaps

Likelihood-ratio test
Assumption: . nested within full_asaps

 LR chi2(1) = 140.67
Prob > chi2 = 0.0000
```

b. Assuming you have found that the new measure adds to your original prediction model, test whether addition of asaps adds significant prediction to the model in terms of the AROC (i.e., compare your full prediction model (with asaps) to your original prediction model).

In terms of AROC the addition of asaps can be statistically tested using using DeLongs test. The command **roccomp** is used in stata. According to DeLongs test, the null hypothesis is defined as: “There is no difference between the AROCs of the curves that are obtained when the model is run with and without the variable asaps”. This test calculates the Z statistics value (Difference in AROCs / Root of variance of the difference). A p-value will then be computed using the normal distribution to determine the statistical significance of the difference. The usual pattern of testing is then followed to check if the AROCs are the same or not between the two curves. The P value in this case is 0.0001 which is <<<0.05. Since it shows high level of significance, there exists a significant difference in the AROC where the Area of the curve for the model with asaps is more than the one without AROC (as the difference is always between full and reduced model). So, the model with asaps is better in prediction as per DeLongs Method.

```
.
. roccomp died prob_p_org prob_p_asaps

                              ROC                    Asymptotic normal
                    Obs       area     Std. err.      [95% conf. interval]
-------------------------------------------------------------------------
prob_p_org        2,789     0.8804      0.0124        0.85608     0.90474
prob_p_asaps      2,789     0.9233      0.0105        0.90260     0.94391
-------------------------------------------------------------------------
H0: area(prob_p_org) = area(prob_p_asaps)
    chi2(1) =    15.58       Prob>chi2 =   0.0001
```

c. Write a short conclusion, comparing your original prediction model to the enhanced prediction model.

Parameter	Enhanced model	Original Model
Area under ROC	0.923	0.867
AIC	794.69	934.25
BIC	877.75	1017.23

While discussing about the veracity of a logit model (prediction model) AROC, AIC, BIC are very important because:

- AIC and BIC evaluate model fit appropriately and helps in comparative statistics (i.e choosing the best model among 2 or more available options) while penalizing complexity (imposing penalty when more insignificant parameters are added) to prevent overfitting and maintain “parsimony”. Lower values of AIC and BIC indicate a better model.
- AROC measures the model's ability to classify and discriminate between the levels of outcome variables, with values closer to 1 indicating apt distinction between died=0 and died=1. Greater the AROC better the predictive model.

Based on these pointers, Since the AIC & BIC are lesser and AROC is more for the enhanced model we choose the model with asaps (enhanced model) over the original model without asaps.