Student – Htun Teza
ID – G6238135
Course – RACE 626

I. Data Preprocessing

After summation, we see there are 10 missing values in two separate variables, also with some impossible values. We were instructed not to drop samples but replace them with missing value ".".

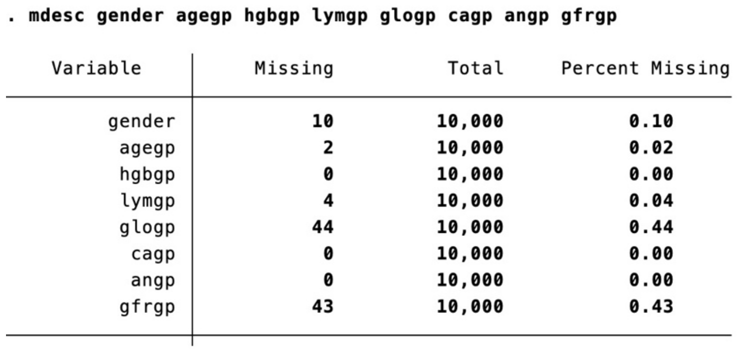
1. For age, the samples range from 0 to 106 years. I will consider 0 year old as impossible and replace with missing value. This action creates 2 missing values.

2. I will encode variable 'sex' into new dichotomous '0,1' variable 'gender', while not using the supplied value '_Isex_2', since the name later conflicts while fitting the final model. The variable 'sex' have 10 missing values from the beginning.

3. For lymphocyte, the samples range from 0 to 588.3 (x 10^9 per liter). While lymphocytopenia (abnormal low lymphocyte count, extremely low in case of AIDS or malnutrition) is possible, I will consider lymphocyte count '0' as impossible and replace it with missing value. This action creates 4 missing values.

4. For globulin, the samples range from 13 to 114 gram per liter. The normal range for globulin is 23 to 35 g/L. I will consider values above '70 g/L' as impossible and replace them with missing value. This action creates 44 missing values.

5. For estimated glomerular filtration rate, the samples range from 3.247472 to 1254.995 milliliter per min per 1.72 meter squared. The normal range for glomerular filtration rate is 80 to 120 mL/min/1.73m^2. I will consider values above '250 mL/min/1.73m^2' as impossible and replace them with missing value. This action creates 33 missing values. The variable 'gfr2' have 10 missing values from the beginning. This results in 43 missing values.



```
. mdesc gender agegp hgbgp lymgp glogp cagp angp gfrgp

    Variable    |     Missing          Total     Percent Missing
----------------+-----------------------------------------------
         gender |          10         10,000           0.10
          agegp |           2         10,000           0.02
          hgbgp |           0         10,000           0.00
          lymgp |           4         10,000           0.04
          glogp |          44         10,000           0.44
           cagp |           0         10,000           0.00
           angp |           0         10,000           0.00
          gfrgp |          43         10,000           0.43
----------------+-----------------------------------------------
```

These actions reduce the total sample size from 10,000 to 9,908, while counting the samples with constraint ‘no missing values in any column’.

Table 1 : Distribution of Patient Characteristics between SPE/UPE positive and negative groups

Factors	Group		P value
	SPE/UPE (-) n = 9191 (91.91%)	SPE/ UPE (+) n = 809 (8.09%)	
Age			<0.001
< 60	4,482 (48.78%)	156 (19.28%)	
≥ 60	4,707 (51.23%)	653 (80.72%)	
Gender			<0.001
Female	5,661 (61.64%)	397 (49.26%)	
Male	3,523 (38.36%)	409 (50.74%)	
Hemoglobin (g/L)			<0.001
< 100	721 (7.84%)	101 (12.48%)	
≥ 100	8,470 (92.16%)	708 (87.52%)	
Lymphocyte (x 10^9/L)			<0.001
< 1.9	4,826 (52.53%)	600 (74.17%)	
≥ 1.9	4,361 (47.47%)	209 (25.83%)	
Globulin (g/L)			<0.001
< 38	5,220 (56.91%)	198 (25.26%)	
≥ 38	3,952 (43.09%)	586 (74.74%)	
Ionized Calcium (mmol/L)			0.431
< 1.35	8,791 (95.65%)	769 (95.06%)	
≥ 1.35	400 (4.35%)	40 (4.94%)	
Anion gap (mmol/L)			0.164
< 10	3,476 (37.82%)	326 (40.30%)	
≥ 10	5,715 (62.18%)	483 (59.70%)	
GFR (mL/min/1.73m2)			<0.001
< 60	2,505 (27.37%)	446 (55.33%)	
≥ 60	6,646 (72.63%)	360 (44.67%)	

II. Univariate Logistic Regression

Table 2 : Factors associated with SPE/UPE: Univariate Logistic Regression Analysis

Factors	OR	95% CI	P value
Age			
< 60	1		
≥ 60	3.99	3.33 - 4.77	<0.001
Gender			
Female	1		
Male	1.66	1.43 – 1.91	<0.001
Hemoglobin (g/L)			
< 100	1.68	1.34 – 2.09	<0.001
≥ 100	1		
Lymphocyte (x 10^9/L)			
< 1.9	2.59	2.20 – 3.05	<0.001
≥ 1.9	1		
Globulin (g/L)			
< 38	1		
≥ 38	3.91	3.31 - 4.62	<0.001
Ionized Calcium (mmol/L)			
< 1.35	1		
≥ 1.35	1.14	0.82 – 1.60	0.431
Anion gap (mmol/L)			
< 10	1.11	0.96 – 1.29	0.164
≥ 10	1		
GFR (mL/min/1.73m2)			
< 60	3.29	2.84 – 3.80	<0.001
≥ 60	1		

Interpretation for Univariate Logistic Regression

1. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects older than 60 years of age is 3.99 times higher than odds of those younger than 60.
2. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for male subjects is 3.99 times higher than odds of female.
3. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with hemoglobin less than 100 g/L is 1.68 times higher than odds of those with hemoglobin equal or more than 100 g/L.
4. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with lymphocyte less than 1.9 x 10^9/L is 2.59 times higher than odds of those with lymphocyte equal or more than 1.9 x 10^9/L.

5. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with globulin equal or more than 38 g/L is 3.91 times higher than odds of those with globulin less than 38 g/L.
6. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with ionized calcium equal or more than 1.35 mmol/L is 1.14 times higher than odds of those with ionized calcium less than 1.35 mmol/L.
7. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with anion gap less than 10 mmol/L is 1.11 times higher than odds of those with anion gap equal or more than 10 mmol/L.
8. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with GFR less than 60 mL/min/1.73m^2 is 1.11 times higher than odds of those with GFR equal or more than 60 mL/min/1.73m^2.

III. Multivariate Logistic Regression

A. Forward selection

Step 0: Start with empty model.

Univariate analysis with constraints applied

Factors	LR test	P-value
Age	274.49	<0.001
Gender	47.54	<0.001
Hemoglobin (g/L)	11.34	<0.001
Lymphocyte (x 10^9/L)	138.36	<0.001
Globulin (g/L)	296.27	<0.001
Ionized Calcium (mmol/L)	0.52	0.4692
Anion gap (mmol/L)	0.62	0.4297
GFR (mL/min/1.73m2)	241.85	<0.001

If the cut-off point is p=0.15, the variables we should consider Age, Gender, Hemoglobin, Lymphocyte, Globulin, and GFR.

The first predictor selected is 'Globulin'.

Step 1: Include 'Globulin'.

Factors	Est store	LR test	P- value
Age	B	214.72	<0.001
Gender	C	44.50	<0.001
Hemoglobin (g/L)	D	1.86	0.1722
Lymphocyte (x 10^9/L)	E	123.52	<0.001
GFR (mL/min/1.73m2)	F	158.73	<0.001

The second predictor selected is 'Age'.

Step 2: Include 'Globulin' and 'Age'.

Factors	Est store	LR test	P- value
Gender	C2	26.04	<0.001
Hemoglobin (g/L)	D2	0.45	0.5043
Lymphocyte (x 10^9/L)	E2	87.07	<0.001
GFR (mL/min/1.73m2)	F2	60.14	<0.001

The third predictor selected is 'Lymphocyte'.

Step 3: Include 'Globulin', 'Age' and 'Lymphocyte'.

Factors	Est store	LR test	P- value
Gender	C3	20.27	<0.001
Hemoglobin (g/L)	D3	3.46	3.46
GFR (mL/min/1.73m2)	F3	52.93	<0.001

The fourth predictor selected is 'GFR'.

Step 4: Include 'Globulin', 'Age', 'Lymphocyte' and 'GFR'.

Factors	Est store	LR test	P- value
Gender	C4	24.12	<0.001
Hemoglobin (g/L)	D4	9.82	0.0017

The fifth predictor selected is ‘Gender’.

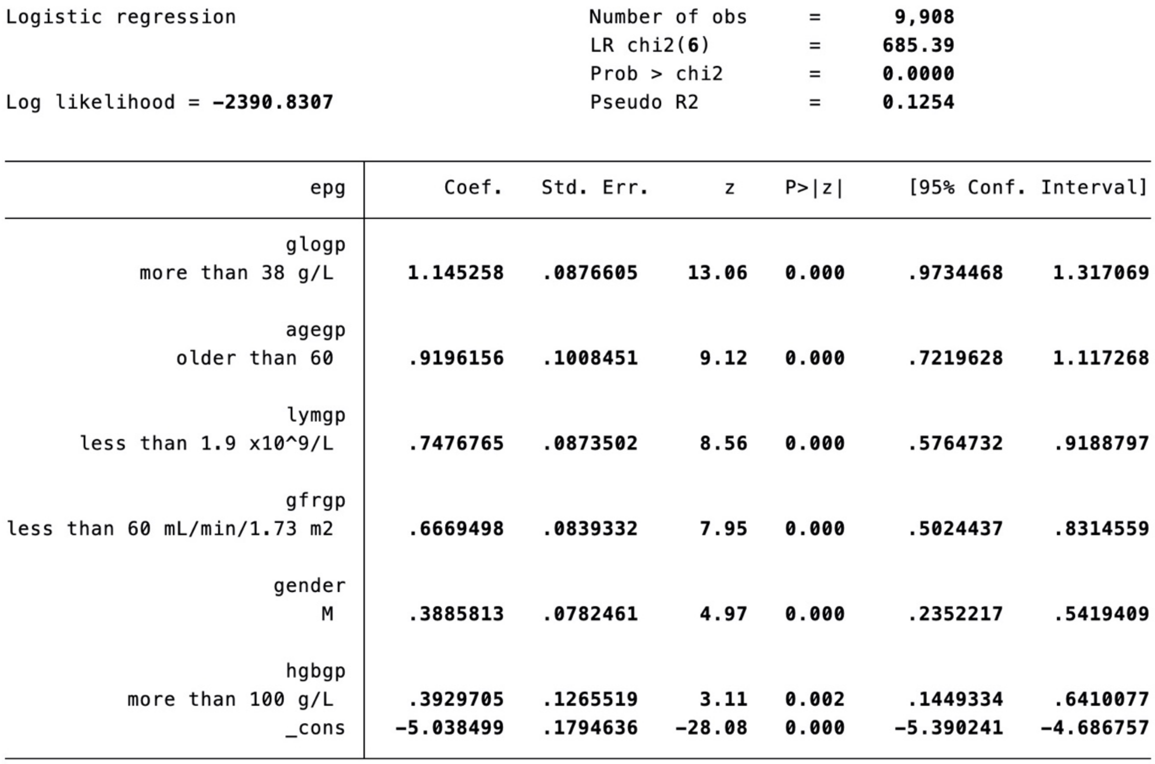
Step 5: Include ‘Globulin’, ‘Age’, ‘Lymphocyte’, ‘GFR’ and ‘Gender’.

Factors	Est store	LR test	P- value
Hemoglobin (g/L)	D5	10.28	0.0013

The fifth predictor selected is ‘Hemoglobin’.

The predictors to be included final multi-variate model are ‘Globulin’, ‘Age’, ‘Lymphocyte’, ‘GFR’, ‘Gender’ and ‘Hemoglobin’.

B. Using Stepwise command



```
Logistic regression                             Number of obs     =      9,908
                                                LR chi2(6)        =     685.39
                                                Prob > chi2       =     0.0000
Log likelihood = -2390.8307                     Pseudo R2         =     0.1254

------------------------------------------------------------------------------------------
                       epg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------------+--------------------------------------------------------------
                     glogp |
          more than 38 g/L |   1.145258   .0876605    13.06   0.000     .9734468    1.317069
                           |
                     agegp |
              older than 60 |   .9196156   .1008451     9.12   0.000     .7219628    1.117268
                           |
                     lymgp |
       less than 1.9 x10^9/L |   .7476765   .0873502     8.56   0.000     .5764732    .9188797
                           |
                     gfrgp |
less than 60 mL/min/1.73 m2 |   .6669498   .0839332     7.95   0.000     .5024437    .8314559
                           |
                    gender |
                         M |   .3885813   .0782461     4.97   0.000     .2352217    .5419409
                           |
                     hgbgp |
         more than 100 g/L |   .3929705   .1265519     3.11   0.002     .1449334    .6410077
                     _cons |  -5.038499   .1794636   -28.08   0.000    -5.390241   -4.686757
------------------------------------------------------------------------------------------
```

Table 3 : Factors associated with SPE/UPE: Multivariate Logistic Regression Analysis

Factors	Coefficient	SE	P value	OR (95% CI)
Age				
<60	0			1
≥ 60	0.92	0.10	<0.001	2.51 (2.06 – 3.06)
Gender				
Female	0			1
Male	0.39	0.08	<0.001	1.47 (1.27 – 1.72)
Hemoglobin (g/L)				
< 100	0			1
≥ 100	0.39	0.13	0.002	1.48 (1.16 – 1.90)
Globulin (g/L)				
<38	0			1
≥ 38	1.15	0.09	<0.001	3.14 (2.65 – 3.73)
Lymphocyte (x 10^9/L)				
<1.9	0.75	0.09	<0.001	2.11 (1.78 – 2.51)
≥ 1.9	0			1
GFR (mL/min/1.73m2)				
<60	0.67	0.08	<0.001	1.95 (1.65 – 2.30)
≥ 60	0			1

Interpretation for Multivariate Logistic Regression

1. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects older than 60 years of age is 2.51 times higher than odds of those younger than 60.
2. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for male subjects is 1.47 times higher than odds of female.
3. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with hemoglobin equal or more than 100 g/L is 1.48 times higher than odds of those with hemoglobin less than 100 g/L.
4. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with globulin equal or more than 38 g/L is 3.14 times higher than odds of those with globulin less than 38 g/L.
5. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with lymphocyte less than 1.9 x 10^9/L is 2.11 times higher than odds of those with lymphocyte equal or more than 1.9 x 10^9/L.
6. The odds of having paraprotein in serum or urine protein electrophoresis (SPE/UPE) for subjects with GFR less than 60 mL/min/1.73m^2 is 1.95 times higher than odds of those with GFR equal or more than 60 mL/min/1.73m^2.

IV. Creating Scoring Scheme Using Regression Coefficient

Factors	Scoring
Age	
<60	0
≥ 60	0.92
Gender	
Female	0
Male	0.39
Hemoglobin (g/L)	
< 100	0
≥ 100	0.39
Globulin (g/L)	
<38	0
≥ 38	1.15
Lymphocyte (x 10^9/L)	
<1.9	0.75
≥ 1.9	0
GFR (mL/min/1.73m2)	
<60	0.67
≥ 60	0

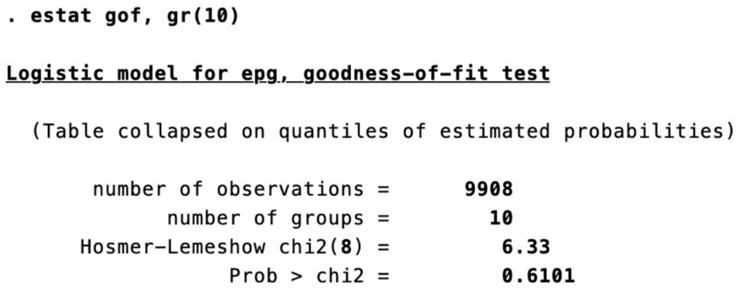
V. Assessing Final Model’s Performance

A. Calibration

Hosmer-Lemeshow goodness of fit test

H_0 : The final model fits well with the given data.

H_A : The final model does not fit well with the given data.

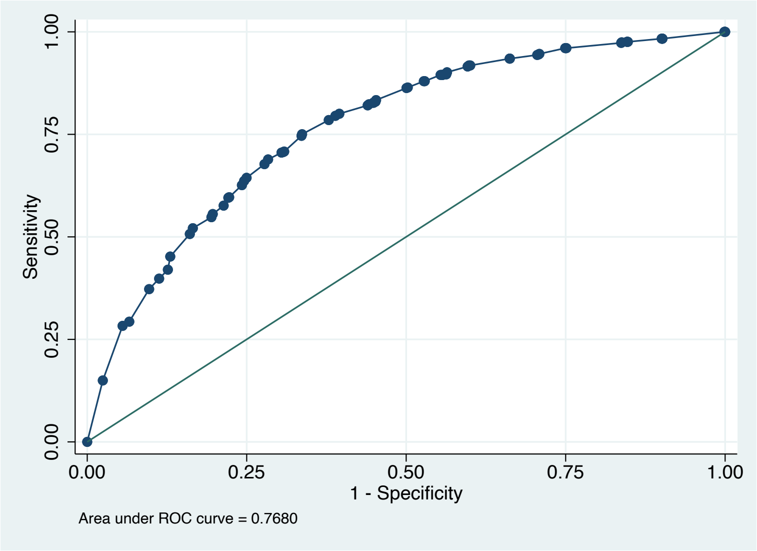


```
. estat gof, gr(10)

Logistic model for epg, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =      9908
             number of groups =        10
      Hosmer–Lemeshow chi2(8) =         6.33
                  Prob > chi2 =         0.6101
```

The p-value is more than 0.05, it is not significant and null hypothesis is failed to reject. The final model fits well with the given data.

B. Discrimination

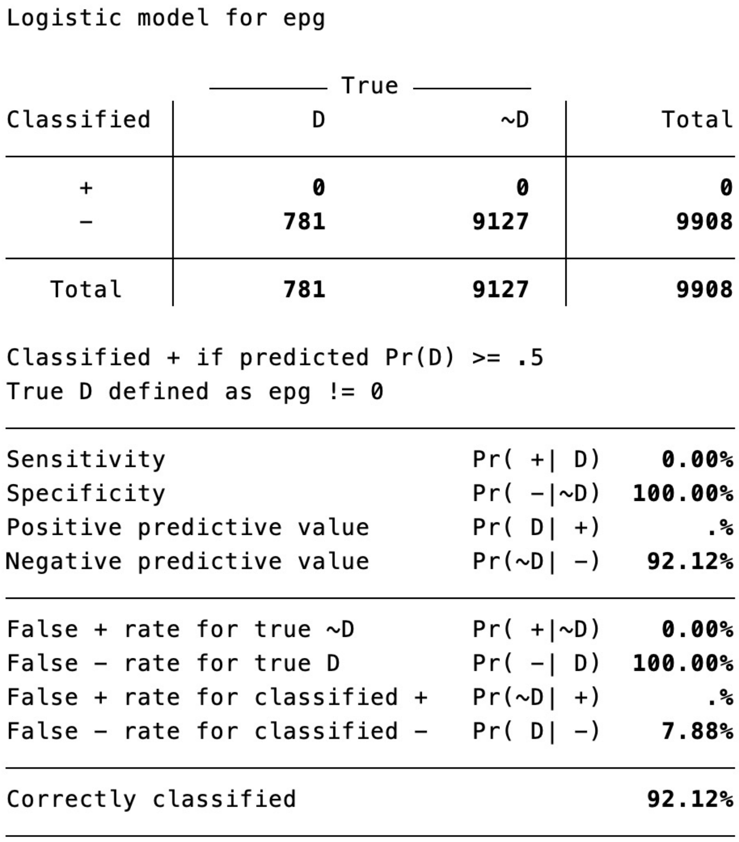
Area Under ROC curve



C statistics = 0.7680 (Acceptable Discrimination)

C. Model Classification

i. Cut off = 0.5 (default)



```
Logistic model for epg

              ──────── True ────────
Classified │         D            ~D  │      Total
───────────┼──────────────────────────┼───────────
     +     │         0             0  │          0
     -     │       781          9127  │       9908
───────────┼──────────────────────────┼───────────
   Total   │       781          9127  │       9908

Classified + if predicted Pr(D) >= .5
True D defined as epg != 0
──────────────────────────────────────────────────
Sensitivity                     Pr( +| D)    0.00%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)       .%
Negative predictive value       Pr(~D| -)   92.12%
──────────────────────────────────────────────────
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)  100.00%
False + rate for classified +   Pr(~D| +)       .%
False - rate for classified -   Pr( D| -)    7.88%
──────────────────────────────────────────────────
Correctly classified                        92.12%
──────────────────────────────────────────────────
```

0.5 is a very bad cut off point for our final model and given data.

ii. Calibrating cut off point

Detailed report of sensitivity and specificity

			Correctly		
Cutpoint	Sensitivity	Specificity	Classified	LR+	LR-
(>= .0599..)	82.07%	56.03%	58.08%	1.8667	0.3199
(>= .0627..)	80.03%	60.48%	62.02%	2.0249	0.3303
(>= .0647..)	79.51%	61.09%	62.55%	2.0437	0.3353.
.					
.					
(>= .1285..)	54.80%	80.53%	78.50%	2.8147	0.5613
(>= .137363)	52.11%	83.44%	80.97%	3.1478	0.5739
(>= .1378..)	50.70%	83.92%	81.30%	3.1524	0.5874.
.					

```
.
( >= .1908.. )    37.26%      90.29%      86.11%            3.8383      0.6949
( >= .2367.. )    29.32%      93.40%      88.35%            4.4455      0.7567
( >= .2375.. )    28.30%      94.46%      89.24%            5.1041    0.7591
( >= .3148.. )    14.98%      97.56%      91.05%            6.1314    0.8715
( >  .3148.. )    0.00%       100.00%     92.12%                      1.0000
```

Cut-off	0.06	0.14	0.237	0.238	0.3148	0.5
Correctly Classified	62.02%	80.97%	88.35%	89.24%	91.05%	92.12%
Sensitivity	80.03%	52.11%	29.32%	28.30%	14.98%	0.00%
Specificity	60.48%	83.44%	93.40%	94.46%	97.56%	100%

We are trying to predict a medical condition with our model. False positives means follow up tests. False negatives means losing early diagnosis and management, so it would be better to having less false negatives at the cost of getting more false positives.

$$sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

False negatives are inversely proportional with sensitivity.

So I would like the model to be more sensitive.

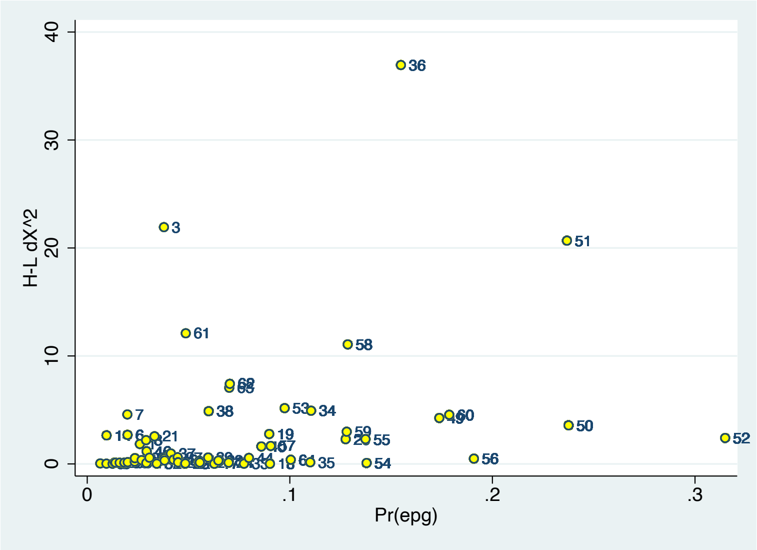
Also, in the assignment, it states that the objective is to develop a simplified clinical decision rule that could increase the yield of serum or urine protein electrophoresis (SPE/UPE) without loss of sensitivity.

0.06 will be considered the new cut off point.

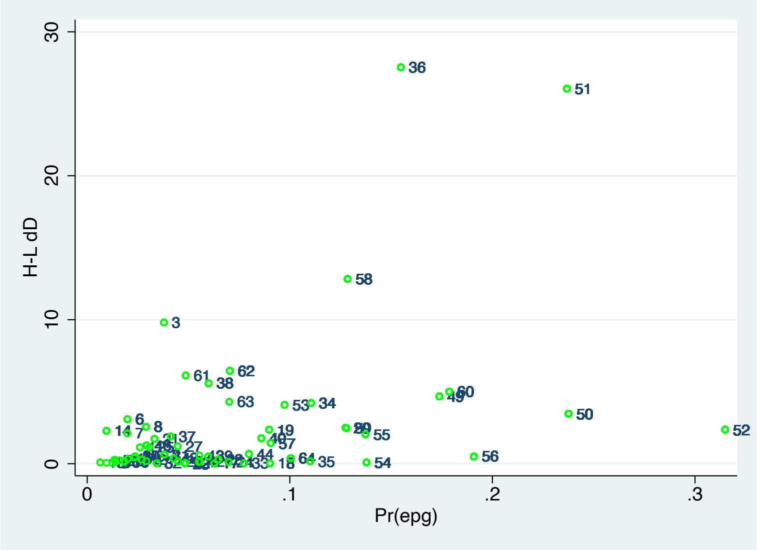
D. Diagnostic Measures

1. Outliers effects on prediction values

a. Pearson Residual Change



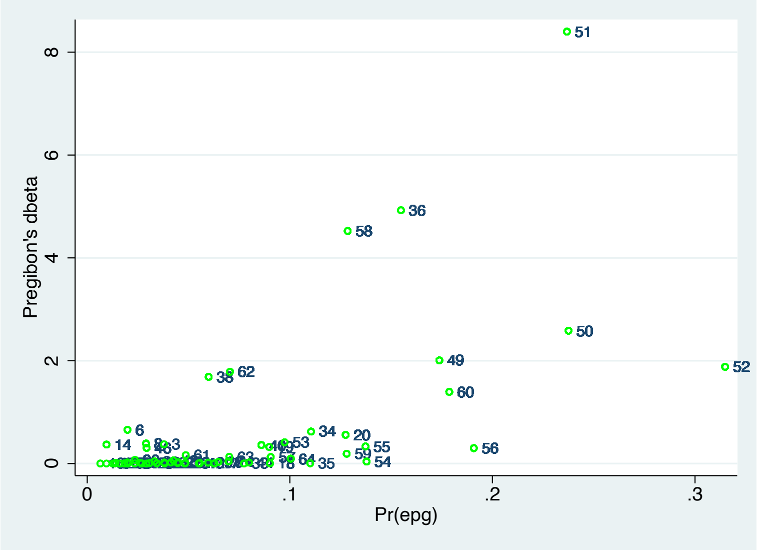
b. Deviance Residual Change



Pattern 36 is the outlier that affects on Prediction Values.

2. Outlier effects on estimation of coefficient

Pregibon-delta coefficient



Pattern 51 is the outlier that affects on estimation of coefficient.

VI. Internal Validation

A. Internal Validation by Splitting Data

i. Data Processing for Validation Dataset

The steps are the same as data preprocessing (I) for derivative dataset except, in validation data set, step (1) is not necessary. Following the rest of the steps, the total sample size is reduced from 4,374 to 4,336, while counting the samples with constraint ‘no missing values in any column’.

ii. Creating Score for validated data

score_final=_b[_cons]+ _b[2.glogp]*2.glogp+ _b[2.agegp]*2.agegp+ _b[1.lymgp]*1.lymgp+ _b[1.gfrgp]*1.gfrgp+ _b[2.gender]*2.gender+ _b[2.hgbgp]*2.hgbgp

score_final=[-5.04]+ [1.15]*2.glogp+ [0.92]*2.agegp+ [0.75]*1.lymgp+ [0.67]*1.gfrgp+ [0.39]*2.gender+ [0.39]*2.hgbgp

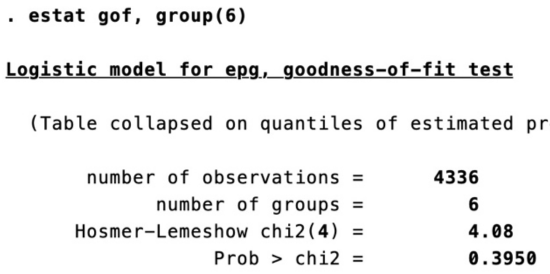
iii. Assessing Model’s Performance

a. Calibration

Hosmer-Lemeshow goodness of fit test

H_0 : The final model fits well with the given data.

H_A : The final model does not fit well with the given data.

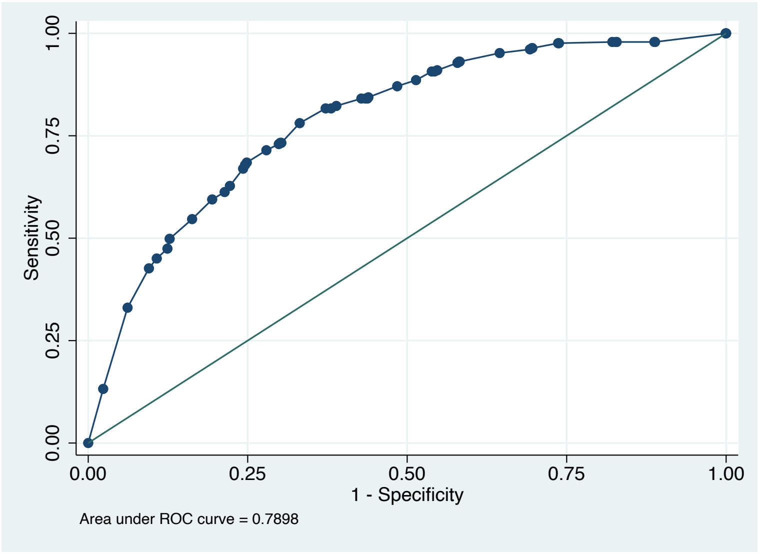


```
. estat gof, group(6)

Logistic model for epg, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =      4336
             number of groups =         6
      Hosmer-Lemeshow chi2(4) =         4.08
                  Prob > chi2 =         0.3950
```

The p-value is more than 0.05, it is not significant and null hypothesis is failed to reject. The final model fits well with the given data.

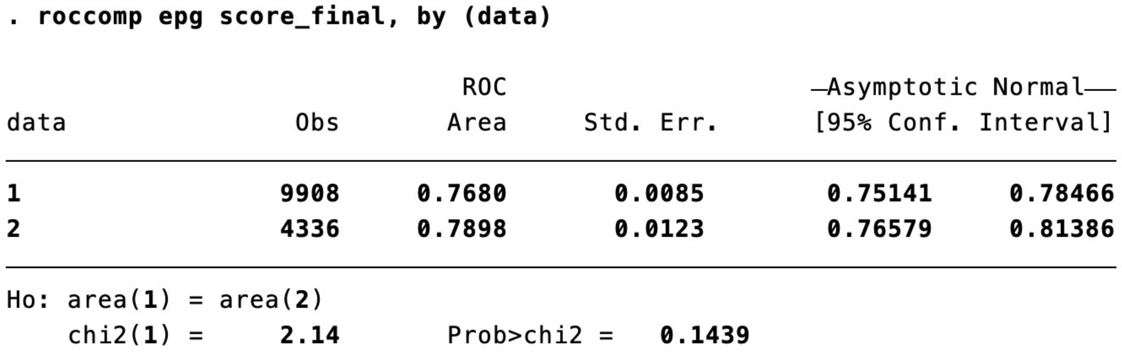
b. Discriminition

Area Under ROC curve



C statistics = 0.7898 (Acceptable Discrimination)

iv. Comparing C statistics between derived and validated model



```
. roccomp epg score_final, by (data)

                              ROC                    —Asymptotic Normal——
data             Obs         Area     Std. Err.      [95% Conf. Interval]
─────────────────────────────────────────────────────────────────────────
1               9908       0.7680       0.0085        0.75141     0.78466
2               4336       0.7898       0.0123        0.76579     0.81386
─────────────────────────────────────────────────────────────────────────
Ho: area(1) = area(2)
    chi2(1) =     2.14       Prob>chi2 =   0.1439
```

The C statistics for derived data (data 1) is 0.7680 (0.7514 – 0.7847)

The C statistics for validated data (data 2) is 0.7898 (0.7658 – 0.8139).

The discriminative ability in the validated data was slightly better than the derived dataset.

But it did not reached statistical significance (p = 0.1439)

It could be interpreted as the predicted model may not work this well in another point of data collection in that center.

B. Internal Validation by Bootstrapping

i. Original C statistics (C_{origin}) and Somers'D coefficient (D_{origin}) from derived model

C_{origin} = 0.7680
D_{origin} = 0.5360

ii. Bootstrapping and estimating C statistics and Somers'D coefficient for each bootstrap

Iteration is set to 200 as stated in module.

iii. Calculating the mean optimism for both C and D

$C_{optism} = C_{origin} – C_{boot}$
$D_{optism} = D_{origin} - D_{boot}$

Mean optimism for C statistics = -0.0345 ((-0.0414) - (-0.0275))
Mean optimism for Somers'D coefficient = -0.0689 ((-0.0829) - (-0.0549))

Optimism closer to zero reflect to good model and less bias.

iv. Bootstrap corrected calibration coefficient

$D_{ccor} = D_{origin.} - D_{optism}$ = 0.6049

VII. Appendix

i. Three Data files are submitted.

a. [Assignment I epg for derive phase.dta] and [Assignment I epg for validation phase.dta] are used in internal validation by splitting data procedure. Because we has to add and drop new columns, it has to be separate procedures with Bootstrapping.

b. [result_boot.dta] is the result file of the bootstrapping procedure, which we used to find out optimum.

ii. Three log files are submitted.

a. [Model_u_m_iv.log] includes

1. Data preprocessing for derived dataset
2. Univariate Logistic Regression
3. Model Selection (Manual and Stepwise)
4. Multivariate Logistic Regression
5. Creating Scoring Scheme
6. Assessing Model Performance
7. Calibration of Cutoff point
8. Data preprocessing for validation dataset
9. Creating Scores
10. Assessing Model Performance
11. Comparing C statistics between derived and validated models

b. [model_u_m_iv_bs.smcl] includes

1. Data preprocessing for derived dataset
2. Univariate Logistic Regression
3. Model Selection (Manual and Stepwise)
4. Multivariate Logistic Regression
5. Creating Scoring Scheme
6. Assessing Model Performance
7. Calibration of Cutoff point
8. Bootstrapping for 200 iterations

(This log file might not have the same results as stated in this report since they were not created at the same time)

c. [after_bootstrapping.log] includes

1. Calculating the mean optimism for both C and D

2. Bootstrap corrected calibration coefficient

(This log file is created together with [result_boot.dta] file and consistent with the results reported here)