Categorical data analysis

Question 1

Write down the equation for the logistic regression model of STA (Vital Status: 0 = Lived,1= Died) on AGE(Years).

```
##
## Call:
## glm(formula = STA ~ AGE, family = binomial, data = table)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9536  -0.7391  -0.6145  -0.3905   2.2854
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.05851    0.69608  -4.394 1.11e-05 ***
## AGE          0.02754    0.01056   2.607  0.00913 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 192.31  on 198  degrees of freedom
## AIC: 196.31
##
## Number of Fisher Scoring iterations: 4
```

The prediction equation is:

$$logit(\hat{\pi}) = -3.05851 + 0.02754 * AGE$$

$$\pi(AGE) = \frac{exp(-3.05851 + 0.02754 * AGE_i)}{1 + exp(-3.05851 + 0.02754 * AGE_i)}$$

Write down the log likelihood for the logistic regression model and the equation for solving the maximum likelihood estimates (MLE) of the parameters.

Log-likelihood for the logistic regression model:

$$l(a,b) = \sum_{i=1}^{200} y_i\, log(\pi) + (1 - y_i) log(1 - \pi) = \sum_{i=1}^{200} y_i\, log(\frac{\pi}{1 - \pi}) + log(1 - \pi)$$

$$log(\frac{\pi}{1 - \pi}) = a + b * AGE_i$$

$$1-\pi = 1 - \frac{exp(a+b*AGE_i)}{1+exp(a+b*AGE_i)} = (1+exp(a+b*AGE_i))^{-1}$$

$$log(1-\pi) = log((1+exp(a+b*AGE_i))^{-1}) = -3.05851 + 0.02754 * AGE_i$$

$$l(a,b) = \sum_{i=1}^{200} STA_i(a+b*AGE_i) - log(1+exp(a+b*AGE_i))$$

$$l(a,b)$$

$$= \sum_{i=1}^{200} STA_i(-3.05851 + 0.02754 * AGE_i) - log(1+exp(-3.05851 + 0.02754 * AGE_i))$$

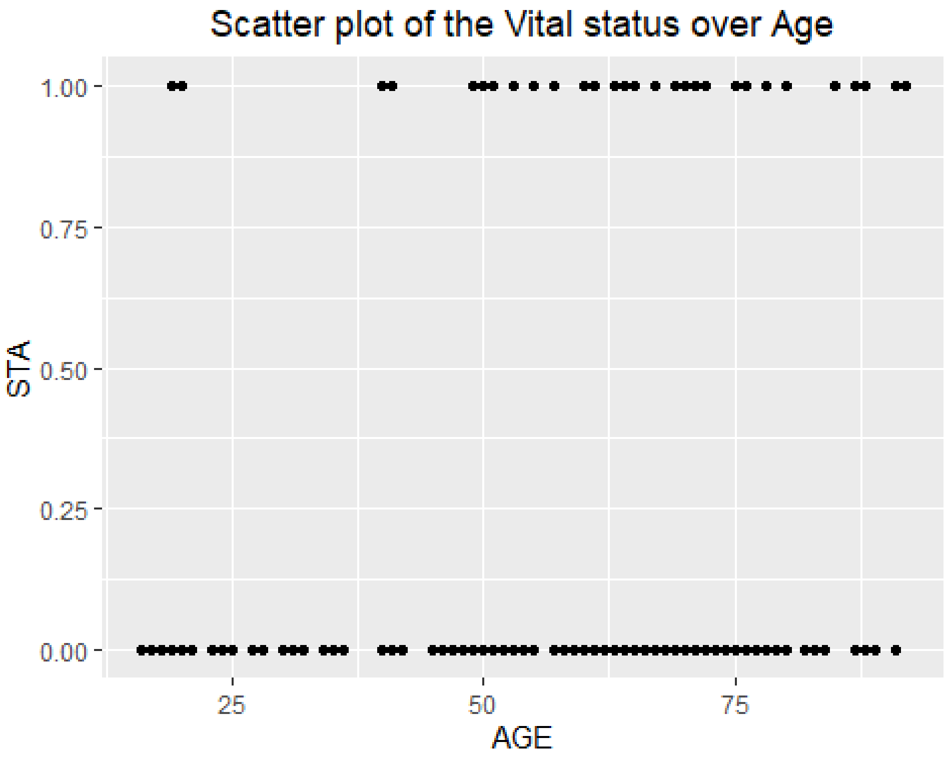
Equation for solving the MLE of the parameters:

$$\frac{\partial l(a,b)}{\partial a} = \sum_{i=1}^{200} (STA_i - \pi) = 0$$

$$\frac{\partial l(a,b)}{\partial b} = \sum_{i=1}^{200} AGE_i(STA_i - \pi) = 0$$

Question 2

We can observe the scatterplot of STA versus AGE:



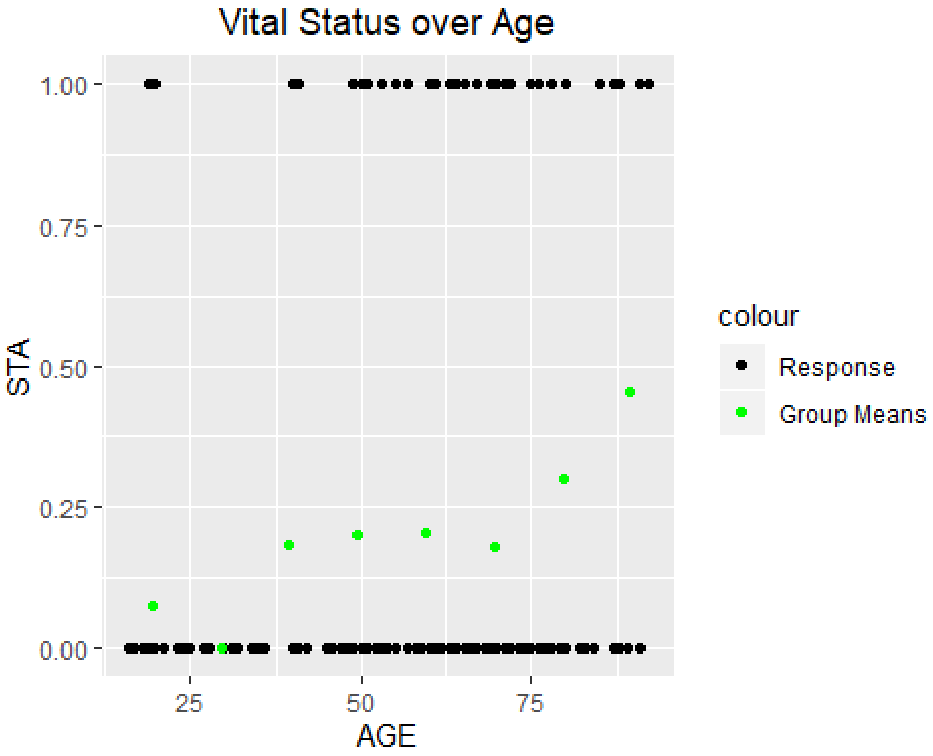
Question 3

We compute the STA mean over subjects within each AGE interval:

The number of people per interval is: [26, 8, 11, 25, 39, 50, 30, 11]

and the percentage of people who died per interval is: [0.0769, 0, 0.1818, 0.2, 0.2051, 0.18, 0.3, 0.4545]

We plot these values of mean STA versus the midpoint of the AGE interval:



Question 4

We obtain the MLE of the parameters of mean STA versus the midpoint of the AGE interval.

Recall: STA is the Vital Status: 0 = Lived,1= Died, AGE is in Years

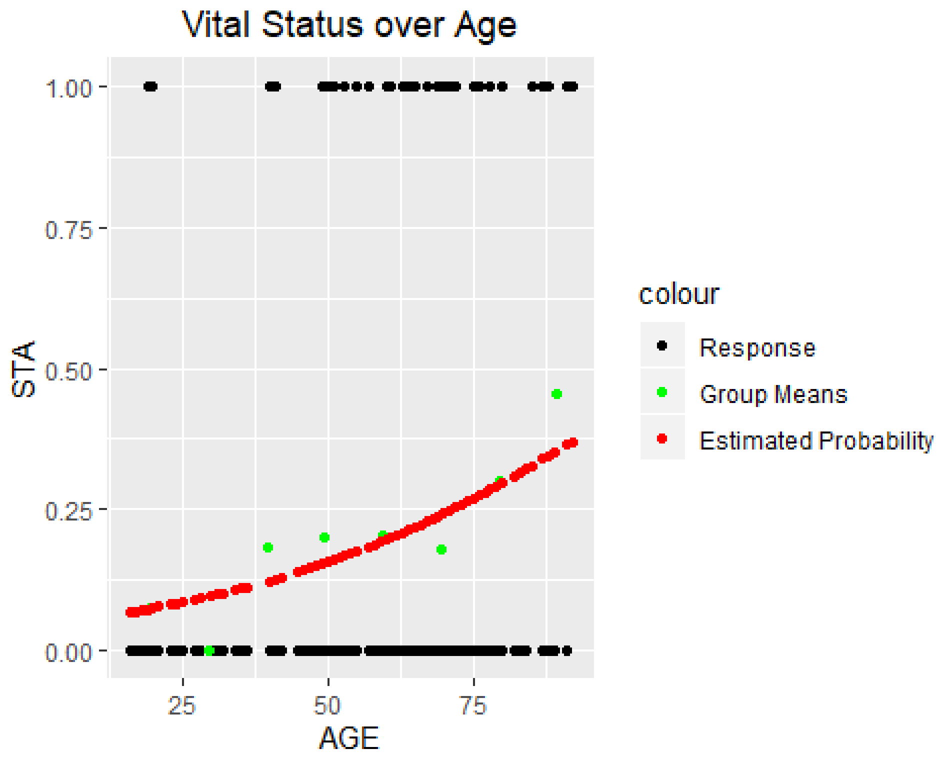
```
##
## Call:
## glm(formula = STAsum/n ~ Age, family = binomial, data = table,
##     weights = n)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2564  -0.1986   0.1380   0.5979   0.6820
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.08609    0.69338  -4.451 8.56e-06 ***
## Age          0.02779    0.01043   2.664  0.00771 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 12.0790  on 7  degrees of freedom
## Residual deviance:  3.8722  on 6  degrees of freedom
## AIC: 30.234
## 
## Number of Fisher Scoring iterations: 4
```

The prediction equation is:

$$logit(\hat{\pi}) = -3.08609 + 0.02779 * Age$$

We plot the estimated probabilities of the STA mean over subjects within each AGE interval:



Does it appear plausible that the logit of the survival probability is linear in AGE?

Yes, it visually appears plausible.

Question 5

1st part

We compute the STA mean over subjects within each SYS interval:

The number of people per interval is:

4 14 46 38 39 41 12 6

and the percentage of people who died per interval is:

0.75 0.6429 0.1087 0.2105 0.2051 0.1219 0.0833 0.1667

We obtain the MLE of the parameters of mean STA versus the midpoint of the SYS interval.

```
##
## Call:
## glm(formula = STA ~ SYS, family = binomial, data = table)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1606  -0.7051  -0.5956  -0.4056   2.6860
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.777195   0.757375   1.026  0.30481
## SYS         -0.017019   0.006001  -2.836  0.00456 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 191.34  on 198  degrees of freedom
## AIC: 195.34
##
## Number of Fisher Scoring iterations: 4
```

The prediction equation is:

$$logit(\hat{\pi}) = 0.777195 + -0.017019 * Sys$$

We also transfrom the data taking the log:

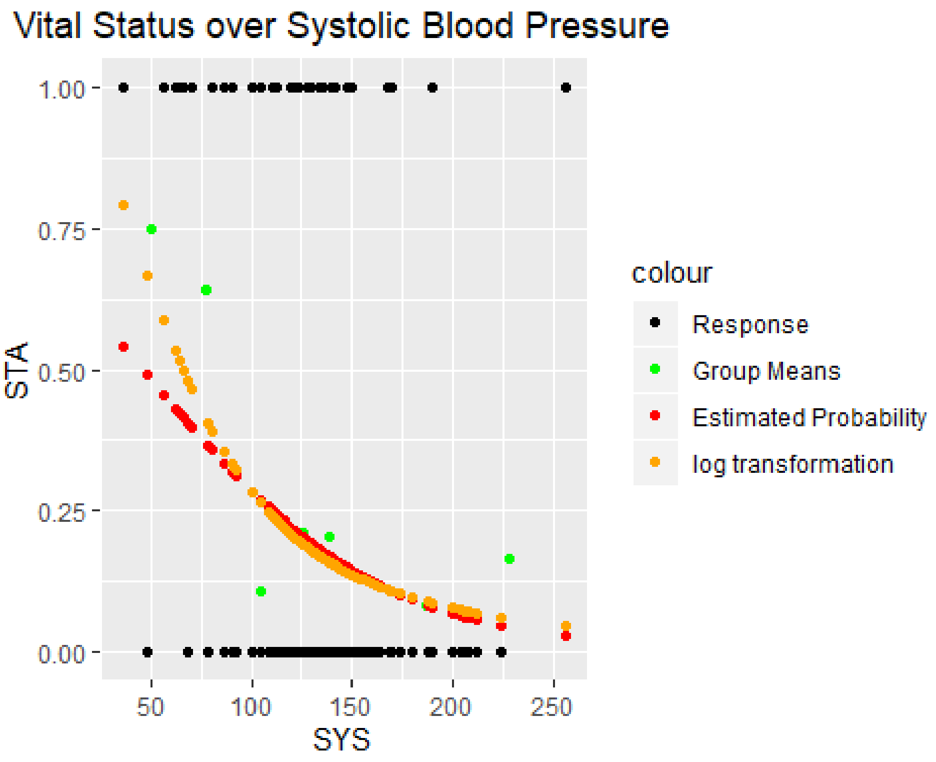
```
##
## Call:
## glm(formula = STA ~ log(SYS), family = binomial, data = table)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4853  -0.6824  -0.5757  -0.4245   2.4788
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   9.3131     3.1990   2.911 0.003600 **
## log(SYS)     -2.2250     0.6686  -3.328 0.000875 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 188.06  on 198  degrees of freedom
## AIC: 192.06
##
## Number of Fisher Scoring iterations: 4
```

The prediction equation is:

$$logit(\hat{\pi}) = 9.3131 + -2.2250 * Sys$$

We plot scatterplot of STA versus SYS, the mean STA versus the midpoint of the SYS interval and the estimated probabilities.



2nd part

We compute the STA mean over subjects within each HRA interval:

The number of people per interval is:

8 29 68 42 29 19 5

and the percentage of people who died per interval is:

0.25 0.1724 0.2059 0.1667 0.2414 0.2105 0.2

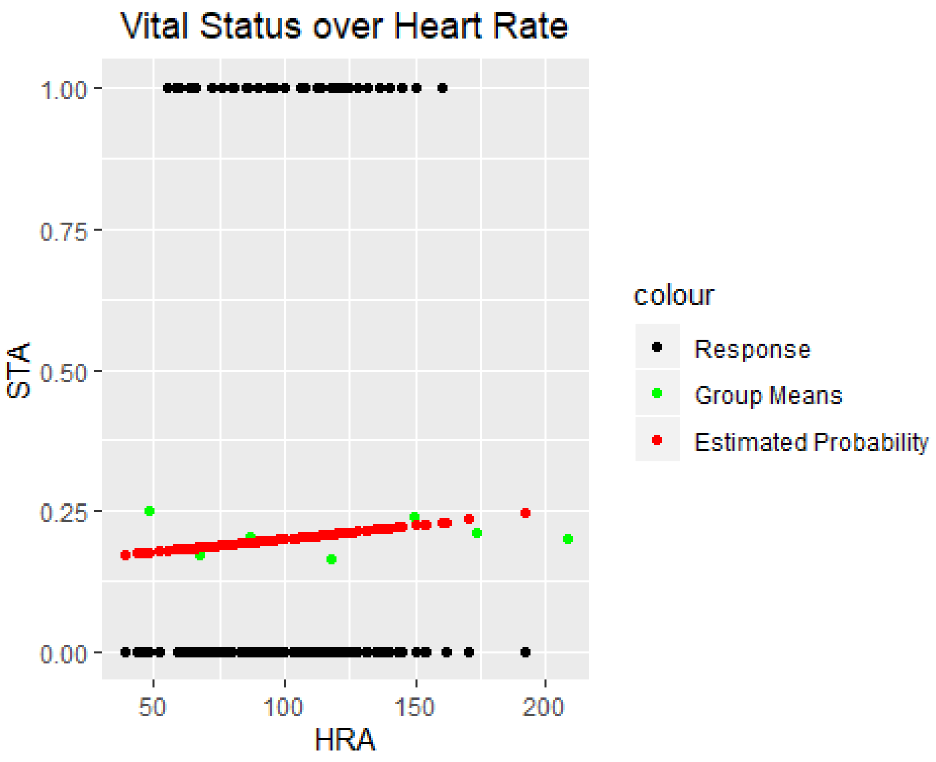
We obtain the MLE of the parameters of mean STA versus the midpoint of the HRA interval:

```
##
## Call:
## glm(formula = STA ~ HRA, family = binomial, data = table)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7533  -0.6784  -0.6579  -0.6339   1.8524
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.679129   0.679863  -2.470   0.0135 *
## HRA          0.002941   0.006552   0.449   0.6535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 199.96  on 198  degrees of freedom
## AIC: 203.96
##
## Number of Fisher Scoring iterations: 4
```

The prediction equation is:

$$logit(\hat{\pi}) = -1.679129 + 0.002941 * Hra$$

We plot scatterplot of STA versus HRA, the mean STA versus the midpoint of the HRA interval and the estimated probabilities.



What are the correct scales for SYS and HRA to enter the model?

The correct scale for SYS seems to be the log scale.

HRA has a slope almost equal to zero. It seems to be that HRA is independent to STA. We will do further analysis in the following question.

Question 6

We look at the MLE of the predictors when we include all of them:

```
##
## Call:
## glm(formula = STA ~ AGE + SEX + RACE + SER + CAN + CRN + INF +
##     CPR + SYS + HRA + PRE + TYP + FRA + PO2 + PH + PCO + BIC +
##     CRE + LOC, family = binomial, data = table)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.50525  -0.53717  -0.17867  -0.00019   3.01708
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.548e+00  2.271e+00  -2.444  0.01454 *
## AGE          5.645e-02  1.848e-02   3.055  0.00225 **
## SEX         -7.215e-01  5.460e-01  -1.321  0.18639
## RACE2       -1.617e+01  1.314e+03  -0.012  0.99018
## RACE3        5.829e-01  1.313e+00   0.444  0.65696    
## SER         -6.739e-01  6.289e-01  -1.071  0.28398    
## CAN          3.483e+00  1.121e+00   3.106  0.00189 ** 
## CRN          1.191e-01  8.449e-01   0.141  0.88786    
## INF         -1.081e-01  5.557e-01  -0.195  0.84573    
## CPR          1.032e+00  9.901e-01   1.043  0.29714    
## SYS         -2.084e-02  9.443e-03  -2.207  0.02732 *  
## HRA         -2.915e-03  1.032e-02  -0.282  0.77761    
## PRE          1.279e+00  7.022e-01   1.822  0.06842 .  
## TYP          3.748e+00  1.342e+00   2.792  0.00523 ** 
## FRA          1.649e+00  1.093e+00   1.509  0.13139    
## PO2         -6.765e-01  9.402e-01  -0.720  0.47179    
## PH           1.771e+00  1.212e+00   1.461  0.14410    
## PCO         -2.084e+00  1.165e+00  -1.789  0.07361 .  
## BIC         -2.623e-01  8.967e-01  -0.293  0.76985    
## CRE          1.004e-01  1.131e+00   0.089  0.92925    
## LOC1         3.771e+01  2.487e+03   0.015  0.98790    
## LOC2         3.458e+00  1.342e+00   2.578  0.00994 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 112.17  on 178  degrees of freedom
## AIC: 156.17
## 
## Number of Fisher Scoring iterations: 17
```

Looking at the summary of all the predictors included, we can see that the predictors: AGE, CAN, SYS, PRE, TYP, PCO and LOC seems to be significant.

Using a model-buildigng strategy, we will select a logistic model for these predictors.

1) Let's start with the backward selection:

We begins with a complex model and sequentially removes terms. At each stage, it selects the term with the largerst p-value for removal. The process stops when any further deletion leads to a significantly poorer fit.

```
## 
## Call:  glm(formula = STA ~ AGE + CAN + SYS + TYP + PH + PCO + LOC, family 
= binomial(link = logit), 
##     data = table)
## 
## Coefficients:
## (Intercept)          AGE          CAN          SYS          TYP  
##    -5.61893      0.04361      2.75119     -0.01879      3.79845  
##          PH          PCO         LOC1         LOC2  
##     1.59249     -1.77718     21.59323      3.15688  
## 
```

```
## Degrees of Freedom: 199 Total (i.e. Null);  191 Residual
## Null Deviance:       200.2
## Residual Deviance: 123.3     AIC: 141.3
```

2) forward selection

At each stage, it selects the term giving the greatest improvement in fit (the term with the smallest p-value). The process stops when further additions do not significantly improve the fit.

```
##
## Call:  glm(formula = STA ~ LOC + TYP + AGE + CAN + SYS, family = binomial,
##     data = table)
##
## Coefficients:
## (Intercept)         LOC1         LOC2          TYP          AGE
##    -5.43308     21.54391      2.41686      3.84786      0.03832
##         CAN          SYS
##     2.59742     -0.01780
##
## Degrees of Freedom: 199 Total (i.e. Null);  193 Residual
## Null Deviance:       200.2
## Residual Deviance: 128.4     AIC: 142.4
```

3) bestgltm with AIC

```
## AIC
## BICq equivalent for q in (0.830250042405567, 0.889798982907676)
## Best Model:
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -0.518536477 0.1580627834 -3.280573 1.231394e-03
## AGE          0.004452841 0.0012484554  3.566680 4.569892e-04
## CAN          0.241609667 0.0880850771  2.742913 6.669297e-03
## SYS         -0.001187431 0.0007483128 -1.586811 1.142095e-01
## PRE          0.097860574 0.0685368969  1.427852 1.549673e-01
## TYP          0.241853086 0.0614508163  3.935718 1.160934e-04
## PH           0.235585276 0.1108888952  2.124516 3.491301e-02
## PCO         -0.226130502 0.0928389763 -2.435728 1.577968e-02
## LOC          0.364505855 0.0551043998  6.614823 3.640248e-10
```

4) bestglm with BIC

```
## BIC
## BICq equivalent for q in (0.132557887868053, 0.559436232609211)
## Best Model:
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -0.555134007 0.103480760 -5.364611 2.273948e-07
## AGE          0.003892268 0.001259688  3.089867 2.293563e-03
## TYP          0.204907278 0.057395027  3.570122 4.488065e-04
## LOC          0.338263643 0.054632949  6.191568 3.425787e-09
```

Sum up:

1) backward selection: STA ~ AGE + CAN + SYS + TYP + PH + PCO + LOC (7 predictors)

2) forward selection: STA ~ LOC + TYP + AGE + CAN + SYS (5 predictords)

3) bestglm with AIC: STA ~ AGE + CAN + SYS + PRE + TYP + PH + PCO + LOC (8 predictors)

4) bestglm with BIC: STA ~ AGE + TYP + LOC (3 predictors)

We test if the best_BIC_mod fits as well as the full model doing anova. (we test this model first because it is composed with the fewest number of predictors with 3 predictors.)

```
## Analysis of Deviance Table
##
## Model 1: STA ~ AGE + TYP + LOC
## Model 2: STA ~ AGE + SEX + RACE + SER + CAN + CRN + INF + CPR + SYS +
##     HRA + PRE + TYP + FRA + PO2 + PH + PCO + BIC + CRE + LOC
##   Resid. Df Resid. Dev Df Deviance
## 1       195     141.87
## 2       178     112.17 17   29.705
```

p-value(29.705, 17) = 0.0285527

Because the p - value < 0.05, we reject the null hypothesis. Therefore, we conclude that the best_BIC_mod does not fit as well as the full model.

We then test if the for_mod fits as well as the full model.

```
## Analysis of Deviance Table
##
## Model 1: STA ~ AGE + CAN + SYS + TYP + LOC
## Model 2: STA ~ AGE + SEX + RACE + SER + CAN + CRN + INF + CPR + SYS +
##     HRA + PRE + TYP + FRA + PO2 + PH + PCO + BIC + CRE + LOC
##   Resid. Df Resid. Dev Df Deviance
## 1       193     128.44
## 2       178     112.17 15   16.273
```

p-value(16.273, 15) = 0.3641436

The p-value is > 0.05 so we do not reject the null hypothesis. Therefore, we conclude that the for_mod fits as well as the full model.

So with the forward selection, we can select the model: STA ~ AGE + CAN + SYS + TYP + LOC.

Question 7

I - goodness of fit

1) classification table

Let's evaluate the confusion matrix with a threshold = 0.5.

```
##                predicted_values
## actual_values   0   1
##             0 157   3
##             1  23  17
```

The sensitivity is 0.425 which means that the ‘true positive rate’ is 42.5% i.e. the probability of predicting that a patient will live (STA=0) after his admission to the ICU is .425 when the patient actually survives.

The specificity is 0.98125 which means that the ‘true negative rate’ is 98% i.e. the probability of predicting that a patient will die (STA=1) after his admission to teh ICU is .98 when the patient actually dies.

With a threshold = 0.5, the specificity is better than the sensitivity.

Let’s evaluate the confusion matrix with a threshold = threshold= mean(STA).

```
##                predicted_values
## actual_values   0   1
##             0 122  38
##             1  10  30
```

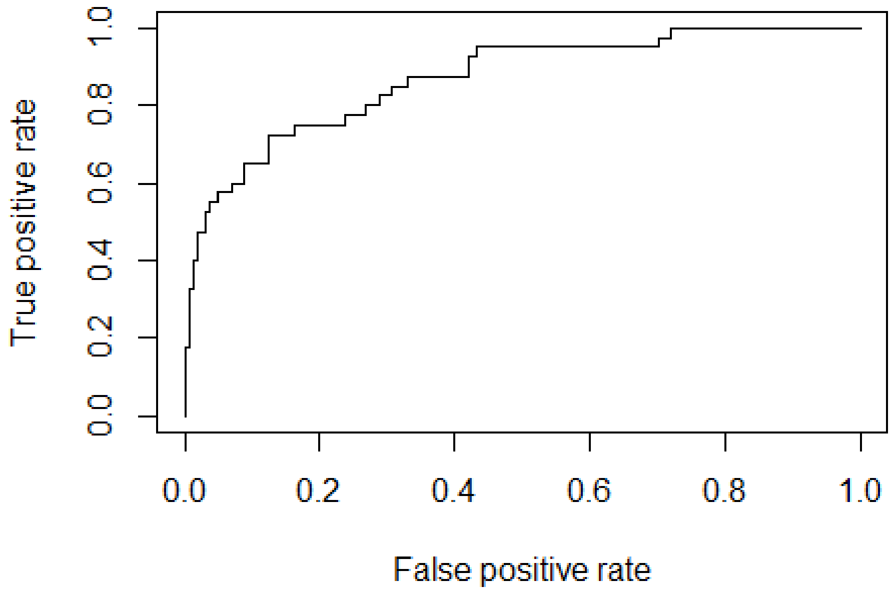
The sensitivity is 0.75 which means that the ‘true positive rate’ is 75% i.e. the probability of predicting that a patient will live (STA=0) after his admission to the ICU is .75 when the patient actually survives.

The specificity is 0.7625 which means that the ‘true negative rate’ is 76% i.e. the probability of predicting that a patient will die (STA=1) after his admission to teh ICU is .76 when the patient actually dies.

With this threshold, the sensitivity and the specificity are equivalent.

2) ROC Curve

We can evaluate the model with the ROC Curve:



In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. This ROC curve is used to show in a graphical way the trade-off between sensitivity and specificity for every possible cut-off for a test or a combination of tests. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

Concordance index = Area under curve = 0.8704687

Therefore, the proportion of pairs of data whose fitted values are pairwise concordant is higher than 87% which indicates that this model fits the data adequately and predict accordingly.

3) Hosmer–Lemeshow test

```
## 
##  Hosmer and Lemeshow goodness of fit (GOF) test
## 
## data:  for_mod$y, fitted(for_mod)
## X-squared = 5.9846, df = 8, p-value = 0.649
```

This gives p - value = 0.649 > 0.05, indicating no evidence of poor fit.

II - influential statistics

We do not observe any value where the Cook's distance and the leverage are influential. Therefore, we may not have any influential points in the data.

Question 8

Looking at the MLE of the model we selected with forward selection, we have:

```
##
## Call:
## glm(formula = STA ~ AGE + CAN + SYS + TYP + LOC, family = binomial,
##     data = table)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.16314  -0.54198  -0.31837  -0.07933   2.63978
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.433e+00  1.792e+00  -3.031  0.00243 **
## AGE          3.832e-02  1.294e-02   2.962  0.00306 **
## CAN          2.597e+00  9.626e-01   2.698  0.00697 **
## SYS         -1.780e-02  7.473e-03  -2.382  0.01722 *
## TYP          3.848e+00  1.270e+00   3.030  0.00244 **
## LOC1         2.154e+01  1.457e+03   0.015  0.98820
## LOC2         2.417e+00  8.743e-01   2.764  0.00570 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 128.44  on 193  degrees of freedom
## AIC: 142.44
##
## Number of Fisher Scoring iterations: 16
```

1) Find the odds ratio of survival for a 40 year old subject compared to a 50 year old subject, controlling for other factors and its 95% confidence interval.

The odds ratio of survival (STA) for a 40 year old subject compared to a 50 year old subject, controlling for other factors is exp(3.832e-02*40 - 3.832e-02*50) = 0.6816766.

The 95% Confidence interval is [exp(-0.3832 - 1.96*1.294e-02*10), exp(-0.3832 + 1.96*1.294e-02*10)] = [0.5289698, 0.8784678].

2) Find the odds ratio of survival and its 95% confidence interval for a subject who was in coma at ICU admission and a subject in deep stupor at admission, controlling for other factors.

recall: LOC: Level of Consciousness 0 = No Coma or Deep Stupor, LOC at ICU Admission 1 = Deep Stupor, 2= Coma

Because we are comparing LOC=2 with LOC=1, we change the baseline for LOC=2.

It turns out there is quasi complete separation for the LOC variable at level 1 (deep stupor).

The the odds ratio of survival for a subject who was in coma at ICU admission and a subject in deep stupor at admission, controlling for other factors. is exp(1.913e+01) = 203260710.

The 95% Confidence interval is [-56.15217480, 420.351020370].