ComputerAssignment4

Gabriel Lindqvist, Jiahui Li

2024-12-05

In this exercise you will carry out several analyses of The ICU Study data (Hosmer Lemeshow (1989):Applied Logistic Regression).The data consists of records on 200 patients admitted to an Intensive Care Unit (ICU)during a certain period and the overall task is to use multiple logistic regression and decision tree models to identify factors that affect the survival of such patients.

Exercise4:1 (Multiple logistic regression)

Which variables affect the survival of a patient admitted to ICU?And how do they affect the survival?To answer that,you should create a multiple logistic regression model for the probability to not survive (i.e.the probability to die).For variable selection,stepwise methods can be used.We will assume that there are no interactions between the variables (you are free to investigate whether this assumption is reasonable).

Task1

Report the model selection process in short. Based on your chosen model, which factors affect the probability to not survive? Report odds ratio with confidence interval for the most important variables/factors and interpret them. Use the variable names in the table above when you report this (not V3,V4,…)

1.1 Data prepration

```
data4 <- read.csv("data_ca4.csv")
# Combine categories
table(data4$v5)
```

```
##
##   1   2   3
## 175  15  10
```

```
data4$v5[data4$v5>1] <- 0
table(data4$v5)
```

```
##
##   0   1
##  25 175
```

```r
table(data4$v21)
```

```
##
##   0   1   2
## 185   5  10
```

```r
data4$v21[data4$v21>0] <- 1
table(data4$v21)
```

```
##
##   0   1
## 185  15
```

1.2 Fitting multiple logistic regression models

```r
m_empty <- glm(v2~1, family=binomial,data=data4) # Empty model (only intercept)
m_full <- glm(v2~ .-v1, family=binomial,data=data4) # Full model (all explanatory variables, remove v1)
summary(m_full)
```

```
##
## Call:
## glm(formula = v2 ~ . - v1, family = binomial, data = data4)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.80396  -0.56064  -0.20440  -0.08635   2.97729
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.086979   2.260515  -2.693  0.00709 **
## v3           0.056393   0.018624   3.028  0.00246 **
## v4          -0.639725   0.531393  -1.204  0.22864
## v5           0.565729   0.926828   0.610  0.54160
## v6          -0.673522   0.601902  -1.119  0.26315
## v7           3.107051   1.045846   2.971  0.00297 **
## v8          -0.035708   0.801647  -0.045  0.96447
## v9          -0.204933   0.553191  -0.370  0.71104
## v10          1.053483   1.006614   1.047  0.29530
## v11         -0.015472   0.008497  -1.821  0.06864 .
## v12         -0.002769   0.009607  -0.288  0.77317
## v13          1.131942   0.671450   1.686  0.09183 .
## v14          3.079583   1.081584   2.847  0.00441 **
## v15          1.411402   1.029705   1.371  0.17047
## v16          0.073822   0.857044   0.086  0.93136
## v17          2.354078   1.208804   1.947  0.05148 .
## v18         -3.018442   1.253448  -2.408  0.01604 *
## v19         -0.709284   0.909777  -0.780  0.43561
## v20          0.295143   1.116925   0.264  0.79159
## v21          5.232292   1.226303   4.267 1.98e-05 ***
## ---
```

```
## 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: 120.78  on 180  degrees of freedom
## AIC: 160.78
##
## Number of Fisher Scoring iterations: 6
```

1.3 Stepwise model selection - forward, backward and both - with AIC or BIC as inclusion criteria

```
mstep1_AIC <- step(m_empty, direction="forward", scope = list(upper = m_full), trace = TRUE)
```

```
## Start:  AIC=202.16
## v2 ~ 1
##
##        Df Deviance    AIC
## + v21   1   165.56 169.56
## + v14   1   185.05 189.05
## + v11   1   191.34 195.34
## + v10   1   192.23 196.23
## + v3    1   192.31 196.31
## + v6    1   193.24 197.24
## + v9    1   193.59 197.59
## + v8    1   194.74 198.74
## + v20   1   195.40 199.40
## <none>      200.16 202.16
## + v19   1   198.56 202.56
## + v5    1   198.89 202.89
## + v16   1   198.92 202.92
## + v17   1   199.25 203.25
## + v13   1   199.92 203.92
## + v12   1   199.96 203.96
## + v4    1   200.08 204.08
## + v7    1   200.16 204.16
## + v18   1   200.16 204.16
## + v15   1   200.16 204.16
##
## Step:  AIC=169.56
## v2 ~ v21
##
##        Df Deviance    AIC
## + v14   1   154.51 160.51
## + v3    1   159.48 165.48
## + v11   1   159.76 165.76
## + v6    1   159.79 165.79
## + v9    1   161.03 167.03
## + v20   1   162.42 168.42
## + v17   1   163.16 169.16
## + v8    1   163.22 169.22
```

```
## + v19   1   163.40 169.40
## <none>      165.56 169.56
## + v5    1   163.69 169.69
## + v10   1   164.07 170.07
## + v12   1   164.24 170.24
## + v13   1   164.50 170.50
## + v16   1   164.80 170.80
## + v18   1   164.98 170.98
## + v7    1   165.07 171.07
## + v15   1   165.55 171.55
## + v4    1   165.55 171.55
##
## Step:  AIC=160.51
## v2 ~ v21 + v14
##
##        Df Deviance    AIC
## + v3    1   145.31 153.31
## + v7    1   149.45 157.45
## + v11   1   150.77 158.77
## + v9    1   151.71 159.71
## + v5    1   152.07 160.07
## <none>      154.51 160.51
## + v20   1   152.65 160.65
## + v13   1   152.99 160.99
## + v8    1   153.02 161.02
## + v17   1   153.09 161.09
## + v19   1   153.60 161.60
## + v6    1   153.71 161.71
## + v18   1   153.85 161.85
## + v10   1   153.88 161.88
## + v16   1   154.00 162.00
## + v12   1   154.05 162.05
## + v4    1   154.32 162.32
## + v15   1   154.42 162.42
##
## Step:  AIC=153.31
## v2 ~ v21 + v14 + v3
##
##        Df Deviance    AIC
## + v7    1   139.13 149.13
## + v11   1   141.94 151.94
## <none>      145.31 153.31
## + v18   1   143.83 153.83
## + v17   1   144.20 154.20
## + v20   1   144.33 154.33
## + v5    1   144.35 154.35
## + v13   1   144.41 154.41
## + v9    1   144.45 154.45
## + v10   1   144.56 154.56
## + v4    1   144.66 154.66
## + v6    1   144.80 154.80
## + v8    1   144.86 154.86
## + v19   1   144.91 154.91
## + v15   1   144.98 154.98
```

```
## + v12   1   145.25 155.25
## + v16   1   145.26 155.26
##
## Step:  AIC=149.13
## v2 ~ v21 + v14 + v3 + v7
##
##        Df Deviance    AIC
## + v11   1   135.61 147.61
## + v18   1   137.11 149.11
## <none>      139.13 149.13
## + v4    1   137.45 149.45
## + v13   1   137.50 149.50
## + v17   1   138.11 150.11
## + v9    1   138.23 150.23
## + v5    1   138.43 150.43
## + v19   1   138.50 150.50
## + v15   1   138.50 150.50
## + v6    1   138.53 150.53
## + v20   1   138.60 150.60
## + v10   1   138.66 150.66
## + v8    1   138.77 150.77
## + v12   1   139.04 151.04
## + v16   1   139.10 151.10
##
## Step:  AIC=147.61
## v2 ~ v21 + v14 + v3 + v7 + v11
##
##        Df Deviance    AIC
## + v18   1   132.88 146.88
## <none>      135.61 147.61
## + v13   1   133.99 147.99
## + v4    1   134.38 148.38
## + v17   1   134.62 148.62
## + v8    1   134.77 148.77
## + v6    1   134.97 148.97
## + v10   1   135.17 149.17
## + v15   1   135.19 149.19
## + v20   1   135.25 149.25
## + v19   1   135.31 149.31
## + v5    1   135.39 149.39
## + v9    1   135.40 149.40
## + v16   1   135.61 149.61
## + v12   1   135.61 149.61
##
## Step:  AIC=146.88
## v2 ~ v21 + v14 + v3 + v7 + v11 + v18
##
##        Df Deviance    AIC
## + v17   1   128.44 144.44
## <none>      132.88 146.88
## + v13   1   131.44 147.44
## + v4    1   131.71 147.71
## + v6    1   131.81 147.81
## + v10   1   131.93 147.93
```

```
## + v15   1   132.10 148.10
## + v8    1   132.11 148.11
## + v9    1   132.50 148.50
## + v5    1   132.59 148.59
## + v20   1   132.72 148.72
## + v16   1   132.73 148.73
## + v19   1   132.75 148.75
## + v12   1   132.82 148.82
##
## Step:  AIC=144.44
## v2 ~ v21 + v14 + v3 + v7 + v11 + v18 + v17
##
##        Df Deviance    AIC
## <none>      128.44 144.44
## + v13   1   126.87 144.87
## + v4    1   127.13 145.13
## + v10   1   127.58 145.58
## + v15   1   127.68 145.68
## + v6    1   127.86 145.86
## + v19   1   127.90 145.90
## + v5    1   128.31 146.31
## + v9    1   128.38 146.38
## + v16   1   128.40 146.40
## + v8    1   128.40 146.40
## + v20   1   128.41 146.41
## + v12   1   128.44 146.44
```

```
mstep2_AIC <- step(m_full, direction="backward", trace = FALSE)
mstep3_AIC <- step(m_full, direction="both", trace = FALSE)
```

```
mstep1_BIC <- step(m_empty, direction="forward", scope = list(upper = m_full), k = log(nrow(data4)), tra
```

```
## Start:  AIC=205.46
## v2 ~ 1
##
##        Df Deviance    AIC
## + v21   1   165.56 176.15
## + v14   1   185.05 195.65
## + v11   1   191.34 201.93
## + v10   1   192.23 202.82
## + v3    1   192.31 202.90
## + v6    1   193.24 203.84
## + v9    1   193.59 204.18
## + v8    1   194.74 205.33
## <none>      200.16 205.46
## + v20   1   195.40 205.99
## + v19   1   198.56 209.16
## + v5    1   198.89 209.49
## + v16   1   198.92 209.52
## + v17   1   199.25 209.85
## + v13   1   199.92 210.52
## + v12   1   199.96 210.56
## + v4    1   200.08 210.67
```

```
## + v7    1   200.16 210.76
## + v18   1   200.16 210.76
## + v15   1   200.16 210.76
##
## Step:  AIC=176.15
## v2 ~ v21
##
##        Df Deviance    AIC
## + v14   1   154.51 170.41
## + v3    1   159.48 175.37
## + v11   1   159.76 175.66
## + v6    1   159.79 175.69
## <none>      165.56 176.15
## + v9    1   161.03 176.92
## + v20   1   162.42 178.32
## + v17   1   163.16 179.05
## + v8    1   163.22 179.11
## + v19   1   163.40 179.29
## + v5    1   163.69 179.59
## + v10   1   164.07 179.96
## + v12   1   164.24 180.13
## + v13   1   164.50 180.39
## + v16   1   164.80 180.70
## + v18   1   164.98 180.88
## + v7    1   165.07 180.97
## + v15   1   165.55 181.45
## + v4    1   165.55 181.45
##
## Step:  AIC=170.41
## v2 ~ v21 + v14
##
##        Df Deviance    AIC
## + v3    1   145.31 166.50
## <none>      154.51 170.41
## + v7    1   149.45 170.64
## + v11   1   150.77 171.96
## + v9    1   151.71 172.90
## + v5    1   152.07 173.26
## + v20   1   152.65 173.84
## + v13   1   152.99 174.18
## + v8    1   153.02 174.21
## + v17   1   153.09 174.28
## + v19   1   153.60 174.79
## + v6    1   153.71 174.90
## + v18   1   153.85 175.04
## + v10   1   153.88 175.07
## + v16   1   154.00 175.19
## + v12   1   154.05 175.24
## + v4    1   154.32 175.52
## + v15   1   154.42 175.62
##
## Step:  AIC=166.5
## v2 ~ v21 + v14 + v3
##
```

```
##        Df Deviance    AIC
## + v7    1   139.13 165.63
## <none>      145.31 166.50
## + v11   1   141.94 168.43
## + v18   1   143.83 170.32
## + v17   1   144.20 170.69
## + v20   1   144.33 170.82
## + v5    1   144.35 170.84
## + v13   1   144.41 170.90
## + v9    1   144.45 170.94
## + v10   1   144.56 171.06
## + v4    1   144.66 171.15
## + v6    1   144.80 171.29
## + v8    1   144.86 171.35
## + v19   1   144.91 171.41
## + v15   1   144.98 171.47
## + v12   1   145.25 171.74
## + v16   1   145.26 171.75
##
## Step:  AIC=165.63
## v2 ~ v21 + v14 + v3 + v7
##
##        Df Deviance    AIC
## <none>      139.13 165.63
## + v11   1   135.61 167.40
## + v18   1   137.11 168.90
## + v4    1   137.45 169.24
## + v13   1   137.50 169.29
## + v17   1   138.11 169.90
## + v9    1   138.23 170.02
## + v5    1   138.43 170.22
## + v19   1   138.50 170.29
## + v15   1   138.50 170.29
## + v6    1   138.53 170.32
## + v20   1   138.60 170.39
## + v10   1   138.66 170.45
## + v8    1   138.77 170.56
## + v12   1   139.04 170.83
## + v16   1   139.10 170.89
```

```
mstep2_BIC <- step(m_full, direction="backward", k = log(nrow(data4)), trace = FALSE)
mstep3_BIC <- step(m_full, direction="both", k = log(nrow(data4)), trace = FALSE)
```

1.4 Model Choosing

In this stepwise regression process, both **AIC** (Akaike Information Criterion) and **BIC** (Bayesian Information Criterion) were used to select the best model. The basic approach for both is the same, with variables being added or removed step by step to optimize the model. However, **AIC** focuses more on model fit, while **BIC** penalizes model complexity more strongly, generally favoring simpler models.

AIC Stepwise Regression:

- The AIC stepwise regression started with an empty model (only the intercept), and variables were added step by step, selecting those that reduced the AIC value.
- The final best model is: ** v2 ~ v21 + v14 + v3 + v7 + v11 + v18 + v17 **
- This model has an AIC value of **144.44**, including more variables, as AIC tends to favor models with better fit, even if they are more complex.

BIC Stepwise Regression:

- The BIC stepwise regression also started with an empty model, but in selecting variables, BIC penalizes model complexity more strongly. Therefore, it tends to select models with fewer variables.
- The final best model is: ** v2 ~ v21 + v14 + v3 + v7 **
- This model has an AIC value of **165.63**, and a smaller BIC, which is why it includes fewer variables compared to the AIC model.

Final Model Choice

Based on the results from both AIC and BIC: - **If the goal is to achieve better model fit and the complexity is not a primary concern**, the **AIC model** should be chosen, as it has a lower AIC value (144.44) and includes more explanatory variables. - **If the goal is to avoid overfitting and prefer a simpler model**, the **BIC model** should be selected, as it includes fewer variables due to the stronger penalty for complexity.

Between the two, **the AIC model** provides a better fit and includes more variables, while **the BIC model** is more parsimonious.

Thus, **the final chosen model** should be the **AIC model with AIC=144.44**: ** v2 ~ v21 + v14 + v3 + v7 + v11 + v18 + v17 **

1.5 Report odds ratio with confidence interval for the most important variables/factors and interpret them

```
summary(mstep1_AIC)
```

```
##
## Call:
## glm(formula = v2 ~ v21 + v14 + v3 + v7 + v11 + v18 + v17, family = binomial,
##     data = data4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9610  -0.5484  -0.2877  -0.1198   2.6921
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.326431   1.607500  -3.313 0.000921 ***
## v21          4.496229   1.014153   4.433 9.27e-06 ***
## v14          3.027648   0.955149   3.170 0.001525 **
## v3           0.042684   0.013707   3.114 0.001845 **
## v7           2.402019   0.888072   2.705 0.006835 **
## v11         -0.014538   0.007083  -2.053 0.040101 *
## v18         -2.253218   0.994392  -2.266 0.023456 *
```

```
## v17          1.854290   0.874168   2.121 0.033904 *
## ---
## 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 192  degrees of freedom
## AIC: 144.44
##
## Number of Fisher Scoring iterations: 6
```

```r
# 1. Extract the model coefficients and standard errors
model_summary <- summary(mstep1_AIC)

# 2. Calculate the odds ratios (exp(coefficient)) and 95% confidence intervals
coefficients <- model_summary$coefficients[, 1]  # Coefficients
std_errors <- model_summary$coefficients[, 2]    # Standard errors

# Calculate the odds ratios (exp of the coefficients)
odds_ratios <- exp(coefficients)

# Calculate the 95% confidence intervals
conf_int_lower <- exp(coefficients - 1.96 * std_errors)
conf_int_upper <- exp(coefficients + 1.96 * std_errors)

# 3. Create a data frame with the results
results <- data.frame(
  Variable = names(coefficients),
  Odds_Ratio = odds_ratios,
  CI_Lower = conf_int_lower,
  CI_Upper = conf_int_upper
)

# Display the results
print(results)
```

```
##                Variable   Odds_Ratio     CI_Lower    CI_Upper
## (Intercept) (Intercept)  0.004861388 2.081749e-04   0.1135252
## v21                 v21 89.678346769 1.228637e+01 654.5632782
## v14                 v14 20.648619829 3.175787e+00 134.2550604
## v3                   v3  1.043608337 1.015944e+00   1.0720256
## v7                   v7 11.045449672 1.937500e+00  62.9687591
## v11                 v11  0.985566747 9.719798e-01   0.9993436
## v18                 v18  0.105060635 1.496223e-02   0.7377066
## v17                 v17  6.387161506 1.151334e+00  35.4335204
```

Based on the `mstep1_AIC` model, we can report the **odds ratios (OR)** and their **95% confidence intervals (CI)** for the most significant variables in predicting survival.

Model Results: v21 (Consciousness Level):
Odds ratio = **89.68** (CI: **12.29 to 654.56**).

Higher levels of consciousness significantly increase the odds of survival by **89.68 times** compared to lower levels.

v14 (Type of Admission):
Odds ratio = **20.65** (CI: **3.18 to 134.26**).
Patients with acute admission are **20.65 times** more likely to survive compared to those with non-acute admission.

v3 (Age):
Odds ratio = **1.04** (CI: **1.016 to 1.072**).
Each year increase in age increases the odds of survival by **4%**. This indicates a small but statistically significant positive effect.

v7 (Cancer):
Odds ratio = **11.05** (CI: **1.94 to 62.97**).
Patients with cancer have **11.05 times** higher odds of survival compared to those without cancer.

v11 (Blood Pressure):
Odds ratio = **0.99** (CI: **0.972 to 0.999**).
Each unit increase in blood pressure slightly decreases the odds of survival by **1%**. The effect is small but statistically significant.

v18 (Blood Carbon Dioxide):
Odds ratio = **0.105** (CI: **0.015 to 0.738**).
Patients with blood carbon dioxide levels above 45 have an **89.5% decrease in odds of survival**, suggesting a strong negative impact.

v17 (Blood pH):
Odds ratio = **6.39** (CI: **1.15 to 35.43**).
Higher blood pH levels (above 7.25) increase the odds of survival by **6.39 times** compared to lower levels.

Key Insights:

1. **Strong positive predictors**:
    - **Consciousness Level (v21)**: Patients with higher consciousness levels are significantly more likely to survive.
    - **Type of Admission (v14)**: Acute admissions are strongly associated with higher survival odds.
    - **Blood pH (v17)**: Maintaining a higher blood pH improves survival odds.
2. **Negative predictors**:
    - **Blood Carbon Dioxide (v18)**: Elevated carbon dioxide levels are associated with substantially lower survival odds.
    - **Blood Pressure (v11)**: Slightly elevated blood pressure reduces survival odds, though the effect is minimal.
3. **Age (v3)**: While age is statistically significant, its effect is relatively small, with each additional year increasing survival odds by 4%.
4. **Cancer (v7)**: Surprisingly, having cancer appears to positively impact survival in this dataset, potentially reflecting specific interventions or biases in the cohort.

Conclusion: The most influential factors in predicting survival include **Consciousness Level (v21)**, **Type of Admission (v14)**, and **Blood pH (v17)**. Conversely, elevated **Blood Carbon Dioxide (v18)** significantly reduces survival odds, indicating the need for managing these levels in critical care settings. This analysis highlights the complex interplay of physiological and contextual factors in determining patient outcomes.

Task2

How well does your chosen model fits the data? In assignment 3,the deviance was used to assess model fit but since the data now is on individual level,deviance is not suitable and instead,the Hosmer-Lemeshow goodness-of-fit test (Agresti 5.2.5) should be used.Perform this test in R(see the R-instruction)and interpret the result.

```
predprob<- predict(mstep1_AIC, type = "response") # Obtain predicted probabilities for the best model
library(ResourceSelection)
```

```
## ResourceSelection 0.3-6   2023-06-27
```

```
hoslem_result <- hoslem.test(data4$v2, predprob, g = 10)  # Use v2 as the response variable

print(hoslem_result)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data4$v2, predprob
## X-squared = 7.2202, df = 8, p-value = 0.5131
```

```
hoslem_result$p.value
```

```
## [1] 0.5130765
```

The **p-value** of **0.5131** from the **Hosmer-Lemeshow test** indicates that the model **fits the data well**. Since the p-value is greater than 0.05, we **fail to reject the null hypothesis**, meaning there is no significant difference between the observed and predicted values. This suggests the model's predictions align closely with the actual outcomes.

Task3

Now we shift the focus to prediction.How well does your multiple logistic regression models predict the outcome?There exists several different metrics of prediction performance for binary responses and here we will use the following:'Accuracy', 'Sensitivity (True positive rate),'Specificity (True negative rate)and 'AUC (area under the ROC curve).To explain these measures we first notice that the model predictions (or fitted values)are probabilities to not survive.If we let patients be classified as 'Survive'or 'Not survive'when theirs predicted probability is below or above a certain cut-off value (threshold value),a confusion matrix (a 2x2 classification table)can be constructed,as shown below.

Actual	**Predicted Survive**	**Predicted Not Survive**	**Total**
Actual Survive	n11	n12	n1.
Actual Not Survive	n21	n22	n2.
Total			n..

Calculate the metrics: - **Accuracy** = $\frac{n_{11}+n_{22}}{n}$ - **Sensitivity (True Positive Rate)** = $\frac{n_{11}}{n_{1.}}$ - **Specificity (True Negative Rate)** = $\frac{n_{22}}{n_{2.}}$

```r
# Apply a threshold of 0.5 to classify as Survive or Not survive
predicted_class <- ifelse(predprob >= 0.5, 1, 0)  # 1 for Survive, 0 for Not Survive

# Construct the confusion matrix
actual_class <- data4$v2  # Adjust to the correct actual variable name if needed

# Confusion matrix
conf_matrix <- table(Actual = actual_class, Predicted = predicted_class)


print(conf_matrix)
```

```
##       Predicted
## Actual   0   1
##      0 157   3
##      1  23  17
```

```r
# Calculate the metrics
n11 <- conf_matrix[1, 1]  # True positives
n12 <- conf_matrix[1, 2]  # False negatives
n21 <- conf_matrix[2, 1]  # False positives
n22 <- conf_matrix[2, 2]  # True negatives

# Calculate accuracy, sensitivity, and specificity
accuracy <- (n11 + n22) / sum(conf_matrix)
sensitivity <- n11 / (n11 + n12)
specificity <- n22 / (n21 + n22)

# Display the results
cat("Accuracy: ", accuracy, "\n")
```

```
## Accuracy:  0.87
```

```r
cat("Sensitivity: ", sensitivity, "\n")
```

```
## Sensitivity:  0.98125
```

```r
cat("Specificity: ", specificity, "\n")
```

```
## Specificity:  0.425
```

Interpretation of the Results:

1. **Confusion Matrix**:
    - **True Negatives (TN)** = 157: These are the cases where the model correctly predicted "Not Survive" (Actual = 0 and Predicted = 0).
    - **False Positives (FP)** = 3: These are the cases where the model incorrectly predicted "Survive" (Actual = 0 but Predicted = 1).
    - **False Negatives (FN)** = 23: These are the cases where the model incorrectly predicted "Not Survive" (Actual = 1 but Predicted = 0).

- **True Positives (TP)** = 17: These are the cases where the model correctly predicted "Survive" (Actual = 1 and Predicted = 1).

2. **Metrics**:
   - **Accuracy = 0.87**: This means that 87% of the total predictions were correct. Accuracy is the overall proportion of correct predictions, including both "Survive" and "Not Survive" cases.
   - **Sensitivity (True Positive Rate) = 0.98125**: This indicates that 98.13% of actual survivors (Actual = 1) were correctly predicted as "Survive" (Predicted = 1). A high sensitivity means the model is good at identifying survivors.
   - **Specificity (True Negative Rate) = 0.425**: This indicates that only 42.5% of actual non-survivors (Actual = 0) were correctly predicted as "Not Survive" (Predicted = 0). A low specificity means that the model is not very good at identifying non-survivors, leading to a relatively high false positive rate.

Conclusion:

- The model is performing well in terms of **sensitivity** (98.13%), meaning it is effective at identifying individuals who survive.
- However, the model has a **low specificity** (42.5%), meaning it struggles to correctly identify individuals who do not survive, as evidenced by a relatively high number of false positives.
- The **accuracy** of 87% suggests that the overall performance is reasonable, but the low specificity indicates that there might be room for improvement, especially in predicting non-survivors.

Task4

As shown above,the sensitivity and specificity dependens on the cut-off setting. A ROC-curve is a plot of sensitivity (true positive rate)against 1-specificity (false postive rate)at each possible cut-off setting.AUC is the area under the ROC curve and can be used as a summary metsure of a binary classifier's performance. AUC=1 indicates perfect classification and AUC=0.5 indicates that the model's performance is no better than random guessing.Create plots of ROC curves and calculate the AUC for the full model and two more models of your choice (see the R-instruction) Which of the model performs best based on AUC?

```
# Step 1: Install and load the necessary libraries
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
##     'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
# Step 2: Obtain predicted probabilities from the models
# For the full model (m_full)
predprob_full <- predict(m_full, type = "response")

#The chosen two models
predprob_model1 <- predict(mstep1_AIC, type = "response")
predprob_model2 <- predict(mstep1_BIC, type = "response")
```

```r
# Step 3: Calculate ROC and AUC for the full model
roc_full <- roc(data4$v2, predprob_full)  # 'v2' is the actual response variable
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

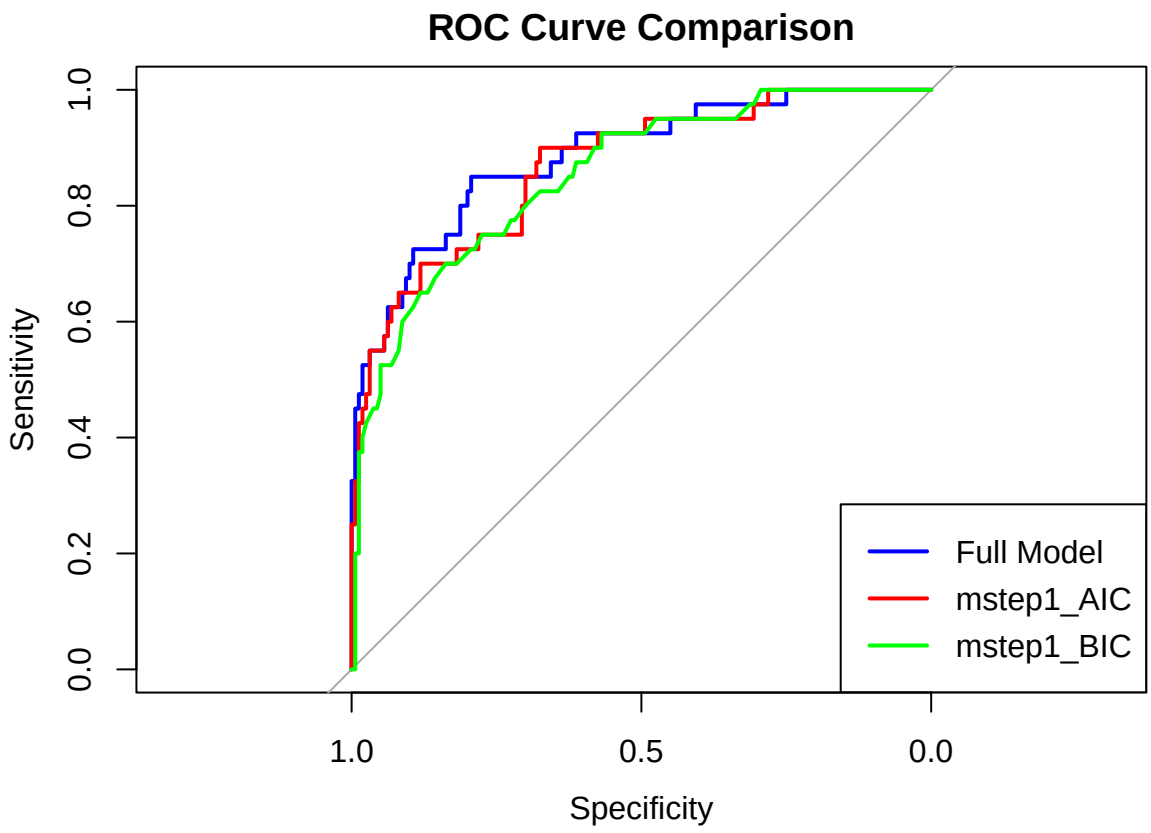
```r
auc_full <- auc(roc_full)

# Step 4: Calculate ROC and AUC for the first additional model
roc_model1 <- roc(data4$v2, predprob_model1)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```r
auc_model1 <- auc(roc_model1)

# Step 5: Calculate ROC and AUC for the second additional model
roc_model2 <- roc(data4$v2, predprob_model2)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```r
auc_model2 <- auc(roc_model2)

# Step 6: Plot the ROC curves for all models
plot(roc_full, col = "blue", main = "ROC Curve Comparison", lwd = 2)
lines(roc_model1, col = "red", lwd = 2)
lines(roc_model2, col = "green", lwd = 2)

# Add a legend
legend("bottomright", legend = c("Full Model", "mstep1_AIC", "mstep1_BIC"),
       col = c("blue", "red", "green"), lwd = 2)
```




```
# Step 7: Display AUC values for comparison
cat("AUC for Full Model: ", auc_full, "\n")
```

```
## AUC for Full Model:  0.8846875
```

```
cat("AUC for mstep1_AIC: ", auc_model1, "\n")
```

```
## AUC for mstep1_AIC:  0.8678125
```

```
cat("AUC for mstep1_BIC: ", auc_model2, "\n")
```

```
## AUC for mstep1_BIC:  0.8532812
```

- **Full Model** performs the best with the highest AUC, indicating the strongest predictive ability.
- **mstep1_AIC** follows closely behind, but with slightly lower performance.
- **mstep1_BIC** has the lowest performance in terms of AUC.

The full model has the highest AUC but these three models' AUC are very close.

Task5

The AUC-values obtained in the previous exercise may be subject to optimistic bias, as they were computed for the same dataset that was used for model fitting. This might result in overfitting,where the model

become well adapted to the training data but may not perform as good to new data.To get a more reliable AUC-value, a validation method can be used.We will use the Leave-One-Out-Cross-Validation (LOOCV).For LOOCV,a model is fitted (or trained)using all observations except one,which is held out for validation.The model is then used to make a prediction on the held-out observation.This process is repeated for each observation,resulting in LOOCV predicted probabilities for each observation (see the R-instruction).These LOOCV predicted probabilities can then be used to calculate AUC.Do this for the full model and two more models of your choice (the same models as in the previous exercise).Which of the model performs best based on LOOCV-adjusted AUC? Compare with the result in the previous exercise. Note:The stepwise procedure should not be performed within the cross-validation 1oop.

```
# Step 1: Create vectors for storing LOOCV predicted probabilities for each model
predprob_LOOCV_full <- numeric(nrow(data4))  # Full model
predprob_LOOCV_AIC <- numeric(nrow(data4))   # AIC model
predprob_LOOCV_BIC <- numeric(nrow(data4))   # BIC model

# Step 2: Perform LOOCV for each model
for (i in 1:nrow(data4)) {

  # Create training and validation sets
  data_training <- data4[-i, ]  # Exclude the i-th observation for training
  data_validation <- data4[i, , drop = FALSE]  # Keep the i-th observation for validation

  # Full model
  m_full <- glm(v2 ~ ., family = binomial, data = data_training)

  # AIC model
  m_AIC <- glm(v2 ~ v21 + v14 + v3 + v7 + v1 + v18 + v17 + v11, family = binomial, data = data_training

  # BIC model
  m_BIC <- glm(v2 ~ v21 + v14 + v3 + v7, family = binomial, data = data_training)

  # Predict the probability for the held-out observation for each model
  predprob_LOOCV_full[i] <- predict(m_full, newdata = data_validation, type = "response")
  predprob_LOOCV_AIC[i] <- predict(m_AIC, newdata = data_validation, type = "response")
  predprob_LOOCV_BIC[i] <- predict(m_BIC, newdata = data_validation, type = "response")
}

# Step 3: Calculate AUC for each model using LOOCV predicted probabilities
library(pROC)

# ROC and AUC for Full Model
roc_full <- roc(data4$v2, predprob_LOOCV_full)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

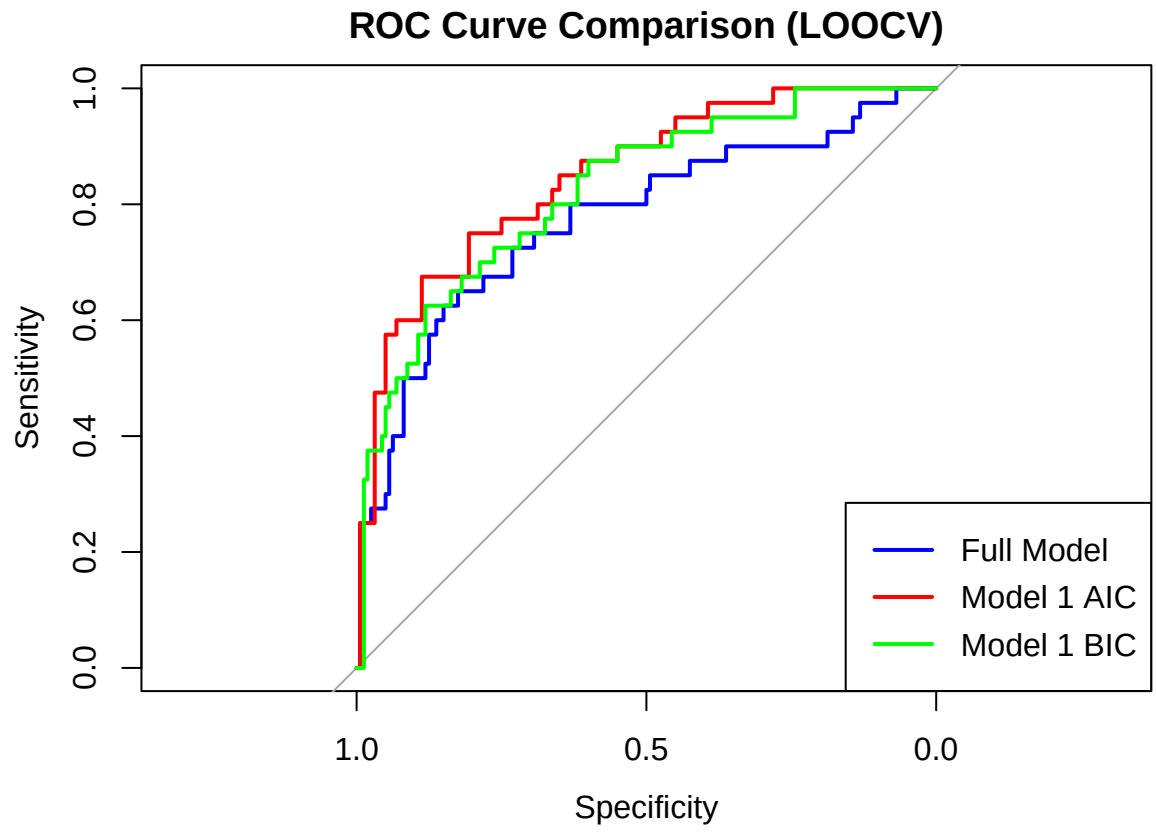
```
auc_full <- auc(roc_full)

# ROC and AUC for AIC Model
roc_AIC <- roc(data4$v2, predprob_LOOCV_AIC)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```r
auc_AIC <- auc(roc_AIC)

# ROC and AUC for BIC Model
roc_BIC <- roc(data4$v2, predprob_LOOCV_BIC)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```r
auc_BIC <- auc(roc_BIC)

# Step 4: Display ROC curves for all models
plot(roc_full, col = "blue", main = "ROC Curve Comparison (LOOCV)", lwd = 2)
lines(roc_AIC, col = "red", lwd = 2)
lines(roc_BIC, col = "green", lwd = 2)

# Add a legend to the plot
legend("bottomright", legend = c("Full Model", "Model 1 AIC", "Model 1 BIC"),
       col = c("blue", "red", "green"), lwd = 2)
```




```r
# Step 5: Display AUC values for comparison
cat("AUC for Full Model: ", auc_full, "\n")
```

```
## AUC for Full Model:  0.7746875
```

```r
cat("AUC for Model 1 (AIC): ", auc_AIC, "\n")
```

```
## AUC for Model 1 (AIC):  0.8495312
```

```r
cat("AUC for Model 1 (BIC): ", auc_BIC, "\n")
```

```
## AUC for Model 1 (BIC):  0.824375
```

- The **AIC model** (AUC = 0.8495312) performs the best, with the strongest discriminative power.
- The **BIC model** (AUC = 0.824375) performs well but is slightly weaker than the AIC model.
- The **Full model** (AUC = 0.7746875) performs relatively poorly, indicating that its predictive ability is not as strong as the AIC and BIC models.

Based on the **LOOCV-adjusted AUC**, the **AIC model** is the best choice.