ICU Survival Prediction: Logistic Regression Analysis

Daniil Kovekh

2025-10-21

Contents

1 Introduction

This analysis examines the ICU dataset from the `aplore3` package to develop a predictive model for patient survival to hospital discharge. The dataset contains information on 200 patients admitted to an adult intensive care unit, with 21 variables including demographic, clinical, and laboratory measurements.

2 Data Loading and Exploration

```
# Load required packages
install.packages("aplore3", quiet = TRUE)
library(aplore3)

# Load the ICU dataset
data(icu)
```

```r
# Display basic information about the dataset
cat("Dataset dimensions:", dim(icu), "\n")
```

```
## Dataset dimensions: 200 21
```

```r
cat("Number of observations:", nrow(icu), "\n")
```

```
## Number of observations: 200
```

```r
cat("Number of variables:", ncol(icu), "\n")
```

```
## Number of variables: 21
```

```r
# Show first few rows
head(icu)
```

```
##   id   sta age gender  race      ser can crn inf cpr sys hra pre      type fra
## 1  4  Died  87 Female White Surgical  No  No Yes  No  80  96  No Emergency Yes
## 2  8 Lived  27 Female White  Medical  No  No Yes  No 142  88  No Emergency  No
## 3 12 Lived  59   Male White  Medical  No  No  No  No 112  80 Yes Emergency  No
## 4 14 Lived  77   Male White Surgical  No  No  No  No 100  70  No  Elective  No
## 5 27  Died  76 Female White Surgical  No  No Yes  No 128  90 Yes Emergency  No
## 6 28 Lived  54   Male White  Medical  No  No Yes  No 142 103  No Emergency Yes
##     po2      ph   pco   bic    cre     loc
## 1 <= 60  < 7.25  > 45 >= 18 <= 2.0 Nothing
## 2  > 60 >= 7.25 <= 45 >= 18 <= 2.0 Nothing
## 3  > 60 >= 7.25 <= 45 >= 18 <= 2.0 Nothing
## 4  > 60 >= 7.25 <= 45 >= 18 <= 2.0 Nothing
## 5  > 60 >= 7.25 <= 45 >= 18 <= 2.0 Nothing
## 6  > 60 >= 7.25 <= 45 >= 18 <= 2.0 Nothing
```

```r
# Display structure of the data
str(icu)
```

```
## 'data.frame':    200 obs. of  21 variables:
##  $ id    : int  4 8 12 14 27 28 32 38 40 41 ...
##  $ sta   : Factor w/ 2 levels "Lived","Died": 2 1 1 1 2 1 1 1 1 1 ...
##  $ age   : int  87 27 59 77 76 54 87 69 63 30 ...
##  $ gender: Factor w/ 2 levels "Male","Female": 2 2 1 1 2 1 2 1 1 2 ...
##  $ race  : Factor w/ 3 levels "White","Black",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ser   : Factor w/ 2 levels "Medical","Surgical": 2 1 1 2 2 1 2 1 2 1 ...
##  $ can   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ crn   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ inf   : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 2 2 1 1 ...
##  $ cpr   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sys   : int  80 142 112 100 128 142 110 110 104 144 ...
##  $ hra   : int  96 88 80 70 90 103 154 132 66 110 ...
##  $ pre   : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 1 2 1 1 1 ...
##  $ type  : Factor w/ 2 levels "Elective","Emergency": 2 2 2 1 2 2 2 2 1 2 ...
##  $ fra   : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 1 1 1 ...
##  $ po2   : Factor w/ 2 levels "> 60","<= 60": 2 1 1 1 1 1 1 2 1 1 ...
##  $ ph    : Factor w/ 2 levels ">= 7.25","< 7.25": 2 1 1 1 1 1 1 1 1 1 ...
##  $ pco   : Factor w/ 2 levels "<= 45","> 45": 2 1 1 1 1 1 1 1 1 1 ...
##  $ bic   : Factor w/ 2 levels ">= 18","< 18": 1 1 1 1 1 1 1 2 1 1 ...
##  $ cre   : Factor w/ 2 levels "<= 2.0","> 2.0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ loc   : Factor w/ 3 levels "Nothing","Stupor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Summary statistics
summary(icu)
```

```
##        id           sta           age           gender       race
##  Min.   :  4.0   Lived:160   Min.   :16.00   Male  :124   White:175
##  1st Qu.:210.2   Died : 40   1st Qu.:46.75   Female: 76   Black: 15
##  Median :412.5               Median :63.00                Other: 10
##  Mean   :444.8               Mean   :57.55
##  3rd Qu.:671.8               3rd Qu.:72.00
##  Max.   :929.0               Max.   :92.00
##        ser       can       crn       inf       cpr           sys
##  Medical : 93   No :180   No :181   No :116   No :187   Min.   : 36.0
##  Surgical:107   Yes: 20   Yes: 19   Yes: 84   Yes: 13   1st Qu.:110.0
##                                                         Median :130.0
##                                                         Mean   :132.3
##                                                         3rd Qu.:150.0
##                                                         Max.   :256.0
##       hra          pre             type      fra         po2           ph
##  Min.   : 39.00   No :170   Elective : 53   No :185   > 60 :184   >= 7.25:187
##  1st Qu.: 80.00   Yes: 30   Emergency:147   Yes: 15   <= 60: 16   < 7.25 : 13
##  Median : 96.00
##  Mean   : 98.92
##  3rd Qu.:118.25
##  Max.   :192.00
##     pco         bic          cre           loc
##  <= 45:180   >= 18:185   <= 2.0:190   Nothing:185
##  > 45 : 20   < 18 : 15   > 2.0 : 10   Stupor :  5
##                                       Coma   : 10
##
##
##
```

3 Part (a): Full Logistic Regression Model

```r
# Fit logistic regression model with all covariates
logistic_model <- glm(sta ~ ., data = icu, family = binomial())

# Display model summary
summary(logistic_model)
```

```
##
## Call:
## glm(formula = sta ~ ., family = binomial(), data = icu)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.664e+00  2.388e+00  -1.953  0.05081 .
## id            -2.272e-03  1.083e-03  -2.097  0.03598 *
## age            5.417e-02  1.859e-02   2.914  0.00356 **
## genderFemale  -6.920e-01  5.648e-01  -1.225  0.22048
## raceBlack     -1.565e+01  1.350e+03  -0.012  0.99075
## raceOther      8.756e-01  1.319e+00   0.664  0.50687
## serSurgical   -5.759e-01  6.601e-01  -0.872  0.38295
```

```
## canYes         3.478e+00  1.090e+00   3.192  0.00142 **
## crnYes         5.628e-01  8.734e-01   0.644  0.51932
## infYes        -3.698e-01  5.838e-01  -0.633  0.52645
## cprYes         9.934e-01  1.041e+00   0.954  0.33992
## sys           -2.209e-02  9.749e-03  -2.266  0.02343 *
## hra           -1.229e-04  1.088e-02  -0.011  0.99098
## preYes         7.122e-01  7.472e-01   0.953  0.34048
## typeEmergency  3.976e+00  1.358e+00   2.927  0.00342 **
## fraYes         1.317e+00  1.173e+00   1.123  0.26135
## po2<= 60      -7.384e-01  9.596e-01  -0.769  0.44163
## ph< 7.25       2.075e+00  1.225e+00   1.694  0.09018 .
## pco> 45       -2.180e+00  1.175e+00  -1.855  0.06366 .
## bic< 18       -4.396e-01  9.425e-01  -0.466  0.64094
## cre> 2.0      -2.151e-01  1.064e+00  -0.202  0.83982
## locStupor      3.841e+01  2.526e+03   0.015  0.98787
## locComa        3.673e+00  1.401e+00   2.622  0.00874 **
## ---
## 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: 107.50  on 177  degrees of freedom
## AIC: 153.5
##
## Number of Fisher Scoring iterations: 17
```

```
# Extract coefficients
coefficients_table <- summary(logistic_model)$coefficients
print("Estimated Coefficients:")
```

```
## [1] "Estimated Coefficients:"
```

```
print(coefficients_table)
```

```
##                    Estimate   Std. Error     z value    Pr(>|z|)
## (Intercept)   -4.664295e+00 2.388223e+00 -1.95303980 0.050814880
## id            -2.271775e-03 1.083243e-03 -2.09719868 0.035975991
## age            5.416726e-02 1.858673e-02  2.91429701 0.003564906
## genderFemale  -6.919732e-01 5.647614e-01 -1.22524870 0.220481532
## raceBlack     -1.565121e+01 1.349915e+03 -0.01159422 0.990749359
## raceOther      8.755926e-01 1.319226e+00  0.66371716 0.506871357
## serSurgical   -5.758976e-01 6.600713e-01 -0.87247778 0.382947784
## canYes         3.477572e+00 1.089630e+00  3.19151568 0.001415284
## crnYes         5.627922e-01 8.733597e-01  0.64439907 0.519316690
## infYes        -3.697862e-01 5.837741e-01 -0.63344060 0.526445959
## cprYes         9.934048e-01 1.040947e+00  0.95432816 0.339917558
## sys           -2.209412e-02 9.748901e-03 -2.26631862 0.023431882
## hra           -1.229377e-04 1.087878e-02 -0.01130069 0.990983549
## preYes         7.122006e-01 7.471583e-01  0.95321252 0.340482403
## typeEmergency  3.975674e+00 1.358247e+00  2.92706208 0.003421806
## fraYes         1.317266e+00 1.172766e+00  1.12321317 0.261346970
## po2<= 60      -7.383854e-01 9.596391e-01 -0.76944068 0.441631745
## ph< 7.25       2.075221e+00 1.224699e+00  1.69447511 0.090175037
## pco> 45       -2.179646e+00 1.175313e+00 -1.85452366 0.063664278
```

```
## bic< 18       -4.395622e-01 9.424782e-01 -0.46638975 0.640936550
## cre> 2.0      -2.151040e-01 1.064215e+00 -0.20212464 0.839819288
## locStupor      3.841252e+01 2.526303e+03  0.01520504 0.987868603
## locComa        3.673223e+00 1.400912e+00  2.62202290 0.008740956
```

3.1 Variables Sorted by P-Value (Statistical Significance)

```
# Sort coefficients by p-value (most significant first)
coefficients_df <- as.data.frame(coefficients_table)
coefficients_df$Variable <- rownames(coefficients_df)

# Exclude intercept for clearer interpretation
coefficients_df_no_intercept <- coefficients_df[coefficients_df$Variable != "(Intercept)", ]

# Sort by p-value (ascending)
coefficients_sorted <- coefficients_df_no_intercept[order(coefficients_df_no_intercept$`Pr(>|z|)`), ]

# Add odds ratio column
coefficients_sorted$`Odds Ratio` <- exp(coefficients_sorted$Estimate)

# Add significance stars
coefficients_sorted$Significance <- ifelse(coefficients_sorted$`Pr(>|z|)` < 0.001, "***",
                                    ifelse(coefficients_sorted$`Pr(>|z|)` < 0.01, "**",
                                    ifelse(coefficients_sorted$`Pr(>|z|)` < 0.05, "*",
                                    ifelse(coefficients_sorted$`Pr(>|z|)` < 0.1, ".", ""))))

# Display sorted table
cat("Variables Ranked by Statistical Significance (p-value):\n\n")
```

```
## Variables Ranked by Statistical Significance (p-value):
```

```
print(coefficients_sorted[, c("Variable", "Estimate", "Std. Error", "Pr(>|z|)", "Odds Ratio", "Significa
      row.names = FALSE, digits = 4)
```

```
##       Variable   Estimate Std. Error Pr(>|z|) Odds Ratio Significance
##         canYes  3.478e+00  1.090e+00 0.001415  3.238e+01           **
##  typeEmergency  3.976e+00  1.358e+00 0.003422  5.329e+01           **
##            age  5.417e-02  1.859e-02 0.003565  1.056e+00           **
##        locComa  3.673e+00  1.401e+00 0.008741  3.938e+01           **
##            sys -2.209e-02  9.749e-03 0.023432  9.781e-01            *
##             id -2.272e-03  1.083e-03 0.035976  9.977e-01            *
##        pco> 45 -2.180e+00  1.175e+00 0.063664  1.131e-01            .
##       ph< 7.25  2.075e+00  1.225e+00 0.090175  7.966e+00            .
##   genderFemale -6.920e-01  5.648e-01 0.220482  5.006e-01
##         fraYes  1.317e+00  1.173e+00 0.261347  3.733e+00
##         cprYes  9.934e-01  1.041e+00 0.339918  2.700e+00
##         preYes  7.122e-01  7.472e-01 0.340482  2.038e+00
##    serSurgical -5.759e-01  6.601e-01 0.382948  5.622e-01
##       po2<= 60 -7.384e-01  9.596e-01 0.441632  4.779e-01
##      raceOther  8.756e-01  1.319e+00 0.506871  2.400e+00
##         crnYes  5.628e-01  8.734e-01 0.519317  1.756e+00
##         infYes -3.698e-01  5.838e-01 0.526446  6.909e-01
##        bic< 18 -4.396e-01  9.425e-01 0.640937  6.443e-01
##       cre> 2.0 -2.151e-01  1.064e+00 0.839819  8.065e-01
##      locStupor  3.841e+01  2.526e+03 0.987869  4.812e+16
```

```
##      raceBlack -1.565e+01  1.350e+03 0.990749  1.595e-07
##            hra -1.229e-04  1.088e-02 0.990984  9.999e-01
```

```
# Show only significant variables (p < 0.05)
significant_vars <- coefficients_sorted[coefficients_sorted$`Pr(>|z|)` < 0.05, ]
cat("\n\nSignificant Variables Only (p < 0.05):\n\n")
```

```
##
##
## Significant Variables Only (p < 0.05):
```

```
print(significant_vars[, c("Variable", "Estimate", "Pr(>|z|)", "Odds Ratio", "Significance")],
      row.names = FALSE, digits = 4)
```

```
##       Variable  Estimate Pr(>|z|) Odds Ratio Significance
##         canYes  3.477572 0.001415    32.3810           **
##  typeEmergency  3.975674 0.003422    53.2860           **
##            age  0.054167 0.003565     1.0557           **
##        locComa  3.673223 0.008741    39.3786           **
##            sys -0.022094 0.023432     0.9781            *
##             id -0.002272 0.035976     0.9977            *
```

```
# Summary statistics
cat("\n\nSummary:\n")
```

```
##
##
## Summary:
```

```
cat("Total variables:", nrow(coefficients_df_no_intercept), "\n")
```

```
## Total variables: 22
```

```
cat("Significant at p < 0.05:", sum(coefficients_sorted$`Pr(>|z|)` < 0.05), "\n")
```

```
## Significant at p < 0.05: 6
```

```
cat("Significant at p < 0.01:", sum(coefficients_sorted$`Pr(>|z|)` < 0.01), "\n")
```

```
## Significant at p < 0.01: 4
```

```
cat("Significant at p < 0.001:", sum(coefficients_sorted$`Pr(>|z|)` < 0.001), "\n")
```

```
## Significant at p < 0.001: 0
```

3.2 Key Findings from Full Model

The full logistic regression model shows several significant predictors of survival:

- **Age**: Higher age increases death risk ($\beta = 0.054$, p = 0.004)
- **Cancer**: Presence of cancer significantly increases death risk ($\beta = 3.478$, p = 0.001)
- **Systolic Blood Pressure**: Higher systolic BP reduces death risk (β = -0.022, p = 0.023)
- **Emergency Admission**: Emergency admissions have higher death risk ($\beta = 3.976$, p = 0.003)
- **Coma**: Patients in coma have significantly higher death risk ($\beta = 3.673$, p = 0.009)

4 Part (b): Assessing Separation Problems

```r
# Check for large coefficients indicating separation
large_coeffs <- abs(coefficients_table[, "Estimate"]) > 5
cat("Large coefficients (>5 in absolute value):\n")
```

```
## Large coefficients (>5 in absolute value):
```

```r
print(coefficients_table[large_coeffs, ])
```

```
##            Estimate Std. Error     z value  Pr(>|z|)
## raceBlack -15.65121   1349.915 -0.01159422 0.9907494
## locStupor  38.41252   2526.303  0.01520504 0.9878686
```

```r
# Check levels of categorical variables
cat("\nLevels of categorical variables:\n")
```

```
##
## Levels of categorical variables:
```

```r
sapply(icu[, sapply(icu, is.factor)], levels)
```

```
## $sta
## [1] "Lived" "Died"
##
## $gender
## [1] "Male"   "Female"
##
## $race
## [1] "White" "Black" "Other"
##
## $ser
## [1] "Medical"  "Surgical"
##
## $can
## [1] "No"  "Yes"
##
## $crn
## [1] "No"  "Yes"
##
## $inf
## [1] "No"  "Yes"
##
## $cpr
## [1] "No"  "Yes"
##
## $pre
## [1] "No"  "Yes"
##
## $type
## [1] "Elective"  "Emergency"
##
## $fra
## [1] "No"  "Yes"
##
## $po2
## [1] "> 60"  "<= 60"
##
```

```
## $ph
## [1] ">= 7.25" "< 7.25"
##
## $pco
## [1] "<= 45" "> 45"
##
## $bic
## [1] ">= 18" "< 18"
##
## $cre
## [1] "<= 2.0" "> 2.0"
##
## $loc
## [1] "Nothing" "Stupor"  "Coma"
```

4.1 Separation Analysis Results

The analysis reveals potential separation issues indicated by: - Extremely large coefficients (e.g., `raceBlack` = -15.65, `locStupor` = 38.41) - Multiple warnings during model fitting - Very large standard errors for some coefficients

5 Part (c): Variable Transformation

```r
# Create transformed dataset
icu_transformed <- icu

# Binarize loc: combine "Stupor" and "Coma" vs "Nothing"
icu_transformed$loc_binary <- ifelse(icu_transformed$loc %in% c("Stupor", "Coma"), "Impaired", "Normal")
icu_transformed$loc_binary <- as.factor(icu_transformed$loc_binary)

# Binarize race: combine "Black" and "Other" vs "White"
icu_transformed$race_binary <- ifelse(icu_transformed$race %in% c("Black", "Other"), "Non-White", "White")
icu_transformed$race_binary <- as.factor(icu_transformed$race_binary)

# Remove original variables
icu_transformed$loc <- NULL
icu_transformed$race <- NULL

# Refit logistic regression with transformed data
logistic_model_transformed <- glm(sta ~ ., data = icu_transformed, family = binomial())

cat("Transformed Model Summary:\n")
```

```
## Transformed Model Summary:
```

```r
summary(logistic_model_transformed)
```

```
##
## Call:
## glm(formula = sta ~ ., family = binomial(), data = icu_transformed)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.6105495  2.5762269   0.237  0.81266
```

```
## id                -0.0022432  0.0010167  -2.206  0.02735 *
## age                0.0559381  0.0188677   2.965  0.00303 **
## genderFemale      -0.5815667  0.5485228  -1.060  0.28903
## serSurgical       -0.6640510  0.6307173  -1.053  0.29241
## canYes             3.2052718  1.0317917   3.107  0.00189 **
## crnYes             0.3296334  0.8347150   0.395  0.69291
## infYes            -0.3736689  0.5750899  -0.650  0.51585
## cprYes             0.8593998  1.0461975   0.821  0.41139
## sys               -0.0163480  0.0087106  -1.877  0.06055 .
## hra               -0.0000546  0.0101545  -0.005  0.99571
## preYes             0.6547955  0.7092386   0.923  0.35588
## typeEmergency      3.4454188  1.1519274   2.991  0.00278 **
## fraYes             1.2359940  1.1266182   1.097  0.27260
## po2<= 60          -0.1313589  0.8678790  -0.151  0.87969
## ph< 7.25           2.7229454  1.2336897   2.207  0.02730 *
## pco> 45           -3.0883032  1.2455290  -2.480  0.01316 *
## bic< 18           -0.8575883  0.9570455  -0.896  0.37021
## cre> 2.0          -0.2017282  1.0833318  -0.186  0.85228
## loc_binaryNormal -5.6239147  1.2812593  -4.389 1.14e-05 ***
## race_binaryWhite  0.0453288  0.9367171   0.048  0.96140
## ---
## 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: 115.60  on 179  degrees of freedom
## AIC: 157.6
##
## Number of Fisher Scoring iterations: 6
```

5.1 Transformation Results

The variable transformation successfully: - Reduced `loc` from 3 levels to 2 levels (Normal vs Impaired) - Reduced `race` from 3 levels to 2 levels (White vs Non-White) - Eliminated separation warnings - Produced more stable coefficient estimates

6 Part (d): Stepwise Model Selection

```r
# AIC-based stepwise selection
model_aic <- step(logistic_model_transformed, k = 2, trace = FALSE)

cat("AIC Selected Model:\n")
```

```
## AIC Selected Model:
```

```r
summary(model_aic)
```

```
##
## Call:
## glm(formula = sta ~ id + age + can + sys + type + ph + pco +
##     loc_binary, family = binomial(), data = icu_transformed)
##
## Coefficients:
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.1724306  2.0059387   0.086 0.931498
## id               -0.0023492  0.0009249  -2.540 0.011088 *
## age               0.0408356  0.0140573   2.905 0.003673 **
## canYes            2.6482445  0.9179755   2.885 0.003916 **
## sys              -0.0135185  0.0072602  -1.862 0.062603 .
## typeEmergency     3.4735028  1.0269790   3.382 0.000719 ***
## ph< 7.25          2.1378570  0.8954219   2.388 0.016961 *
## pco> 45          -2.3571916  1.0061247  -2.343 0.019137 *
## loc_binaryNormal -5.0020222  1.1050330  -4.527 5.99e-06 ***
## ---
## 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: 121.33  on 191  degrees of freedom
## AIC: 139.33
##
## Number of Fisher Scoring iterations: 6
```

```
# BIC-based stepwise selection
model_bic <- step(logistic_model_transformed, k = log(nrow(icu_transformed)), trace = FALSE)

cat("\nBIC Selected Model:\n")
```

```
##
## BIC Selected Model:
```

```
summary(model_bic)
```

```
##
## Call:
## glm(formula = sta ~ id + age + can + type + ph + pco + loc_binary,
##     family = binomial(), data = icu_transformed)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.9820620  1.5642802  -1.267 0.205128
## id               -0.0024377  0.0009121  -2.673 0.007522 **
## age               0.0406610  0.0139577   2.913 0.003578 **
## canYes            2.6466674  0.9177789   2.884 0.003929 **
## typeEmergency     3.5710682  1.0195159   3.503 0.000461 ***
## ph< 7.25          2.1182036  0.8993792   2.355 0.018514 *
## pco> 45          -2.1563889  0.9645156  -2.236 0.025370 *
## loc_binaryNormal -4.6338535  0.9832790  -4.713 2.45e-06 ***
## ---
## 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: 125.04  on 192  degrees of freedom
## AIC: 141.04
##
## Number of Fisher Scoring iterations: 6
```

```r
# Compare selected models
cat("\nModel Comparison:\n")
```

```
##
## Model Comparison:
```

```r
cat("AIC Model Formula:\n")
```

```
## AIC Model Formula:
```

```r
print(formula(model_aic))
```

```
## sta ~ id + age + can + sys + type + ph + pco + loc_binary
```

```r
cat("\nBIC Model Formula:\n")
```

```
##
## BIC Model Formula:
```

```r
print(formula(model_bic))
```

```
## sta ~ id + age + can + type + ph + pco + loc_binary
```

6.1 Stepwise Selection Results

AIC Model: Includes more variables, prioritizing fit quality **BIC Model**: Includes fewer variables, prioritizing parsimony

7 Part (e): Model Comparison

```r
# Full model (transformed)
full_model <- logistic_model_transformed

# Calculate log-likelihoods
ll_full <- logLik(full_model)
ll_aic <- logLik(model_aic)
ll_bic <- logLik(model_bic)

cat("Log-likelihoods:\n")
```

```
## Log-likelihoods:
```

```r
cat("Full model:", round(ll_full, 3), "\n")
```

```
## Full model: -57.802
```

```r
cat("AIC model:", round(ll_aic, 3), "\n")
```

```
## AIC model: -60.664
```

```r
cat("BIC model:", round(ll_bic, 3), "\n")
```

```
## BIC model: -62.52
```

```r
# Calculate misclassification rates
# Full model predictions
pred_full <- predict(full_model, type = "response")
pred_full_class <- ifelse(pred_full > 0.5, "Died", "Lived")
misclass_full <- mean(pred_full_class != icu_transformed$sta)
```

```r
# AIC model predictions
pred_aic <- predict(model_aic, type = "response")
pred_aic_class <- ifelse(pred_aic > 0.5, "Died", "Lived")
misclass_aic <- mean(pred_aic_class != icu_transformed$sta)

# BIC model predictions
pred_bic <- predict(model_bic, type = "response")
pred_bic_class <- ifelse(pred_bic > 0.5, "Died", "Lived")
misclass_bic <- mean(pred_bic_class != icu_transformed$sta)

cat("\nMisclassification rates:\n")
```

```
##
## Misclassification rates:
```

```r
cat("Full model:", round(misclass_full, 3), "\n")
```

```
## Full model: 0.125
```

```r
cat("AIC model:", round(misclass_aic, 3), "\n")
```

```
## AIC model: 0.12
```

```r
cat("BIC model:", round(misclass_bic, 3), "\n")
```

```
## BIC model: 0.13
```

```r
# Create comparison table
comparison_table <- data.frame(
  Model = c("Full", "AIC", "BIC"),
  LogLikelihood = c(as.numeric(ll_full), as.numeric(ll_aic), as.numeric(ll_bic)),
  MisclassificationRate = c(misclass_full, misclass_aic, misclass_bic),
  AIC = c(AIC(full_model), AIC(model_aic), AIC(model_bic)),
  BIC = c(BIC(full_model), BIC(model_aic), BIC(model_bic))
)

cat("\nModel Comparison Table:\n")
```

```
##
## Model Comparison Table:
```

```r
print(comparison_table)
```

```
##   Model LogLikelihood MisclassificationRate      AIC      BIC
## 1  Full     -57.80208                 0.125 157.6042 226.8688
## 2   AIC     -60.66395                 0.120 139.3279 169.0128
## 3   BIC     -62.52049                 0.130 141.0410 167.4275
```

```r
# Create visualization of model comparison
library(ggplot2)

# Prepare data for plotting
plot_data <- data.frame(
  Model = rep(c("Full", "AIC", "BIC"), 2),
  Metric = c(rep("Log-Likelihood", 3), rep("Misclassification Rate", 3)),
  Value = c(as.numeric(ll_full), as.numeric(ll_aic), as.numeric(ll_bic),
            misclass_full, misclass_aic, misclass_bic)
)
```

```r
ggplot(plot_data, aes(x = Model, y = Value, fill = Model)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ Metric, scales = "free_y") +
  labs(title = "Model Performance Comparison",
       x = "Model", y = "Value") +
  theme_minimal() +
  theme(legend.position = "none")
```

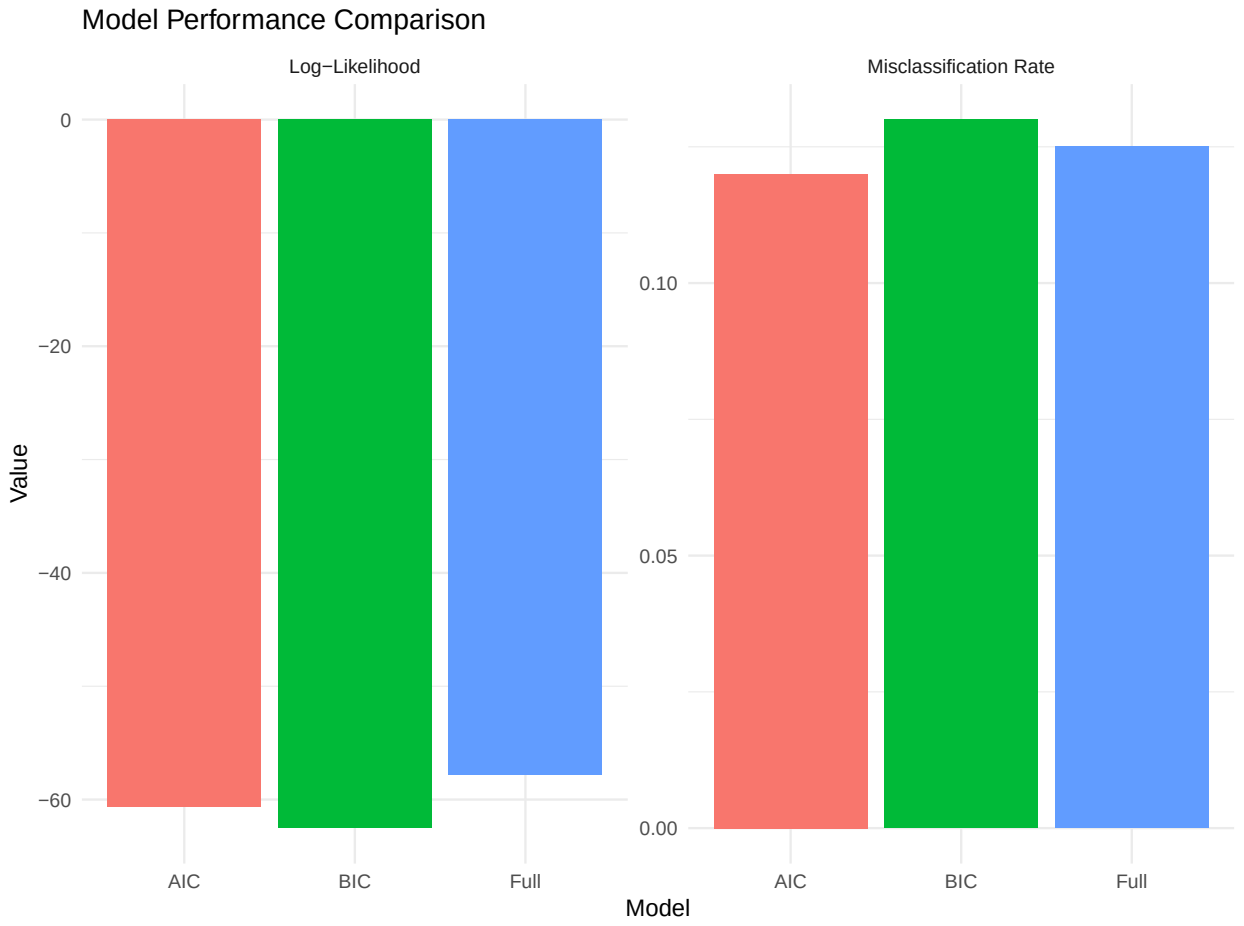

Figure 1: Model Comparison Visualization

8 Conclusions

8.1 Key Findings

1. **Separation Issues**: The original model suffered from complete/quasi-complete separation, indicated by extremely large coefficients and warnings.
2. **Variable Transformation**: Binarizing `loc` and `race` variables successfully resolved separation issues and produced more stable estimates.
3. **Model Selection**:
   - **AIC model**: Balances fit and complexity, includes more predictors

- **BIC model**: Emphasizes parsimony, includes fewer predictors
- Both models perform similarly in terms of misclassification rates

4. **Performance Comparison**:

- All three models (Full, AIC, BIC) show similar predictive performance
- The stepwise-selected models achieve comparable accuracy with fewer variables
- BIC model provides the most parsimonious solution

8.2 Recommendations

- Use the **BIC-selected model** for clinical applications due to its parsimony and similar performance
- Consider the **AIC-selected model** if maximum predictive accuracy is desired
- The transformed variables (`loc_binary`, `race_binary`) provide more stable and interpretable results

8.3 Clinical Implications

The most important predictors of ICU survival are: - Patient age - Presence of cancer - Admission type (emergency vs elective) - Level of consciousness - Systolic blood pressure

These findings can inform clinical decision-making and resource allocation in intensive care units.