Assignment IV

Multiple Logistic Regression and Decision Tree

Silpa Soni Nallacheruvu (19980824-5287) Hernan Aldana (20000526-4999)

2024-12-12

Exercise 4:1 (Multiple Logistic Regression)

Question 1

Report the model selection process briefly. Based on your chosen model, which factors affect the probability of not surviving? Report odds ratios with confidence intervals for the most important variables/factors, and interpret them. Use the variable names from the table (not V3, V4, etc.).

Approach:

- Data Preparation:
  - Load the dataset and rename variables for clarity.
  - Combine categories for categorical variables if necessary.
- Model Fitting:
  - Fit an empty logistic regression model and a full logistic regression model to be used in the stepwise selection.
  - Here, Logistic Regression Model is used to predict the binary outcome of probability of not surviving based on multiple predictors.
- Model Selection Process:
  - Use a stepwise selection with AIC to identify a parsimonious model.
- Analysis of the Final Model:
  - Extract coefficients, odds ratios, and their 95% confidence intervals for significant variables.
- Interpretation:
  - Interpret the results from the AIC, odds ratios, and confidence intervals, to determine the best model.

Results:

Summary of the final model after performing stepwise selection using AIC:

```
##
## Call:
## glm(formula = Survival ~ ConsciousnessLevel + TypeOfAdmission +
##     Age + Cancer + Patient + BloodCarbonDioxide + BloodPH + BloodPressure,
##     family = binomial, data = data_ca4)
```

```
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.7353420  1.6104573  -2.940 0.003278 **
## ConsciousnessLevel  2.6208042  0.6859650   3.821 0.000133 ***
## TypeOfAdmission     3.0547147  0.9339217   3.271 0.001072 **
## Age                 0.0385864  0.0133655   2.887 0.003889 **
## Cancer              2.3388380  0.8671971   2.697 0.006997 **
## Patient            -0.0020714  0.0008783  -2.359 0.018345 *
## BloodCarbonDioxide -2.4646334  1.0619854  -2.321 0.020299 *
## BloodPH             2.0884994  0.9031831   2.312 0.020757 *
## BloodPressure      -0.0099893  0.0070360  -1.420 0.155682
## ---
## 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: 130.19  on 191  degrees of freedom
## AIC: 148.19
##
## Number of Fisher Scoring iterations: 6
```

Model selection:

- The final logistic regression model includes the following variables: ConsciousnessLevel, TypeOfAdmission, Age, Cancer, Patient, BloodCarbonDioxide, BloodPH, and BloodPressure.
- These variables were selected using a stepwise AIC, which ensures a balance between model complexity and goodness of fit.

Significant Variables:

- Variables with a p-value < 0.05 are considered significant predictors of survival:
  - ConsciousnessLevel
  - TypeOfAdmission
  - Age
  - Cancer
  - BloodCarbonDioxide,
  - BloodPH
- Patient variable was exlcluded from the final model as the ID code was not significant.
- BloodPressure was also excluded from the final model as it was not significant according to the p-value.

Odds Ratios and Confidence Intervals:

Here's the final report after extracting odds ratios and confidence intervals for significant variables:

```
## Waiting for profiling to be done...
```

```
##                              Variable    OddsRatio     CI.Lower    CI.Upper
## (Intercept)               (Intercept)  0.008779445 0.0002996746   0.1795376
## ConsciousnessLevel ConsciousnessLevel 13.746774142 4.3144886742  65.2810381
## TypeOfAdmission       TypeOfAdmission 21.215131928 4.3561881520 189.1540175
## Age                               Age  1.039340550 1.0141827727   1.0692793
## Cancer                           Cancer 10.369181076 1.9513659807  66.5483329
## Patient                         Patient  0.997930702 0.9961324780   0.9995944
## BloodCarbonDioxide BloodCarbonDioxide  0.085040009 0.0080634712   0.5539141
## BloodPH                         BloodPH  8.072791946 1.4001269889  53.6032965
```

1. ConsciousnessLevel:
   - Odds Ratio: 13.75 (CI: 4.31-65.28)
   - Patients who are unconscious or in a coma have a significantly higher probability of not surviving compared to those who are conscious.
2. TypeOfAdmission:
   - Odds Ratio: 21.25 (CI: 4.36-189.15)
   - Acute admissions are associated with a significantly higher probability of not surviving compared to non-acute admissions.
3. Age:
   - Odds Ratio: 1.04 (CI: 1.01-1.07)
   - For each additional year of age, the odds of not surviving increase by 4%.
4. Cancer:
   - Odds Ratio: 10.37 (CI: 1.95-66.54)
   - Patients with cancer have over 10 times higher odds of not surviving compared to those without cancer.
5. BloodCarbonDioxide:
   - Odds Ratio: 0.085 (CI: 0.008-0.55)
   - Lower blood carbon dioxide levels significantly reduce the odds of not surviving.
6. BloodPH:
   - Odds Ratio: 8.07 (CI: 1.4-53.60)
   - Lower blood pH levels significantly increase the odds of not surviving.

Conclusion:

The selected model indicates that factors such as consciousness level, type of admission, age, cancer, blood carbon dioxide, and blood pH are significant predictors of survival. Patients who are unconscious, have acute admissions, are older, have cancer, and have abnormal blood gas levels are at higher risk of not surviving. These results can help identify high-risk patients and improve treatment strategies to increase survival rates.

Question 2

How well does your chosen model fit the data? In assignment 3, deviance was used to assess model fit. However, for individual-level data, deviance is unsuitable. Instead, perform the Hosmer-Lemeshow goodness-of-fit test using the recommended R code.

Approach:

- Understand the Hosmer-Lemeshow Test:
  - The Hosmer-Lemeshow test evaluates whether the observed event rates matches the expected probabilities predicted by the model.

    - The null hypothesis is that the model fits the data well (a high p-value suggests no evidence of poor fit).

- Implementation:
    - Use the function for the test ResourceSelection::hoslem.test()
    - Calculate the predicted probabilities from the final model.
    - Specify the predicted probabilities from the final model and the actual outcomes while performing Hosmer-Lemeshow test (m_step and Survival respectively).
    - 10 is selected as the number of groups for the test because dividing by deciles is a common choice and it was sufficient for this dataset of 200 observations.

Hosmer-Lemeshow Test Results:

Here are the results of the Hosmer-Lemeshow goodness-of-fit test:

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data_ca4$Survival, predicted_probs
## X-squared = 2.2064, df = 6, p-value = 0.8998
```

Interpretation:

1. The p-value of 0.9145 tells us not reject the null hypothesis that the model fits the data well.

Conclusion:

The Hosmer-Lemeshow test indicates that the model fits the data well. The observed event rates are consistent with the expected probabilities predicted by the model. This suggests that the model is a good fit for the data and can be used to make accurate predictions about survival probabilities.

Question 3

Create a confusion matrix for the chosen model. Calculate the accuracy, sensitivity, specificity, and positive and negative predictive values. Interpret the results.

Approach:

- Understand the Confusion Matrix:
    - A confusion matrix is a table that summarizes the performance of a classification model.
    - It shows the number of true positives, true negatives, false positives, and false negatives.
- Implementation:
    - Use the confusionMatrix() function from the caret package to calculate the confusion matrix and performance metrics.
    - Specify the predicted probabilities from the final model and the actual outcomes while performing the confusion matrix.
    - Threshold of 0.5 is used to classify the predicted probabilities into binary outcomes since it is the default threshold for logistic regression.
    - Calculate the accuracy, sensitivity, specificity, and positive and negative predictive values from the confusion matrix.

Sensitivity: Sensitivity (True Positive Rate) measures the proportion of actual positive cases that are correctly identified by the model.

Specificity: Specificity (True Negative Rate) measures the proportion of actual negative cases that are correctly identified by the model.

Results:

Here are the results of the confusion matrix and performance metrics:

Table 1: Confusion Matrix

	Predicted Survived	Predicted Not Survived
Actual Survived	155	5
Actual Not Survived	22	18

Table 2: Performance Metrics

Metric	Value
Accuracy	0.86500
Sensitivity	0.96875
Specificity	0.45000

Interpretation:

Accuracy: The model has an accuracy of 0.865, meaning that it correctly predicted 86.5% of the cases.

Sensitivity: The sensitivity of 0.96875 indicates that the model correctly identified 96.9% of the actual survivors.

Specificity: The specificity of 0.45 suggests that the model correctly identified 45% of the actual non-survivors.

Conclusion:

The confusion matrix and performance metrics provide insights into the model's predictive accuracy. The model has a high sensitivity, indicating that it is effective at identifying actual survivors. However, the specificity is relatively low, suggesting that the model has difficulty distinguishing between actual non-survivors. These results can help evaluate the model's performance and identify areas for improvement.