Scott Jue
MSDS 410
2/19/2023

Modeling Assignment #7: Logistic Regression Basics

Introduction

In this analysis we will be modeling data obtained from a sample of 200 subjects randomly selected from a much larger study on the survival of patients following admission to an adult intensive care unit (ICU) in a major metropolitan city. The objective of this assignment is to use this data to construct a logistic regression model to predict the probability of survival of patients through an adult ICU experience. A logistic regression model is appropriate for this kind of analysis over an OLS regression model due to the target response variable, STA, being a dichotomous variable. As such, this analysis will explore how well a logistic regression model fits the data with certain predictor variables as well as the logit transformation of the model. This type of modeling introduces the use of conditional probability and odds for the likelihood of an event occurring given the occurrence of another event. Therefore, the model will predict the probability that the patient does not survive after being admitted in to an adult ICU (i.e. STA = 1). Then we can add in other conditions such as age, admission type, etc. to see how this impacts the patients' probability.

Tasks

1. The population of interest for this analysis is adult ICU patients in a major metropolitan city. In the original sample data there are observations that are younger than 18 years old. Therefore, I have removed any observations with an AGE that is less than 18 as they are not considered adults. This results in 195 observations remaining. Of these 195 observations there are a total of 40 patients that died and 155 patients that survived. Then, I also looked at RACE and noticed that 172 observations are white patients, so the majority of the sample is of white patients. Therefore, the population of interest can be further defined as also a predominately white population. Additionally, the sample population has a mean age of 56 years old and ranges from 18 to 92 years old.

2. In analyzing the gender (SEX) variable I created a 2x2 contingency table that relates the patient's gender (SEX) to Status (STA). Additionally, the odds and probabilities of survival for males and females have also been computed along with the odds ratio of survival that compares them.

Contingency Table:

	Female	Male
Lived	58	97
Died	16	24

Probability and Odds Ratio:

	Probability of Survival	Odds
Men	0.802	4.051
Women	0.784	3.63

The odds ratio of survival that compares males to females is 1.116. This means that odds of survival is 1.116 times greater if the ICU patient is male compared to female. The survival probability and odds ratio for both men and women are similar with men being slightly higher. Because of the similarities, gender does not seem to be a distinguishable characteristic in determining the survival outcome of ICU patients.

3. In analyzing the Type of Admission (TYP) variable I created a 2x2 contingency table that relates Type of Admission (TYP) to Status (STA). Additionally, the odds and probabilities of survival among the different Types of Admission have also been computed along with the odds ratio of survival that compares them.

Contingency Table:

	Elective	Emergency
Lived	51	104
Died	2	38

Probability and Odds Ratio:

	Probability of Survival	Odds
Elective	0.962	25.316
Emergency	0.732	2.731

The odds ratio of survival that compares elective admission type to emergency admission type is 9.27. This can be interpreted as the odds of survival is 9.27 times greater if the ICU patient's admission type is elective compared to emergency. This could be a variable of interest that could add predictive value to the model.

4. Next, I will create a logistic regression model using AGE as the predictor variable.

a. The general equation for the logistic regression model of STA (Y) using AGE (X) and the general equation for the logit transformation of this logistic regression model can be seen below.

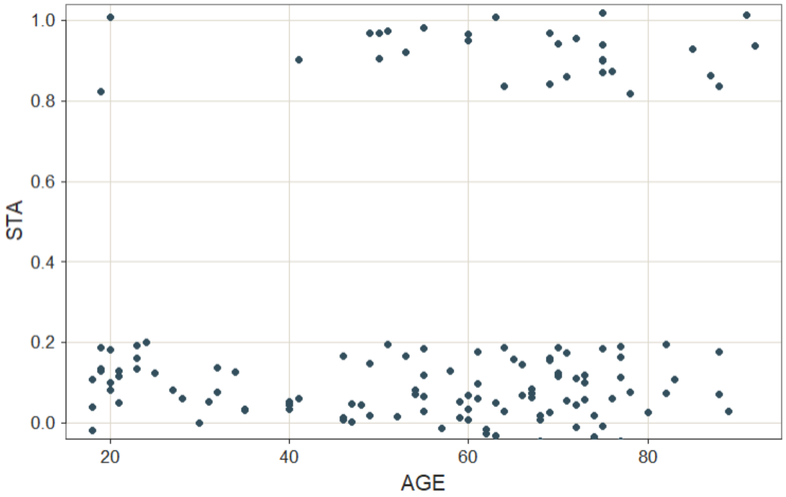
Logistic Regression model of STA (Y) using AGE(X):

$$\pi_{STA} = \frac{e^{(\beta_0+\beta_1*AGE)}}{1+e^{(\beta_0+\beta_1*AGE)}}$$

Logit Transformation:

$$logit(\pi_{STA}) = \log(\frac{\pi_{STA}}{1-\ \pi_{STA}})$$

b. To determine if the patient's age would be a good discriminator between levels of STA, I created the below scatterplot to visualize the relationship.



Based on the scatterplot, it appears that age seems to be a good discriminator between levels of STA. Of the ICU patients that died, a large majority are older patients who are older than 50 years old. Except for 5 patients, almost all of the patients younger than 50 years old survived. Therefore, age has the potential to be a predictor for STA(Y).

c. With the below conditions, I created a new variable AGE_CAT that discretized the age variable.

AGE_CAT = 1 if AGE is in the interval [15,24]

AGE_CAT = 2 if AGE is in the interval [25,34]

AGE_CAT = 3 if AGE is in the interval 3 = [35,44]

AGE_CAT = 4 if AGE is in the interval 4 = [45,54]

AGE_CAT = 5 if AGE is in the interval 5 = [55,64]

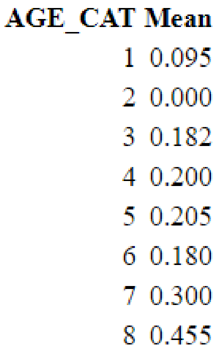
AGE_CAT = 6 if AGE is in the interval 6 = [65,74]

AGE_CAT = 7 if AGE is in the interval 7 = [75,84]

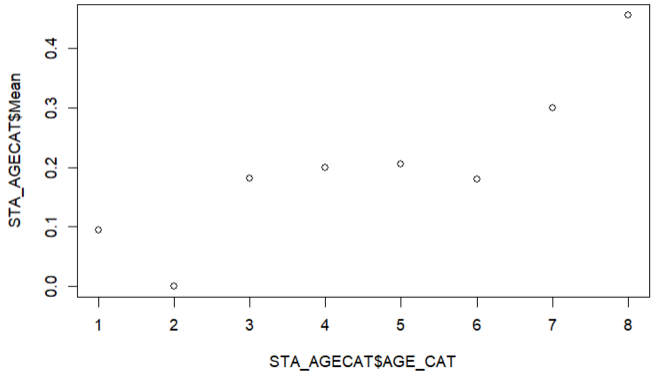
AGE_CAT = 8 if AGE is in the interval 8 = [85,94]

AGE_CAT = 9 if AGE is in the interval 9 = 95 and over

Using this categorical variable, I computed the STA mean (i.e. proportion) for each of the age intervals. Below is a summary of the AGE_CAT and the STA means for each category. Additionally, these means and age categories have been plotted in the scatterplot below. We can see that the older age categories (7 and 8) have higher STA means.



AGE_CAT	Mean
1	0.095
2	0.000
3	0.182
4	0.200
5	0.205
6	0.180
7	0.300
8	0.455



d. Next, I fit a a logistic regression model to predict STA using the original continuous AGE variable. Below is the summary of the coefficients for the model.

Model Coefficient Summary Table:

```
Call:
glm(formula = STA ~ AGE, family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9379  -0.7396  -0.6304  -0.4146   2.2451

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) -2.92275    0.71242  -4.103 0.0000409 ***
AGE          0.02560    0.01081   2.368    0.0179 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 197.90  on 194  degrees of freedom
Residual deviance: 191.55  on 193  degrees of freedom
AIC: 195.55

Number of Fisher Scoring iterations: 4
```

The intercept for the logistic regression model is -2.923. This means that log of odds of the STA variable being equal to 1 (i.e. ICU patient dying) is equal to -2.923 or a probability of .05 when AGE is equal to 0. As such, this intercept cannot be interpreted outside the context of this model since a 0-year-old patient is out of the range of the data. Furthermore, having a 0 AGE is also not a feasible age to include in this analysis of adult ICU experience. The coefficient for the AGE variable is 0.026. This can be interpreted as for every one year increase in the patient's age then the predicted odds of dying (i.e. STA equal to 1) increases by 0.026.

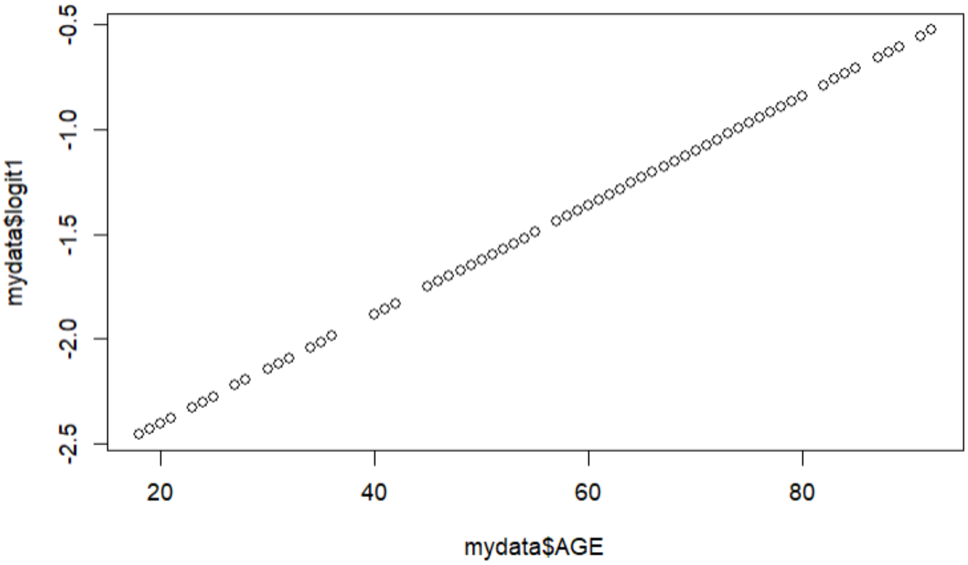
e. Below are the hypothesis test results for the likelihood-ratio test of the fitted logistic regression model using AGE.

   Null Hypothesis →H_0: $\beta_1 = 0$
   Alternative Hypothesis → H_A: $\beta_1 \neq 0$

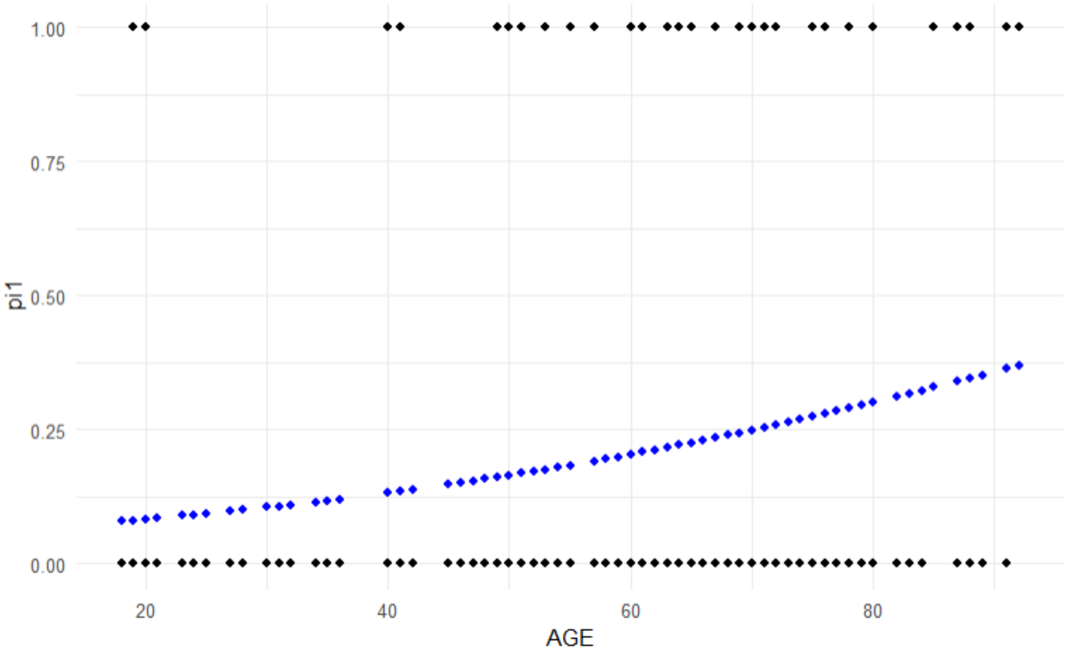
   For the likelihood-ratio test, I calculated a p-value of overall Chi-Square statistic is 0.012. Since this p-value is less than .05 significance level, we can reject the null hypothesis and conclude that there is a statistically significant relationship between age and ICU patient survival outcome.

f. The AIC of the model 195.55 and the BIC is 202.101. The value of the deviance for the fitted model is 191.55 on 193 degrees of freedom.

g. Using the fitted model, I then predicted logit values for each record in the dataset. Below is a scatterplot of the predicted logits(Y) by AGE (X). We can see the points forming a diagonal line meaning that as the age of the patient increases the odds of dying increases in a linear pattern. This is somewhat expected as the majority of deaths occurred in the older patients.



h. Using the predict logits, I converted these values to probabilities of dying (i.e. probability of having STA = 1). The below scatter plot shows the predicted probabilities by age in blue and the raw data for the actual STA outcome in black. We can see that the blue points do not display the typical 'S' shaped logistic curve. Additionally, we can see that none of the predicted probabilities are over 0.5 meaning that all the predictions would result in a 0 for STA with no 1's, if using a 0.5 threshold.



i. Using my own age (34), the model predicts that the probability that STA will be a 1 is 0.115 which in survival rate terms would be 0.885 survival rate. This is consistent with the scatter plot above. Additionally, this prediction seems reasonable given, what I observed in Tasks 1 and 2. The probability of survival for men from task 2 is .802 and given that I am towards the younger end of the range it makes sense that my survival rate is slightly higher than the mean survival rate for all men.

As noted previously, the model does not show a typical 'S' shape and all of the predicted probabilities are less than 0.5. As such, I do not believe we have the correct model yet. Ideally, we have a model that has prediction probabilities that span between the full range 0 and 1. By doing this, we can use an ifelse statement to convert the predicted probabilities to the dichotomous variables used in the STA outcome variable using 0.5 as the cutoff threshold. This issue could be caused by an imbalance in the STA outcome variable. Additionally, when creating the AGE_CAT variable I noticed that none of the categories have a probability more than 0.5. So even though the older patients have a higher rate of death compared to the younger patients, the majority of them still lived.

5. The next steps for this analysis would be to improve model fit. The current model that uses the continuous AGE variable does not produce prediction probabilities high enough to correctly classify if an ICU patient survives. Therefore, I would look into using the AGE_CAT variable instead to see if that improves the model. Next, I would then look to add other predictor variables to the model to see if these other variables improve the model's fit. In task 3, when looking at the proportional differences of outcomes for the two different admission types (TYP) it was observed that the elective odds of survival were 9.27 times greater than the odds of survival for the emergency admission ICU patients, so this would be a variable I would consider adding to the model to possibly improve the STA outcome predictions. Lastly, I would also recommend to resample the data if possible to see we can obtain a more balanced dataset in regards to the STA variable.

Conclusion

Overall, I got a good grasp of the basics of logistic regression models. I don't have a lot of experience with these types of models or dealing with converting probabilities to odds or vice versa, so it was a very good learning experience in working with this type of data and modeling technique. However, the model I created using the continuous AGE variable was a bit of a head-scratcher to me. I'm not sure if the poor fit was due to something I did incorrectly or if the poor fit was done by design to show that even though a variable by appear to be a "good" predictor that it can still produce a poorly fitting model. Based on the scatterplot of AGE vs STA, I would have thought the patient's age would be a good discriminator. But after creating the model using AGE and looking at the predicted probabilities and subsequent scatterplot, it appears that age alone is not a good enough predictor to determine the probability of an ICU patient's survival outcome. None of the predicted probabilities are over 0.5 which results in the plot of predicted probabilities not having a 'S' shaped curve. At first, I thought my model might have been wrong, but after further analysis the results make sense as even the oldest age category has a survival probability of 0.545 meaning that over half of them survive, although being the group with the lowest survival probability. So, it is plausible that a model that only looks at age would predict all of the observations as having under 0.5 probability for not surviving (i.e. STA = 1). However, there are several other variables in the dataset that could be added to the model as potential predictor variables, so that would be something I would like to explore in future work of this model to try and see if I can get the predicted probabilities to go above 0.5 and also to resemble more of an 'S' shaped curve.