Reed Ballesteros
MSDS-410-DL, Summer 2022
Dr. Mickelson
8/7/2022

Modeling Assignment 7: Logistic Regression Basics

Tasks

Please complete the tasks listed below and be sure to number your responses relative to the task number.

1. **Familiarize yourself with the codes for each of the variables. The response variable (Y) for this analysis will be the Status variable (STA). Conduct a basic exploratory data analysis to familiarize yourself with the data and the potential predictive relationships here.**

```
> nrow(mydata)
[1] 200
> summary(mydata)
       ID              STA              AGE              SEX              RACE             SER        
 Min.   :  4.0   Min.   :0.0   Min.   :16.00   Min.   :0.00   Min.   :1.000   Min.   :0.000  
 1st Qu.:210.2   1st Qu.:0.0   1st Qu.:46.75   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.000  
 Median :412.5   Median :0.0   Median :63.00   Median :0.00   Median :1.000   Median :1.000  
 Mean   :444.8   Mean   :0.2   Mean   :57.55   Mean   :0.38   Mean   :1.175   Mean   :0.535  
 3rd Qu.:671.8   3rd Qu.:0.0   3rd Qu.:72.00   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :929.0   Max.   :1.0   Max.   :92.00   Max.   :1.00   Max.   :3.000   Max.   :1.000  
      CAN              CRN              INF              CPR              SYS              HRA        
 Min.   :0.0   Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   : 36.0   Min.   : 39.00  
 1st Qu.:0.0   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:110.0   1st Qu.: 80.00  
 Median :0.0   Median :0.000   Median :0.00   Median :0.000   Median :130.0   Median : 96.00  
 Mean   :0.1   Mean   :0.095   Mean   :0.42   Mean   :0.065   Mean   :132.3   Mean   : 98.92  
 3rd Qu.:0.0   3rd Qu.:0.000   3rd Qu.:1.00   3rd Qu.:0.000   3rd Qu.:150.0   3rd Qu.:118.25  
 Max.   :1.0   Max.   :1.000   Max.   :1.00   Max.   :1.000   Max.   :256.0   Max.   :192.00  
      PRE              TYP              FRA              PO2              PH              PCO        
 Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   :0.0  
 1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.0  
 Median :0.00   Median :1.000   Median :0.000   Median :0.00   Median :0.000   Median :0.0  
 Mean   :0.15   Mean   :0.735   Mean   :0.075   Mean   :0.08   Mean   :0.065   Mean   :0.1  
 3rd Qu.:0.00   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.00   3rd Qu.:0.000   3rd Qu.:0.0  
 Max.   :1.00   Max.   :1.000   Max.   :1.000   Max.   :1.00   Max.   :1.000   Max.   :1.0  
      BIC              CRE              LOC        
 Min.   :0.000   Min.   :0.00   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000  
 Median :0.000   Median :0.00   Median :0.000  
 Mean   :0.075   Mean   :0.05   Mean   :0.125  
 3rd Qu.:0.000   3rd Qu.:0.00   3rd Qu.:0.000  
 Max.   :1.000   Max.   :1.00   Max.   :2.000  
> sum(is.na(mydata))
[1] 0
```

The dataset contains 200 observations and contains no null values. ID aside, there are three continuous variables age (AGE), systolic blood pressure (SYS), and heart rate (HRA). Two variables are nominal categorical variables, race (RACE) and level of consciousness (LOC). The remaining variables are binary.

What is the population of interest for this problem?
The population is the complete dataset of 200 observations.

Do we need dropdown conditions of any kind?
No dropdown conditions need to be made on the ICU dataset at this time.

2. **Obtain a 2x2 contingency table that relates gender (SEX) to Status (STA).**

```
# Sex           0=Male, 1=Female  SEX
# Vital Status   0=Lived, 1=Died  STA

xtabs(~ SEX + STA, data=mydata)
#     STA
# SEX   0   1  tot
#   0 100  24  124
#   1  60  16   76
# tot 160  40  200
```

Determine the odds and the probabilities of survival among males and females.
Probabilities:

```
# P(Lived|Male)
round(100/124, digits=3)
# 0.806

# P(Died|Male)
round(24/124, digits=3)
# 0.194

# P(Lived|Female)
round(60/76, digits=3)
# 0.789

# P(Died|Female)
round(16/76, digits=3)
# 0.211
```

Odds:

```
# Odds of lived (Male):
# = P(Lived|Male)/(1-P(Lived|Male))
round((100/124)/(1-(100/124)), digits=3)
round(100/24, digits=3)
# 4.167

# Odds of lived (Female):
# = P(Lived|Female)/(1-P(Lived|Female))
round((60/76)/(1-(60/76)), digits=3)
round(60/16, digits=3)
# 3.75
```

Then compute the odds ratio of survival that compares males to females.

```
# Odds ratio - Males vs. Females
odds_lived_male <- (100/124)/(1-(100/124))
odds_lived_female <- (60/76)/(1-(60/76))
odds_ratio <- round(odds_lived_male/odds_lived_female, digits=3)
odds_ratio
# 1.111
```

Does anything seem interesting here?

With an odds ratio of Males/Female of 1.11, both sexes pretty much have the same chance of survival while at the ICU, but Males have a very slight edge.

3. **Obtain a 2x2 contingency table that relates Type of Admission (TYP) to Status (STA).**

```
# Type of Admission 0 = Elective, 1 = Emergency  TYP
# Vital Status                    0=Lived, 1=Died  STA

xtabs(~ TYP + STA, data=mydata)
#     STA
# TYP   0   1  tot
#   0  51   2   53
#   1 109  38  147
# tot 160  40  200
```

Again, determine the odds and probabilities of survival among the different Types of Admission.

```
# P(Lived|Elective)
round(51/53, digits=3)
# 0.962

# P(Died|Elective)
round(2/53, digits=3)
# 0.038


# P(Lived|Emergency)
round(109/147, digits=3)
# 0.741

# P(Died|Emergency)
round(38/147, digits=3)
# 0.259
```

Then compute and interpret the odds ratio of survival that compares them.

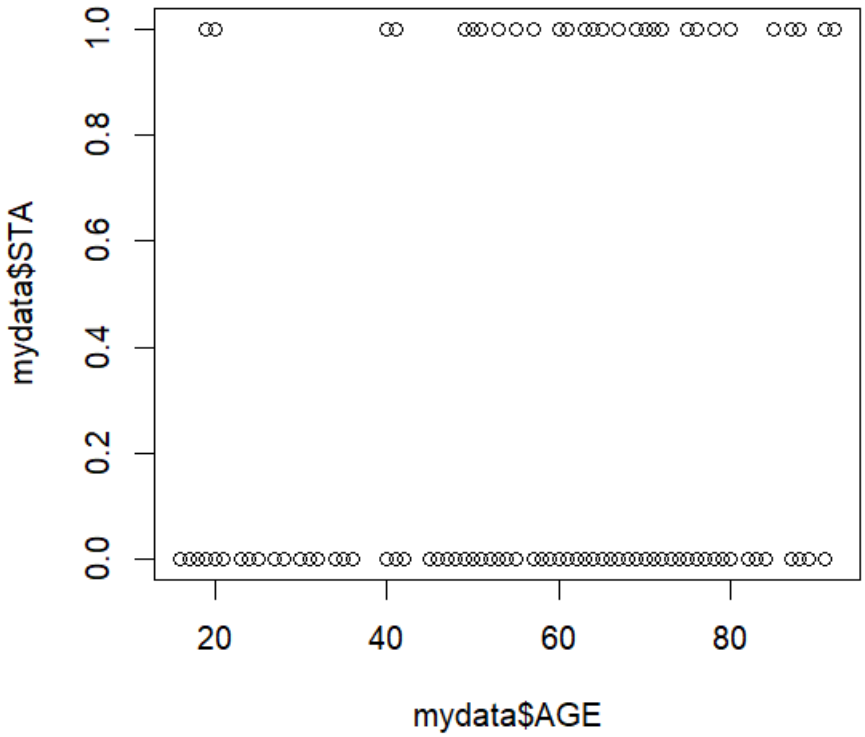
```
# Odds ratio - Elective vs. Emergency
odds_lived_el <- (51/53)/(1-(51/53))
odds_lived_em <- (109/147)/(1-(109/147))
odds_ratio_ee <- round(odds_lived_el/odds_lived_em, digits=3)
odds_ratio_ee
# 8.89
```

With an odds ratio of 8.89, people that have elective surgery have an almost nine times greater chance to survive the ICU than those who have emergency surgery.

4. **Suppose the patient's AGE is considered to be a key determinant of the patient's survival. With this information, complete the following:**
    a. **Write the equation for the logistic regression model of STA (Y) using AGE (X).**
    With pi = P(STA=1|AGE=x), B0 = intercept, B1 = coefficient of AGE:
    $$pi = \frac{e^{B0+B1*x}}{1 + e^{B0+B1*x}}$$

    Write the equation for the logit transformation of this logistic regression model.
    $$\log\left(\frac{pi}{1-pi}\right) = B0 + B1 * x$$

b. Make a scatterplot of STA (Y) by AGE(Y).



Does AGE seem to be a good discriminator between levels of STA?
At this point, AGE seems to be a decent discriminator between levels of STA.

c. Construct a new categorical variable by discretizing AGE into the following intervals:

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

```
> AGECAT <- ifelse(mydata$AGE<=24,1,
+           ifelse(mydata$AGE<=34,2,
+           ifelse(mydata$AGE<=44,3,
+           ifelse(mydata$AGE<=54,4,
+           ifelse(mydata$AGE<=64,5,
+           ifelse(mydata$AGE<=74,6,
+           ifelse(mydata$AGE<=84,7,
+           ifelse(mydata$AGE<=94,8,9)))))))
> mydata$AGECAT <- AGECAT
> table(mydata$AGECAT)

 1  2  3  4  5  6  7  8
26  8 11 25 39 50 30 11
```

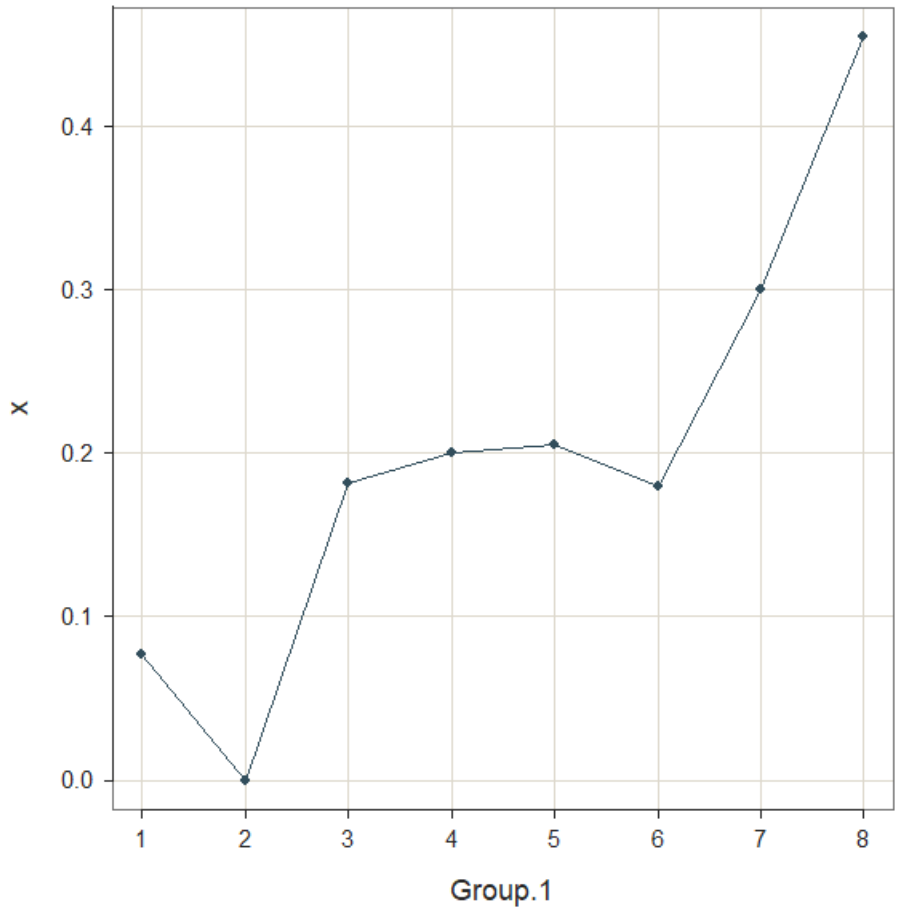
Using this categorical variable, compute the STA mean (i.e. proportion) over subjects in the age interval.

```
> agg_agecat_sta <- data.frame(aggregate(mydata$STA, list(mydata$AGECAT), FUN=mean))
> # Group.1 = AGECAT, x= mean by AGECAT
> agg_agecat_sta
  Group.1          x
1       1 0.07692308
2       2 0.00000000
3       3 0.18181818
4       4 0.20000000
5       5 0.20512821
6       6 0.18000000
7       7 0.30000000
8       8 0.45454545
```

Plot these means versus the categorical variable.

Group.1 = AGECAT, x = mean by AGECAT

```
> Plot(Group.1, x, data=agg_agecat_sta)
>>> Suggestions
Plot(Group.1, x, enhance=TRUE)  # many options
Plot(Group.1, x, color="red")  # exterior edge color of points
Plot(Group.1, x, fit="lm", fit_se=c(.90,.99))  # fit line, stnd errors
Plot(Group.1, x, out_cut=.10)  # label top 10% from center as outliers

>>> Pearson's product-moment correlation

Number of paired values with neither missing, n = 8
Sample Correlation of Group.1 and x: r = 0.884

Hypothesis Test of 0 Correlation:  t = 4.636,  df = 6,  p-value = 0.004
95% Confidence Interval for Correlation:  0.476 to 0.979
```



d. Fit a logistic regression model to predict STA using the original continuous AGE variable. Report and interpret the coefficients for the model.

```
> model1 <-glm(STA ~ AGE, data=mydata, family=binomial)
> model1

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

Coefficients:
(Intercept)          AGE
   -3.05851      0.02754

Degrees of Freedom: 199 Total (i.e. Null);  198 Residual
Null Deviance:      200.2
Residual Deviance: 192.3        AIC: 196.3
```

Converting the AGE coefficient of 0.02754 as a percentage:

```
> model1AgePercent <- exp(0.02754) - 1
> round(model1AgePercent * 100, digits=3)
[1] 2.792
```

For each additional year in age, the odds of death increases by 2.792%.

e. Report and interpret all hypothesis test results. What do you conclude?

H0: model1 is adequate relative to the null model (model1 is better than the null model).

HA: model1 is not adequate relative to the null model.

```
> anova(model1, test="LR")
Analysis of Deviance Table

Model: binomial, link: logit

Response: STA

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                   199     200.16
AGE   1   7.8546       198     192.31 0.005069 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

With the analysis of deviance table above, the deviance difference between the null and residual deviance is 7.8546, which is also the Chi-square statistic. With a confidence level of 95%, the Chi-squared p-value 0.005069 is less than the critical p-value of 0.05 such that we do not reject the null hypothesis and that logistic regression model1 is statistically more significant than the null model, thus a better model.

f. Report the AIC and BIC values.

```
> AIC(model1)
[1] 196.3064
> BIC(model1)
[1] 202.903
```

What is the value of the deviance for the fitted model?

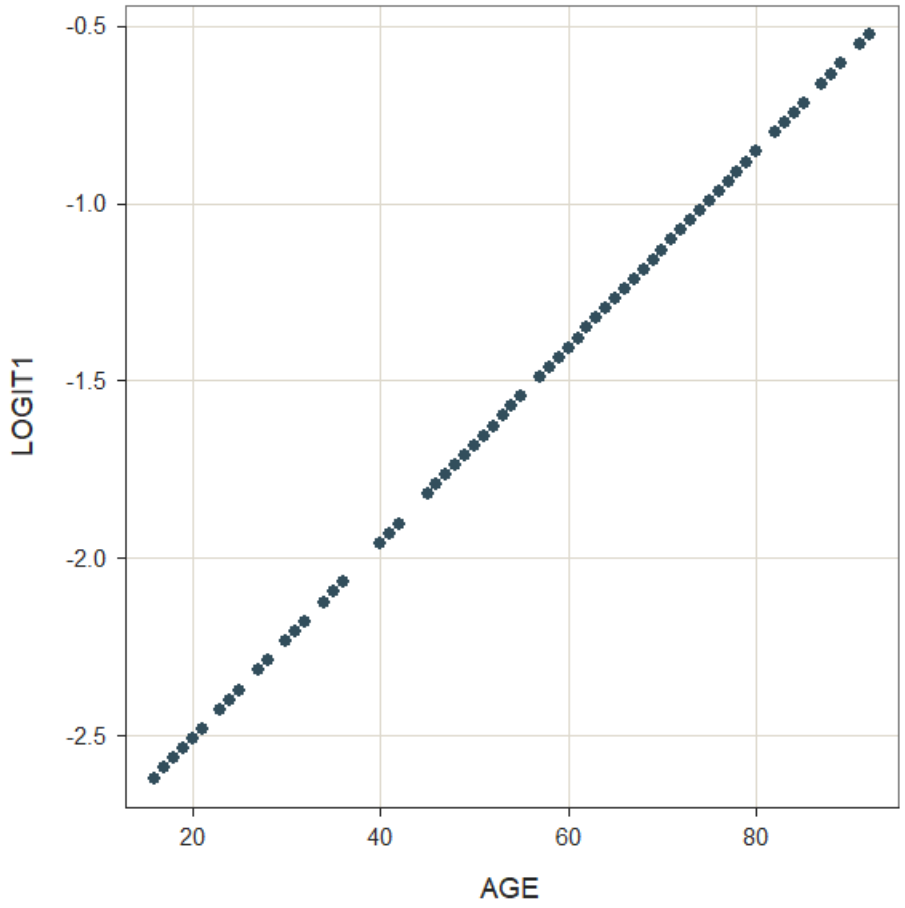
```
> summary(model1)

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

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 0.0000111 ***
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 null deviance is 200.16 and the residual deviance is 192.31

g. **Use the fitted model to predict logit values for each record in the dataset. Save the logits to your analysis file.**

```
logit1 <- -3.05851 + 0.02754*mydata$AGE
mydata$LOGIT1 <- logit1
```

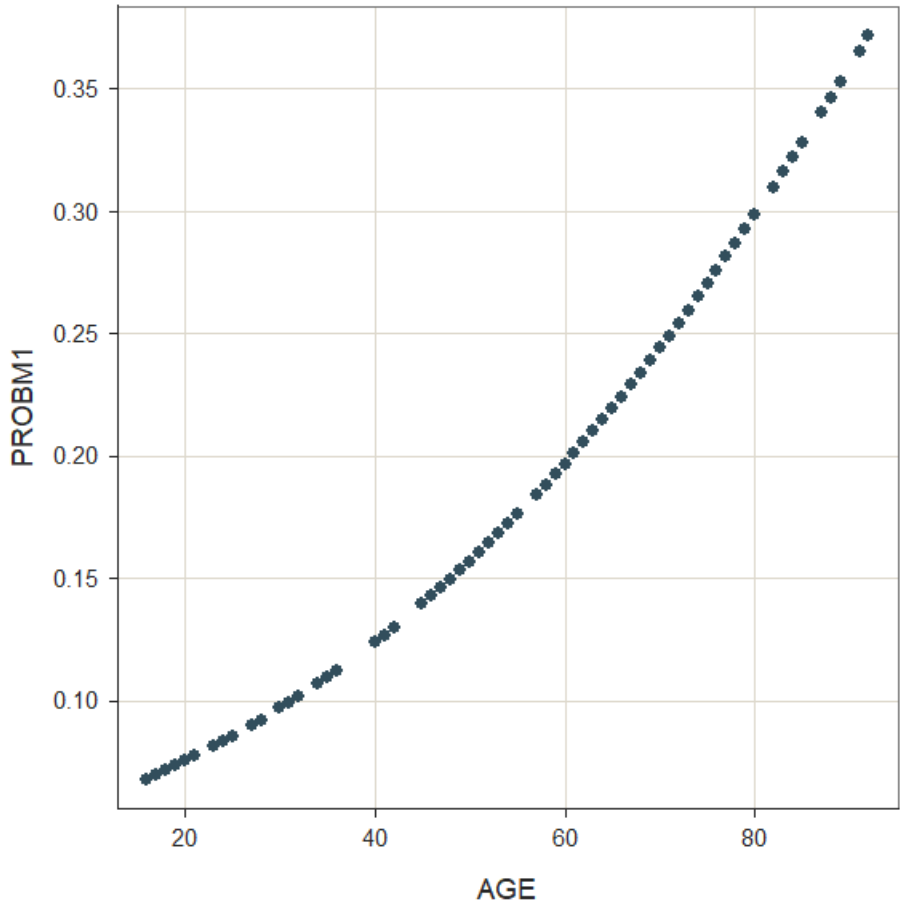
Then make a scatterplot of the predicted logits(Y) by AGE (X). Discuss the scatterplot.



h. **Write a line or two or three of R-code to compute the probabilities of survival (pi) from the logits. Save the predicted probabilities to your analysis file.**

```
oddsratio1 < -exp(logit1)
PROBM1 < -exp(logit1)/(1 + exp(logit1))
mydata$PROBM1 <- PROBM1
```

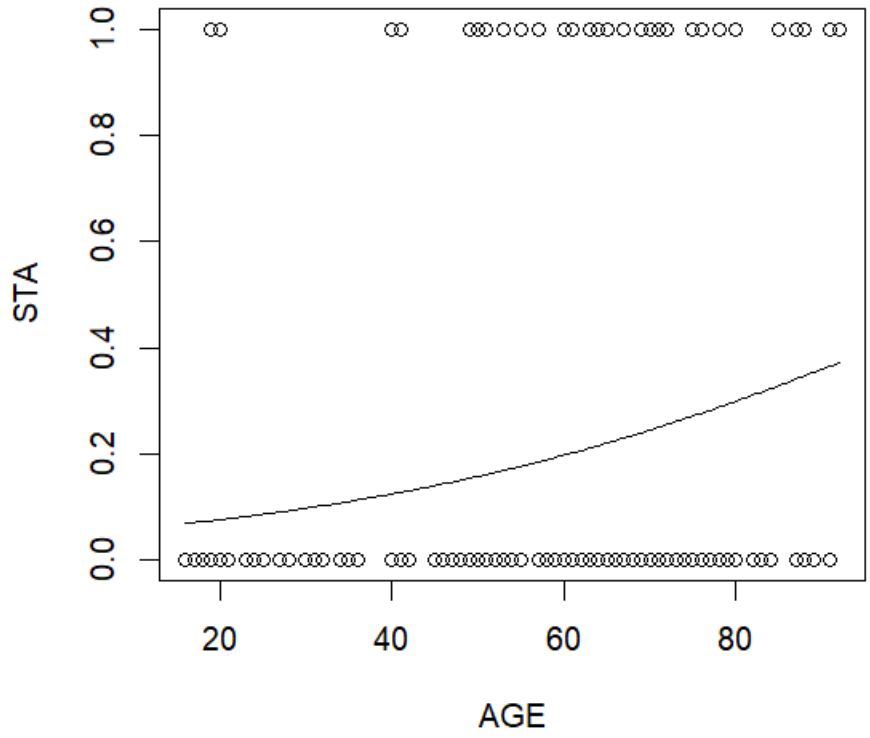
Then make a scatterplot of the predicted probabilities (Y) by AGE (X).



Do you see the typical 'S' shaped logistic curve?

We see a curve but not a complete S curve, which stops well short of the 0.5 threshold for the probabilities.

If possible, overlay the raw data of Y=STA on top of your predicted values of probability of Survival.



i. **Use the logistic model you developed to predict the probability of survival for someone your age (47).**

```
> myagedata <- data.frame(AGE=c(47))
> myagedata_yhat <- predict(model1, newdata=newdataoldguy, type="response")
> round(myagedata_yhat, digits=3)
    1
0.146
```

The calculated probability of survival for someone of age 47 is 0.146.

```
> OutcomeMyAge <- ifelse(myagedata_yhat > 0.5,1,0)
> OutcomeMyAge
1
0
```

Given the probability of survival and a threshold of 0.5, we get a STA outcome of 0, which means that someone my age (47) should survive the visit to the ICU.

Is this prediction consistent with what you see in the scatterplot above? Does this seem like a reasonable prediction given what you observed in Tasks 1 and 2?
The prediction is consistent with the scatterplot such that all ages are plotted underneath the threshold of 0.5, showing that anyone, regardless of age, can survive the visit to the ICU, despite the dataset showing otherwise.

Do we have the correct model yet?
We do not have a correct model yet as the plot of probabilities does not display a full S-curve, and as mentioned earlier, the model shows that anyone, regardless of age, can survive the visit to the ICU, despite the dataset showing otherwise.

5. **Given what you have learned from this modeling endeavor so far, what are the next steps for our analysis? What is your recommended plan for the next phase of modeling?**

I feel that 200 observations are not yet a large enough sample to create a logistic regression of STA based on AGE. Much more data needs to be collected. Another route can also venture into multivariate logistic regression incorporating other continuous variables such as blood pressure (SYS), heart rate (HRA), and other binary can categorical variables in the dataset. That being said, multivariate logistic regression is out of the scope of this module's lessons and look forward to later modules to understand the concept.