Kay Quiballo | MSDS 410 | 02.13.2023

Modeling Assignment 7: Logistic Regression Basics

Assignment Overview

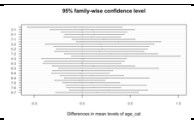
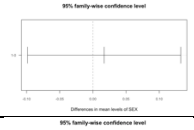
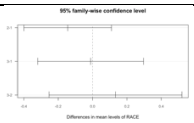
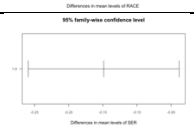
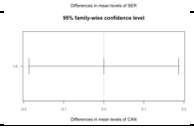
Construct logistic regression models to predict survival of patients through an adult ICU experience. The data for this assignment is the ICU data set: ICU.xlsx. It consists of a sample of 200 subjects randomly selected from a much larger study on the survival of patients following admission to an adult intensive care unit (ICU) in a major metropolitan city. The goal of this project was to develop models to predict the probability of survival to hospital discharge of the patients seen. This data is made available for free online by John Wiley & Sons, Inc.

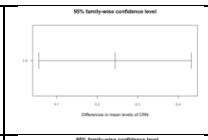
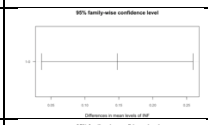
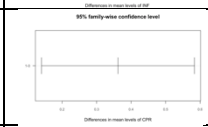
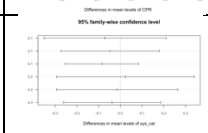
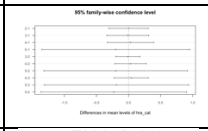
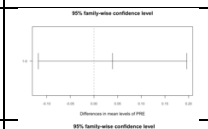
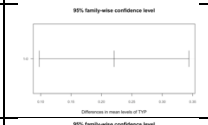
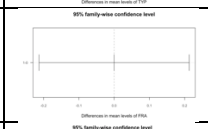
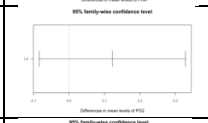
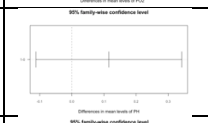
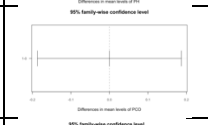
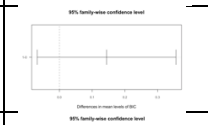
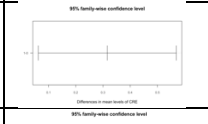
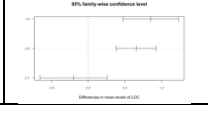
Tasks

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. What is the population of interest for this problem? Do we need dropdown conditions of any kind?

The population of interest is ICU patients in a major metropolitan city who have or have not survived after being discharged from the hospital. The specific year and location are unknown and will be noted in interpretation. There are no missing data points and no infeasible values. Dropdown conditions will be revisited if any outliers need to be removed.

Variable	**Proportion Table for Column %**	**ANOVA difference of means**	**Do Col # and ANOVA indicate difference in means?**
AGE (Age) *age_cat according to section 4c*	age_cat 2 3 4 5 6 7 liv 0.89 1.00 0.91 0.81 0.80 0.75 die 0.11 0.00 0.09 0.19 0.20 0.25		No
SEX (Sex)	0 1 liv 0.81 0.79 die 0.19 0.21		No
RACE (Race)	1 2 3 liv 0.79 0.93 0.80 die 0.21 0.07 0.20		No
SER (Service at ICU)	med surg liv 0.72 0.87 die **0.28** **0.13**		Yes, those who receive medical services are more likely to die than surgical services
CAN (Cancer part of present problem)	0 1 liv 0.8 0.8 die 0.2 0.2		No

CRN (History of Chronic Renal Failure)	No Yes liv 0.82 0.58 die **0.18 0.42**		Yes, those who have chronic renal failure are more likely to die than those without
INF (Infection Probable at ICU Admission)	No Yes liv 0.86 0.71 die **0.14 0.29**		Yes, those who have an infection are more likely to die than those without
CPR (CPR Prior to ICU Admission)	No Yes liv 0.82 0.46 die **0.18 0.54**		Yes, those who had CPR are more likely to die than those who did not
SYS (Systolic Blood Pressure at ICU Admission) CDC 4 ranges	sys_cat 1 2 3 4 liv 0.75 0.82 0.80 0.84 die 0.25 0.18 0.20 0.16		No
HRA (Heart Rate at ICU Admission) Break at 60, 100, 130, 170 Heart.org ranges	hra_cat 1 2 3 4 5 liv 0.80 0.80 0.81 0.77 1.00 die 0.20 0.20 0.19 0.23 0.00		No
PRE (Previous Admission to ICU within 6 months)	0 1 liv 0.81 0.77 die 0.19 0.23		No
TYP (Type of Admission)	Elec Emer liv 0.96 0.74 die **0.04 0.26**		Yes, those who came into the ICU as an Emergency are more likely to die than those who Elected to come in
FRA (Bone Fracture)	0 1 liv 0.8 0.8 die 0.2 0.2		No
PO2 (PO2 from Initial Blood Gases)	0 1 liv 0.81 0.69 die 0.19 0.31		No
PH (PH from Initial Blood Gases)	0 1 liv 0.81 0.69 die 0.19 0.31		No
PCO (PCO2 from Initial Blood Gases)	0 1 liv 0.8 0.8 die 0.2 0.2		No
BIC (Bicarbonate from Initial Blood Gases)	0 1 liv 0.81 0.67 die 0.19 0.33		No
CRE (Creatinine from Initial Blood Gases)	<=2 >2 liv 0.82 0.50 die **0.18 0.50**		Yes, those with creatine >2 were more likely to die than those with creatine <=2
LOC (Level of Consciousness at ICU Admission)	none stup coma liv 0.85 0.00 0.20 die **0.15** 1.00 0.80		Yes, those in a deep stupor or coma were more likely to die than those who were not

From the proportion tables and ANOVA tables above, here are the 7 predictors that we are most interested in for predicting vital status (STA): SER, CRN, INF, CPR, TYP, CRE, and LOC.

2. Obtain a 2x2 contingency table that relates gender (SEX) to Status (STA). Determine the odds and the probabilities of survival among males and females. Then compute the odds ratio of survival that compares males to females. Does anything seem interesting here?

STA vs SEX

Frequencies

	SEX=Male	SEX=Female
STA=Live	100	60
STA=Die	24	16

Total %

	SEX=Male	SEX=Female
STA=Live	0.50	0.30
STA=Die	0.12	0.08

Row %

	SEX=Male	SEX=Female
STA=Live	0.62	0.38
STA=Die	0.60	0.40

Column %

	SEX=Male	SEX=Female
STA=Live	0.81	0.79
STA=Die	0.19	0.21

Probability of mortality in Males = 24/(24+100) = 19.35%
Probability of mortality in Females = 16/(16+60) = 21.05%
Odds in exposed group (Male) = 24/100 = 0.24
Odds in not exposed group (Female) = 16/60 = 0.2666667
Odds ratio = 0.24 / 0.2666667 = 0.9

This hypothetical group of males has 0.9 times the odds of dying than females.
It is interesting to see that females are more likely to die.

3. Obtain a 2x2 contingency table that relates Type of Admission (TYP) to Status (STA). Again, determine the odds and probabilities of survival among the different Types of Admission. Then compute and interpret the odds ratio of survival that compares them.

STA vs TYP

Frequencies

	TYP=Elective	TYP=Emergency
STA=Live	51	109
STA=Die	2	38

Total %

	TYP=Elective	TYP=Emergency
STA=Live	0.26	0.54
STA=Die	0.01	0.19

Row %

	TYP=Elective	TYP=Emergency
STA=Live	0.32	0.68
STA=Die	0.05	0.95

Column %

	TYP=Elective	TYP=Emergency
STA=Live	0.96	0.74
STA=Die	0.04	0.26

Probability of mortality in Elective = 2/(2+51) = 3.77%
Probability of mortality in Emergency = 38/(38+109) = 25.85%
Odds in exposed group (Elective) = 2/51 = 0.03921569
Odds in not exposed group (Emergency) = 38/109 = 0.3486239
Odds ratio = 0.03921569 / 0.3486239 = 0.1124871

This hypothetical group of individuals who elected to go to the ICU has 0.1124871 times the odds of dying than those who came to the ICU as an emergency.

4. Suppose the patient's AGE is considered to be a key determinant of the patient's survival. With this information, complete the following:

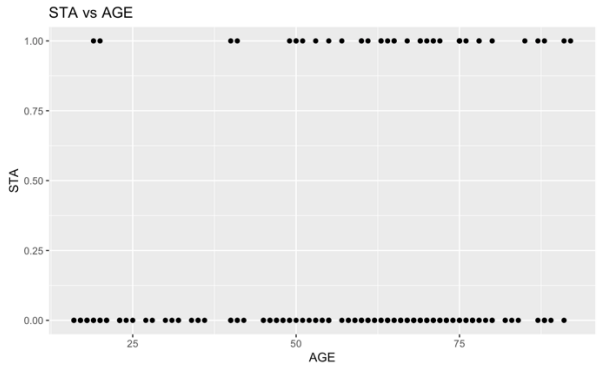
a. Write the equation for the logistic regression model of STA (Y) using AGE (X). Write the equation for the logit transformation of this logistic regression model.

$\pi(x) = E(y|x) = e(\beta 0+\beta 1x) / [1+e(\beta 0+\beta 1x)]$
where π is the expected value of X=STA given Y=AGE.

Odds ratio = $\pi(x) / [(1-\pi(x))]$

$\ln(\pi(x) / (1-\pi(x))) = \beta 0+\beta 1x$
Logit = -3.05851 + 0.02754*AGE

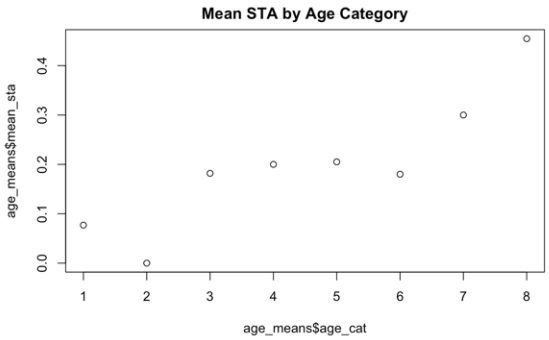
b. Make a scatterplot of STA (Y) by AGE(Y). Does Age seem to be a good discriminator between levels of STA?



There may be slight evidence to suggest that lower ages tend to live after a visit to the ICU. However, the distribution of AGE between STA=0 and STA=1 is difficult to differentiate and may indicate that AGE it is not the best discriminator (which is supported by section 1's ANOVA test).

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

Age_cat	Mean(STA)
[15,24]	0.07692308
[25,34]	0.00000000
[35,44]	0.18181818
[45,54]	0.20000000
[55,64]	0.20512821
[65,74]	0.18000000
[75,84]	0.30000000
[95,]	0.45454545



Younger age groups like [15,24] and [25,34] have a much higher proportion of survival compared to older age groups such as [75,84] and [95,] which have a relatively higher proportion of mortality.

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

Logit = -3.05851 + 0.02754*AGE
For each additional year of age, we estimate the odds of a patient dying after going to the ICU increases by exp(b1) – 1 = exp(0.02754) – 1 = 1.0279 – 1 = 0.0279 or **2.79%**.

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

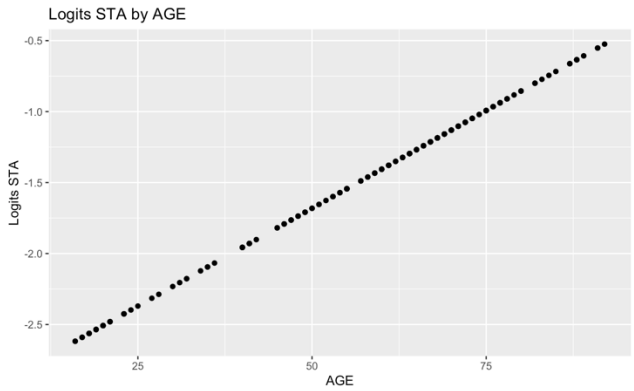
Null Hyp	Alt Hyp	Z and P value	Conclusion
β0 = 0	β0 != 0	-4.394 0.0000111	Fail to reject H0. Conclude that β0 != 0
β1 = 0	β1 != 0	2.607 0.00913	Fail to reject H0. Conclude that β1 != 0

With P values < 0.05, we conclude that both the intercept and coefficient for AGE are significant to the logistic regression and not equal to 0.

f. Report the AIC and BIC values. What is the value of the deviance for the fitted model?

AIC: 196.3064
BIC: 202.903
Null deviance: 200.16 on 199 degrees of freedom
Residual deviance: 192.31 on 198 degrees of freedom

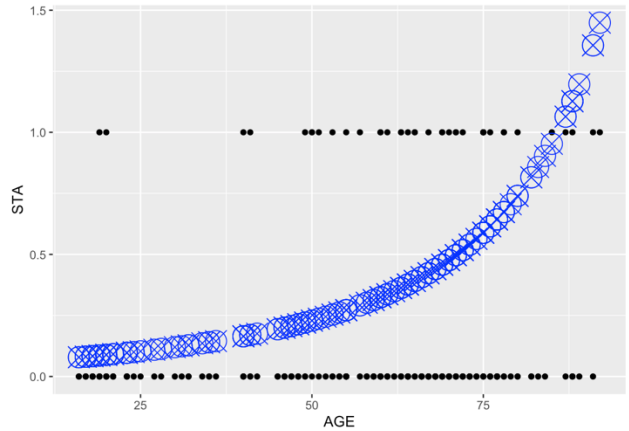
g. Use the fitted model to predict logit values for each record in the dataset. Save the logits to your analysis file. Then make a scatterplot of the predicted logits(Y) by AGE (X). Discuss the scatterplot.



The logistic model predicts a perfectly linear relationship between AGE and Logits of STA. As age increases, we also see an increase in Logits of STA.

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. Then make a scatterplot of the predicted probabilities (Y) by AGE (X). Do you see the typical 'S' shaped logistic curve? If possible, overlay the raw data of Y=STA on top of your predicted values of probability of Survival.

STA vs Age (black) & Predicted Probability of Survival/Mortality vs AGE (blue)



The blue points here indicate the probability of survival/mortality pi vs AGE, and the black points here indicate of STA vs AGE. As we can see, the pi values take on the beginning of an S shaped logistic curve before the S curve would plateaus out.

i. Use the logistic model you developed to predict the probability of survival for someone your age. 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? Do we have the correct model yet?

The predicted pi value (probability of survival/mortality) for AGE=25 is 0.1031194. This means that a 25 year-old individual who goes to the ICU has a predicted value of dying = 10.31%

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?

Based off our previous assignments, I there are several next steps for the analysis that I believe we should take.

Models need tweaking in order to consider better fitting models.

1) This goes back to section 1 of the EDA where we considered the relationship of discrete predictors with STA and tested with 95% pairwise ANOVA tests whether there was a significant difference in means. This EDA informs us of which predictors to test out in models. In this specific assignment, we identified 7 predictors (SER, CRN, INF, CPR, TYP, CRE, and LOC) that have statistically differences in means in STA across levels.
2) After choosing an optimal model, we must also check if any points are influential and highly deviate from the dominant prediction pattern of our model. Note that we don't want to introduce any modeler bias by eliminating a significant number of valuable data. Further analysis must note the elimination of outliers in the dropdown conditions and interpretations of the model.

Once we have tested different combinations of predictors and eliminated any outliers, we have different ways we can validate the model.

1) With testing and training datasets, we can check the in-sample and out-of-sample goodness of fit metrics. Doing so will help us determine which models perform better for predicting logit STA based on different combinations of predictors. We can also determine whether a model is overfitting if its out-of-sample accuracy is relatively lower than its in-sample accuracy or if the model is robust and can be generalized to our population of interest.
2) The final steps will be confirming that the model we selected does not violate any assumptions that go with a logistic regression such as independence of errors, linearity of logits, no multicollinearity, and no outliers. A model cannot be used for interpretation or predicting values unless it satisfies these assumptions.
3) After we have tweaked the and validated the model, we can interpret any coefficients and their meanings. In this example how the predictors affect logit STA and what that means for ICU patient survival. We can also revisit any of the tweaking and validating steps in the analysis if the interpretation of coefficients does not logically make sense or cannot be justified.