

Modeling Assignment 7: Logistic Regression Basics

Assignment Overview

In this assignment we will be calculating the various summative statistics that are associated with logistic regression, as well as fitting logistic regression models and interpreting the results. The objective of this assignment is to begin constructing logistic regression models to predict survival of patients through an adult ICU experience. Students are expected to show all work in their computations. A good practice is to write down the generic formula for any computation and then fill in the values need for the computation from the problem statement. Throughout this assignment keep all decimals to three places, i.e. X.xxx. Students are expected to use correct notation and terminology, and to be clear, complete and concise with all interpretations of results.

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. You can find the link in the Canvas assignment. The actual observed values of this data have been modified to protect subject confidentiality.

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?
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?
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.
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.

- b. Make a scatterplot of STA (Y) by AGE(Y). Does Age seem to be a good 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

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

- d. Fit a logistic regression model to predict STA using the original continuous AGE variable. Report and interpret the coefficients for the model.
- e. Report and interpret all hypothesis test results. What do you conclude?
- f. Report the AIC and BIC values. What is the value of the deviance for the fitted model?
- 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.
- h. Write a line or two or three of R-code to compute the probabilities of survival (π) 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.
- 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?

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?

Assignment Document

Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document you submit to be graded MUST be submitted in pdf format. Please use the naming convention: ModelAssign7_YourLastName.pdf.