Unit 12 Pre-Live

# Introduction

For this pre-live assignment, we are going to be focusing on two main things, interpretation and verifying that logistic regression is a more general technique that has the previous methods as special cases. The data set we will explore is a coronary artery disease data set and the goal is to determine what risk factors would contribute to the likelihood of an individual having the disease or not. The data is below in the SAS code.

Ca coronary artery disease (1 = Yes)

Sex (1 = Male, 0 = Female)

ECG Categorical level for an echocardiogram result (0 = low, 1 = medium, 2 = high)

Age Continuous, in years

# Questions

1. Use the proc freq and proc means procedures to explore if any relationships exist between the response, disease status, and the potential risk factors.

A picture containing table

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Sex by Ca | ECG by CA | ECG by CA (Male) | ECG by CA (Female) |
| A picture containing table  Description automatically generated | A picture containing table  Description automatically generated | A picture containing calendar  Description automatically generated | Table  Description automatically generated |

Graphical user interface, text, application

Description automatically generated

**It appears that all of the variables may contribute to coronary artery disease, and there may also be an interaction between ECG and Sex.**

1. Fit a simple logistic regression using all 3 predictors and verify that the overall model is significant. We will discuss in live session that residual model checking for logistic regression is not very informative due to the categorical nature. One way to do this is with the Hosmer-Lemeshow lack of fit test. Obtain the statistic and p-value from the output. Does the result tell us that our model fit is reasonable or not?

**With a p-value of 0.5806, the Hosmer-Lemeshow test does not provide evidence that the predicted probabilities deviate from the observed probabilities.**

Graphical user interface, application

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

1. Provide an interpretation of the regression coefficient for Age. What does the value represent in terms of the odds of having the disease?

Table

Description automatically generated

**Holding all other values equal, each additional year in age corresponds to a 1.1 time increase in the odds of having coronary artery disease.**

1. What is the odds ratio of a person with an Age = 55 having coronary artery disease with respect to someone with who is 45 years of age given they are both males and have the same ECG result (See 20.2.2 in text)? Pay attention to the output as you may need to do some “flipping” (reciprocal 1/OR).

**The odds of a male age 55 having coronary artery disease are 2.6 times more likely than a male age 45.**

1. Consider a new patient comes into a doctor’s office and has measurement ECG = 2, Age = 50, Sex = Male. What is the probability that this person has the disease? (Here we are trying to predict if the new person has the disease or not.)

1. Run a 2x2 analysis with CA and Sex and then run a logistic regression model just using Sex as an explanatory variable. Determine the odds ratio (disease:no disease) males vs females for the two analysis runs and verify that they are very close. This shows that logistic regression has the simple 2x2 table analysis as a special case when you have one categorical explanatory variable with just two levels.

**I don’t think this is what you’re looking for, but it appears that the odds for a female to have the disease are approximately 75% less than the odds of a male.**

Table

Description automatically generated

1. **Optional**: It is not in the videos, but feel free to do a little google on the topic of ROC curves which we will discuss over the next couple of weeks. ROC stands for “Receiver Operating Characteristic”.

# SAS Code

**data** coronary;

input sex ecg age ca @@ ;

cards;

0 0 28 0 1 0 42 1 0 1 46 0 1 1 45 0

0 0 34 0 1 0 44 1 0 1 48 1 1 1 45 1

0 0 38 0 1 0 45 0 0 1 49 0 1 1 45 1

0 0 41 1 1 0 46 0 0 1 49 0 1 1 46 1

0 0 44 0 1 0 48 0 0 1 52 0 1 1 48 1

0 0 45 1 1 0 50 0 0 1 53 1 1 1 57 1

0 0 46 0 1 0 52 1 0 1 54 1 1 1 57 1

0 0 47 0 1 0 52 1 0 1 55 0 1 1 59 1

0 0 50 0 1 0 54 0 0 1 57 1 1 1 60 1

0 0 51 0 1 0 55 0 0 2 46 1 1 1 63 1

0 0 51 0 1 0 59 1 0 2 48 0 1 2 35 0

0 0 53 0 1 0 59 1 0 2 57 1 1 2 37 1

0 0 55 1 1 1 32 0 0 2 60 1 1 2 43 1

0 0 59 0 1 1 37 0 1 0 30 0 1 2 47 1

0 0 60 1 1 1 38 1 1 0 34 0 1 2 48 1

0 1 32 1 1 1 38 1 1 0 36 1 1 2 49 0

0 1 33 0 1 1 42 1 1 0 38 1 1 2 58 1

0 1 35 0 1 1 43 0 1 0 39 0 1 2 59 1

0 1 39 0 1 1 43 1 1 0 42 0 1 2 60 1

0 1 40 0 1 1 44 1

;

**run**;**quit**;

\*Explore;

**proc** **freq** data=coronary;

tables sex\*ca ecg\*ca sex\*ecg\*ca / chisq relrisk;

**run**;**quit**;

**proc** **means** data=coronary;

class ca sex ecg;

types ca ca\*sex ca\*ecg;

var age;

**run**;

\*Simple proc logistic call to play around with;

**proc** **logistic** data=coronary ;

class sex ecg / param=ref;

model ca(event='1')= sex/ scale=none aggregate influence lackfit;

**run**;