HW 04: Logistic Regression

Dav King

2022-11-20

Set up

> **Note**
>
> No R packages or data needed for this assignment.
> Select this page for Workflow & formatting".

Exercises

Exercise 1

The odds of having an incident neurocognitive disorder (NCD) are expected to be 2.564 (1/.39) times higher for (initially) cognitively intact community-living Chinese elderly who consistently do not consume tea than those who do consistently consume tea, holding age, pre-existing health conditions, diet, and behavioral factors constant.

The odds of having an NCD are expected to be 3.125 (1/.32) times higher for (initially) cognitively intact community-intact Chinese elderly females who consistently do not consume tea than those who do consistently consume tea, holding age, pre-existing health conditions, diet, and behavioral factors constant.

Exercise 2

This statement is not supported by the results of the study. Their claim of 86% comes from subtracting the odds ratio of NCD risk among APOE e4 carriers who regularly consume tea from 1. Since an odds ratio can be greater than 1, this is not an appropriate way to interpret the data - thus, the study does not support this claim.

Exercise 3

The odds that a randomly selected person who has been unemployed over a year is 55 and up are 0.282.

The odds that a randomly selected person who has been unemployed over a year is not 25 to 54 years old are 0.613.

Exercise 4

In order to raise an individual's odds of becoming long-term unemployed by 35 percent, they odds must be multiplied by 1.35. Since logistic regression models give coefficients in terms of log-odds, we must take the log(1.35), which gives us a coefficient for unemployment rate of roughly 0.3.

Exercise 5

Sensitivity: among people who were rearrested, the model correctly classified 86% of them as high risk.

Specificity: among people who were not rearrested, the model correctly classified 24% of them as low risk.

Positive predictive power: among people classified as high risk by the model, 57% of them were actually rearrested.

Negative predictive power: among people classified as low risk by the model, 60% of them were not rearrested.

Exercise 6

The false positive rate is 0.76. This means that 76% of people who the model classified as high risk did not end up getting rearrested.

Exercise 7

Based on the given AUC value, this algorithm is not a good fit at all for the population examined. It performs barely better than a coin flip at predicting whether defendants are high or low risk, and this lack of predictive validity leads to a lot of harmful misclassifications.

Exercise 8

$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -0.81 - 2.64 \times to\_multiple1 + 1.63 \times winneryes - 1.59 \times format1 - 3.05 \times re\_subj1$

The predicted probability of such an observation being spam is 0.316 (-0.81 - 2.64(0) + 1.63(1) - 1.59(1) -3.05(0) yields Odds = -0.77, exp(-0.77) / 1 + exp(-0.77) $\approx$ 0.316).

Without having the data, it is somewhat difficult to find an exactly optimized value here. However, personally, I would set a very high threshold for the probability a message is spam before it is put in a spam folder - somewhere around .85 or .9. There are two main tradeoffs to consider here. If we optimize for sensitivity, we will correctly classify as many as possible of our spam emails as spam, but a number of emails that are not spam will also be classified as spam and thus not seen by the user. On the other hand, if we optimize for specificity, we will correctly classify as many as possible of our non-spam emails as not spam, but a number of emails that are spam will also not be caught by the filter. As a data scientist, I would prioritize minimizing false positives, so that we ensure customers actually see the important emails that are sent to them - thus, I would optimize the model for specificity by choosing a high cutoff value.