Logistic Regression Applications

YOUR NAME

11 December, 2020

Exercises

1. Possum classification

Let's investigate the possum data set again. This time we want to model a binary outcome variable. As a reminder, the common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum. We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia.

We use logistic regression to differentiate between possums in these two regions. The outcome variable, called pop, takes value Vic when a possum is from Victoria and other when it is from New South Wales or Queensland. We consider five predictors: sex, head_l, skull_w, total_l, and tail_l.

a. Explore the data by making histograms or boxplots of the quantitative variables, and bar charts of the discrete variables.
   Are there any outliers that are likely to have a very large influence on the logistic regression model?
b. Build a logistic regression model with all the variable. Report a summary of the model.

c. Using the p-values decide if you want to remove a variable(S) and if so build that model.
d. For any variable you decide to remove, build a 95% confidence interval for the parameter.
e. Explain why the remaining parameter estimates change between the two models.
f. Write out the form of the model. Also identify which of the following variables are positively associated (when controlling for other variables) with a possum being from Victoria: head_l, skull_w, total_l, and tail_l.
g. Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australia. However, the sign does indicate that the possum is male, its skull is about 63 mm wide, its tail is 37 cm long, and its total length is 83 cm. What is the reduced model's computed probability that this possum is from Victoria? How confident are you in the model's accuracy of this probability calculation?

2. Medical school admission

The file MedGPA.csv in the data folder has information on medical school admission status and GPA and standardized test scores gathered on 55 medical school applicants from a liberal arts college in the Midwest.

The variables are:

Accept Status: A=accepted to medical school or D=denied admission Acceptance: Indicator for Accept: 1=accepted or 0=denied Sex: F=female or M=male BCPM: Bio/Chem/Physics/Math grade point average GPA: College grade point average VR: Verbal reasoning (subscore) PS: Physical sciences (subscore) WS: Writing sample (subcore) BS: Biological sciences (subscore) MCAT: Score on the MCAT exam (sum of CR+PS+WS+BS) Apps: Number of medical schools applied to

a. Build a logistic regression model to predict `Acceptance` from `GPA`.
b. Plot `Acceptance` versus `GPA`, add *jitter* in the vertical direction.
c. Repeat the plot in part b but add linear and logistic fitted line to the plot.
d. Check the linearity assumption by plotting `GPA` versus the logit of `Acceptance`, the response on the logit scale.