Home Exercises 8

Your Name

11.11.2024

Write your name at the beginning of the file as "author:".

1. Return to Moodle by **9.00am, Mon 11.11.** (to section "BEFORE").
2. Watch the exercise session video available in Moodle by **10.00am, Mon 11.11.**
3. If you observe during the exercise session that your answers need some correction, return a corrected version to Moodle (to section "AFTER") by **9.00 am, Mon 18.11.**

Problem 1.

Read in the data using `y = read.csv("Arrests.csv", header = TRUE, stringsAsFactors = TRUE)`. Read the explanation of the data: http://vincentarelbundock.github.io/Rdatasets/doc/carData/Arrests.html.

(i) To get familiar with the data, use `str( )` to see which types of variables they are and make a 2 x 4 plotting area by `par(mfrow = c(2,4))` and apply `barplot(table(y$VARIABLE))` to the 5 categorical variables (`released`, `colour`, `sex`, `employed citizen`) and apply `hist(y$VARIABLE)` to the 3 quantitative variables (`year`,`age`,`checks`). Remove the 1st column "X" by setting `y$X = NULL`.

(ii) We want to model whether individual is released with summons. Fit a logistic regression regressing `released` on all other 7 variables. You can use "." notation as in Example 8.2 from Lecture 8. Which variables seem the clearest non-zero predictors from the statistical point of view? Explain for them also which direction the effect goes? "White are more likely to be released" etc.

(iii) What is the probability that an individual was released with summons in year 2000 if individual was White, Employed, Male Citizen, 43 years old, and has 0 previous checks? What is the corresponding probability if the other parameters are the same but individual is Black, unemployed and not a citizen? (Hint: use `predict( ,type = "response")` function with `newdata` parameter.)

Problem 2.

Split arrests data from Problem 1 into three parts.

- `y.nc`, to have all non-citizens.
- `y.tr`, to have 3000 randomly chosen citizens for training
- `y.te`, to have the remaining 1455 citizens for testing

(Hint: Make a vector of all citizen indexes `cit.ind = which(y$citizen == "Yes")` and use `tr.ind = sample(cit.ind , size = 3000)` function to choose a random set of 3000 citizen indexes. Use then `setdiff(cit.ind, tr.ind)` to get the indexes of the remaining citizens to keep as test data.)

(i) Fit a logistic regression model in the training data by regressing `released` on other variables except `citizen` (as training data are all citizens).

(ii) Make three ROC curves in the same Figure by applying the model from Part (i) to the three data sets (non-citizens, training and testing). Do the three ROCs look as you expected relative to each other?

Problem 3.

Read in data using

`y = read.csv("Wells.csv", header = TRUE, stringsAsFactors = TRUE)`.

The data are from Bangladesh. The researchers labelled each well with its level of arsenic and an indication of whether the well was "safe" or "unsafe". Those using unsafe wells were encouraged to switch. After several years, it was determined whether each household using an unsafe well had changed its well. Here we have data on 3020 families that had originally an unsafe well. The question is which factors are associated with whether the family switched to a safer well. The data are explained at http://vincentarelbundock.github.io/Rdatasets/doc/carData/Wells.html.

(a) Plot histograms of `arsenic`, `distance` and `education` and barplots of `switch` and `association` (analogously to plotting in Problem 1).

(b) Fit a logistic regression model `switch ~ distance` to see how the distance to closest safe well affects switching probability. Show `summary()` and interpret the effect. Use `predict()` to make a prediction of switching probabilities for a grid of distances from 0 to 300m, with a step size of 10m. Plot the probabilities as a function of distance. Check from the figure, how does the probability of switching change if distance is a couple of meters compared to if it is 300m.

Problem 4. Continue with the wells data from Problem 3.

Fit a model for switching that includes `distance` and `education` and their **interaction term**. This means that we model the log-odds of switching by the formula

$$a + b_1 \cdot D + b_2 \cdot E + c \cdot D \times E,$$

where b_1 and b_2 are the **main effects** of Distance and Education, respectively, and c is their interaction effect. If $c \neq 0$, then the effect of Distance on switching depends on Education level, that is, Distance and Education are *interacting*. You can fit such a model by simply using formula `switch ~ distance * education` in `glm( )` call.

Take from this model two sets of predicted probabilities, both for the grid of `distance` values from 0 to 300, as in Problem 3. For the first set of probabilities, keep `education = 2` and for the second set keep `education = 10`. Plot these two sets of probabilities in the same figure.

Based on the plot, at which distance is the switch probability the same for a lower educated family as it is for a higher educated family that has the distance of 300m?

(Hint: Use `plot( )` command for the first set of probabilities and `lines( )` command for the second set of probabilities to add them into same figure. To get a constant level of education values (e.g. 2) across the distance grid `dists`, you can use `education = rep(2, length(dists))`.)