Logistic regression

November 6, 2019

Answers for this lab must be submitted before **November 13th at 5pm on Moodle**. In your answer for each question, please include a copy of the R code used (if applicable) and the results obtained.

1. Wells data revisited

Load the *carData* package to access the `Wells` data frame, containing data from a study of wells with high arsenic concentration in Bangladesh.

```r
library(carData)
head(Wells)
```

```
##   switch arsenic distance education association
## 1    yes    2.36   16.826         0          no
## 2    yes    0.71   47.322         0          no
## 3     no    2.07   20.967        10          no
## 4    yes    1.15   21.486        12          no
## 5    yes    1.10   40.874        14         yes
## 6    yes    3.90   69.518         9         yes
```

During the class on logistic regression, we modeled a household's decision to switch wells or not (`switch`) based on the arsenic concentration of the well (`arsenic`, in multiples of 100 $\mu g/L$) and the distance to the nearest safe well (`distance`, in m). The purpose of this exercise is to evaluate the effect of the `education` predictor, representing the number of years of edcuation of the responding adult for the household.

a) Estimate the parameters of a model in which the three variables (concentration of arsenic, distance and education) have an additive effect. Interpret *qualitatively* the coefficient of the variable `education`: according to the sign of this coefficient, what is the effect of education on the decision to switch wells or not?

b) In the context of this study, what would an **interaction** between the level of education and the concentration of arsenic mean for the decision to switch wells? Formulate a hypothesis about the possible direction of this interaction, with a brief explanation of your choice.

c) Estimate the parameters of the model with an interaction between the level of education and the arsenic concentration, then check whether the result corresponds to your hypothesis formulated in (b).

d) Based on the values of the AIC, compare the models in (a) and (c) with the base model fitted during the course (`switch ~ arsenic + distance`). What is the best model? Is it reasonable to make predictions with this model only?

e) From an appropriate diagnostic plot, verify that the residuals of the best model in (d) are distributed according to the logistic regression model.

f) Visualize the probability of switching wells predicted by the best model in (d) for arsenic concentrations between 0.5 and 5 (on the data scale) and for three levels of education (0, 5 and 10 years). The variable `distance` will not appear on the graph, but you can use a constant distance of 50 m for predictions.

2. Incidence of esophageal cancer

The `esoph` data frame included in R contains data from a study on the incidence of esophageal cancer in France as a function of age, alcohol consumption and tobacco use.

```r
head(esoph)
```

```
##   agegp     alcgp    tobgp ncases ncontrols
## 1 25-34 0-39g/day 0-9g/day      0        40
## 2 25-34 0-39g/day    10-19      0        10
## 3 25-34 0-39g/day    20-29      0         6
## 4 25-34 0-39g/day      30+      0         5
## 5 25-34     40-79 0-9g/day      0        27
## 6 25-34     40-79    10-19      0         7
```

The data indicate the number of people diagnosed with cancer (`ncases`) and the number of unaffected people (`ncontrols`) for each combination of categories of the three predictors.

The original table uses ordered factors for each predictor. Since the analysis of ordinal variables is not seen in this course, we convert these predictors to unordered factors (nominal variables).

```r
library(dplyr)
esoph <- mutate(esoph, agegp = factor(agegp, ordered = FALSE),
                alcgp = factor(alcgp, ordered = FALSE),
                tobgp = factor(tobgp, ordered = FALSE))
```

a) Fit a binomial logistic regression model to these data, assuming an additive effect of the three predictors, and then evaluate the McFadden pseudo-R^2 coefficient of the model.

b) Which of the two risk factors (alcohol or tobacco) increases the incidence of esophageal cancer the most for this population? Justify your answer.

c) Answer the following questions based on the coefficients of the model and using the `invlogit` function, which converts the value of the linear predictor into a probability.

```r
invlogit <- function(x) 1/(1 + exp(-x))
```

- How do you interpret the value of the `Intercept` coefficient?
- What is the probability of cancer incidence for a person aged 55 to 64 who consume less than 39 g of alcohol and less than 9 g of tobacco a day?

d) If the number of cases and controls were reversed in the model formula: `cbind(ncontrols, ncases) ~ agegp + alcgp + tobgp`, what would be the effect on the estimated coefficients? Try to predict the answer based on your knowledge of the logistic model, then check with R.