Probability, Statistics and Modelling II

Tutorial 6 –Binary Logistic Regression, Fitted Probabilities

This time, we will be using a new dataset: selected variables from the Crime Survey for England and Wales (CSEW). I will list the name of the variables below, for details on each of them please refer to the uploaded questionnaire and/or use R.

Variables in the dataset:

- year – year of data collection
- nadults – number of adults in the household
- sex – sex of the respondent
- age – age of the respondent
- nchil – number of children (i.e. people below 16) in the household
- yrsarea – years a person lived in the area
- ownbike – bike owners
- perclc2 – perception of changes in crime in the area
- pubeve – frequency of going to pubs
- club – frequency of going to clubs
- inner – living in an inner city
- rural3 – living in a rural area
- rubbcomm – rubbish in the area
- vandcomm – signs of vandalism in the area
- mottheft – motorbike/moped having been stolen
- biktheft – bicycle having been stolen

Please carry out the tasks and answer the questions below.

Congratulations! You have been hired by the Home Office as a government analyst. Your first task is to use the Crime Survey for England and Wales to estimate the probability of someone's bike being stolen to develop an effective prevention strategy.

1.Familiarise yourself with the dataset and the outcome variable. What do you notice?

2.There are several working hypotheses in the Home Office regarding what might increase the probability of a bike being stolen. At each step, incorporate one more of these aspects into your model and interpret your findings by using the odds ratios and confidence intervals. What is the null-hypothesis being tested here?

a) Many people think that bike theft is primarily an “inner-city” problem.

b) Others suggest that the perpetrators might be emboldened in areas where there are clear signs of disorder (e.g. graffiti on the walls, high amount of rubbish on the street, as argued by the broken windows theory).

c) Some believe that living in an area for a longer period might be a protective factor (you learn where (not) to store your bike).

d) Other analysts point out that the number of years a person could have lived in a certain area is strongly associated with age, so controlling for that would be a good idea. In addition, young people are more likely to use bikes, which makes it more likely that they would have their bikes stolen.

e) Finally, there is some anecdotal evidence that people who go to pubs and clubs are more likely to have their bikes stolen.

f) Compare these models to one another. How do they change? What kind of conclusion can you draw from these changes? What kind of policy-recommendation would you propose?

3.It is fairly difficult to interpret log-odds and odds-ratios, predicted probabilities are much more straightforward. Follow the instructions in the R-script and estimate the predicted probabilities of the model. How would you interpret these results?

4.The Home Office would like to know which variable is the most influential from the above list. Estimate the marginal effects to answer this question. What would be your response to the Home Office? How would you interpret the results?

5.Finally, run the different tests of model fit (Hosmer-Lemeshow-test and various pseudo-R^2 values). How much improvement do you notice comparing model 1 and model 5? How would you interpret these results?