Bayes assignment on regression lines

Sean van der Merwe

2022/05/05

Instructions

You are provided with a set of data from an anonymous survey where people with desktop computers were asked how many cookies were set on their favourite browser. Additionally, information was obtained on their internet privacy habits. Your goal is to determine the relationships between the number of cookies and the habits reported.

Specifically, the questions were:

1. What is the number of cookies reported by your favourite browser?
2. How many times a year do you clear your browser cookies and/or history?
3. Do you use a virtual private network (VPN) regularly when connecting to the internet?

Part 1: Visualisation [15]

Read in and visualise your data. Give summary statistics overall and by VPN group, as well as a plot where the groups can be identified. Discuss any apparent patterns without making any firm conclusions yet.

Part 2: First model fit [30]

A Negative Binomial GLM formulation (with log link function) is recommended as a starting point. Fit such a model and show the model fit on a plot along with the data to assess the fit. You must show uncertainty via prediction intervals at least.

In case it isn't clear from the data itself: the dependent variable is the raw number of cookies; while the explanatory variables should be the square root of the number of clears, as well as a binary indicator of VPN use.

Part 3: Changing variance [10]

There are good fundamental reasons to assume that VPN users would vary more in terms of number of cookies (after adjusting for the other factor measured). Calculate a reasonable posterior probability that this is the case (this may require adjusting your model to allow for different variances).

Part 4: Model comparison [25]

Ensure that you have fitted a model that differentiates between VPN and non-VPN users, and one that does not (more than these two is fine, less is not). Compare all your models using at least one good criterion and say which model seems the most parsimonious.

Part 5: Final step [20]

Pretend you answered the survey. What is the predicted number of cookies for you personally, according to the model you think is best? Where does the actual number (as reported by your favourite browser) lie on the predicted distribution?