

Practice 9

You need to use the R packages **faraway** and **nnet** to work on the following questions.

The **hsb** data was collected from the High School and Beyond Study. We want to see how the relevant variables in the data are related to the choice of **program** with 3 types — academic, vocational, or general — that the students pursue in high school. The response variable **prog** may be regarded as following a multinomial distribution with three levels.

1. Type `help(hsb)` to see its description. Conduct an exploratory data analysis on **hsb** to better understand the **hsb** data. For example, check the size of the data, the type of each variable (categorical, factor, ordered factor, numerical), etc.
2. Fit a trinomial logistic model with **prog** as the response and including 1 as the only predictor (i.e. the null model). Save the results into **hsb0**. Then explore **hsb0** using the commands such as `summary`, `anova`, `fitted`, `prediction`, and `deviance` etc. to see whether you understand the R outcomes and are able to interpret them.
3. Fit a trinomial logistic model with **prog** as the response and all other variables except **id** as predictors (untransformed, and no interaction terms). Save the results into **hsb1**. Then explore **hsb1** using the commands such as `summary`, `anova`, `fitted`, `prediction`, and `deviance` etc. to see whether you understand the R outcomes and are able to interpret them. Also compare **hsb1** with **hsb0** using the `anova` command.

Note: Change of deviance between two multinomial logit models can still be used to test the difference between the two models, which approximately follows a χ^2 distribution. But the deviance based χ^2 test cannot be used to reliably test the goodness of fit of a multinomial logit model. Other methods are needed.

4. Perform variable selection based on **hsb1** using `step` function with AIC or BIC option. Save the results into **hsb1.aic** and **hsb1.bic** respectively.
5. Compare **hsb1** with **hsb1.aic** and **hsb1.bic**.
6. There are two students A and B who have the same math, science and social science scores. Student A comes from a high ses class and private school, while student B comes from a low ses class and public school. Consider the model **hsb1.aic**.
 - (a) Estimate the odds ratio of choosing **general** program against **academic** program for student A versus student B. Find an approximate 95% confidence interval for this odds ratio.
 - (b) Estimate the odds ratio of choosing **vocation** program against **academic** program for student A versus student B. Find an approximate 95% confidence interval for this odds ratio.
 - (c) Estimate the odds ratio of choosing **general** program against **vocation** program for student A versus student B. Find an approximate 95% confidence interval for this odds ratio.
7. For the student with id 99, compute the predicted probabilities of the three possible choices based on the best model among **hsb1**, **hsb1.aic** and **hsb1.bic**.