Research Methods in Political Science II – Exercises for Day 3

Q1. Linear Regression

We are using the data set `data-attendance-1.txt` prvided by Matsuura (2016). This data set contains three variables, A, $Score$, and Y. Each row represents a student at some university. A is a dummy variable indicating that a student likes their part-time job. $Score$ measures how much a student likes their classes, ranging from 0 (doesn't like at all) to 200 (likes very much). Lastly, Y is attendance rate of each student.

Using this data set, answer the following questions.

1. Explore the data, and visualize bivariate relationships between variables.
2. Build a normal-linear model to explain the attendance rate. What results do you expect to get by fitting the model?
3. Fit the model with RStan and check the convergence.
4. Visualize and interpret the result.
   (a) Plot the effects of each explanatory variable.
   (b) Visualize the relationship between the outcome and the explanatory variables.
   (c) Visualize the relationship between the predicted outcome and the observed outcome.
   (d) Etc.

Q2. Binomial Model

We are using the data set `data-attendance-2.txt` prvided by Matsuura (2016). This data set contains five variables, $PersonID$, A, $Score$, M, and Y. Each row represents a student at some university. A and $Score$ are same as Q1. PersonID is the ID variable. M is the number of classes that each student is supposed to attend.

Y is the number of attendances out of M.

Using this data set, answer the following questions.

1. Explore the data, and visualize bivariate relationships between variables.
2. Build a binomial model to explain the number of attendances. What do you expect to get by fitting the model?
3. Fit the model with RStan and check the convergence.
4. Visualize and interpret the result.
   (a) Plot the effects of the explanatory variables on the outcome.
   (b) Visualize the relationship between the predicted and observed outcomes.
   (c) Plot the relationship between the predicted attendance *rate* and the observed attendance *rate*.
5. What is the advantage of using the count instead of rate?
6. Etc.

Q3. Bernoulli Model

We are using the data set `data-attendance-3.txt` prvided by Matsuura (2016). This data set contains five variables, $PersonID$, A, $Score$, $Weather$, and Y. Each row represents a student-class. $PersonID$, A, and $Score$ are same as Q2. $Weather$ is a categorical variable that has three categories: A, B, and C represent "Clear", "Cloudy", and "Rainy", respectively. Y is a binary variable indicating that the student attended the class.

Using this data set, answer the following questions.

1. Quantify the categorical variable Weather. For this exercise, let's assume that the bad weather decrease the probability of attendance, and the effect of rain is five times larger than that of cloud. So convert the categories (A, B, C) into the numbers (0, 0.2, 1).
2. Explore the data, and visualize bivariate relationships between variables.

3. Build a Bernoulli model to explain the attendance. What do you expect to get by fitting the model?
4. Fit the model with RStan and check the convergence.
5. Visualize and interpret the result.
   (a) Plot the effects of the explanatory variables on the outcome.
   (b) Evaluate the model by hit rate.
   (c) Visualize the relationship between the explanatory variables, the predicted probability of attendance, and the observed outcome.
   (d) Visualize the relationship between the predicted probability and observed outcome.
   (e) Draw the ROC curve and evaluate the fit.

Reference

Matsuura, K (2016): 松浦健太郎.『R と Stan でベイズ統計モデリング (Bayesian Statistical Modeling with R and Stan)』共立出版.