Group members: Zhijian Liu, Zhongyan Liang

Multilevel Logistic Regression

Introduction

Multilevel modeling is a generalization of regression methods, and as such can be used for a variety of purposes, including prediction, data reduction, and causal inference from experiments and observational studies (Kreft and De Leeuw 1998). It is an common approach that can be used to handle clustered or grouped data.

For example, supposedly we are trying to estimate what affects a child's academic performance at school. The factor involves in our study will be what was taught in school, the socioeconomic status of the child's parents, class level factors, single-sex vs mixed and etc. Multi-level modelling provides a useful framework for thinking about problems with this type of hierarchical structure (Browne, 2004).

Dataset Information

We obtained our data from Machine Learning UCI Repository. The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to build a model to predict if the client will subscribe a term deposit (variable Y). We want to know if marketing campaign will affect the decision of customers to subscribe a term deposit. The marketing campaigns were based on phone calls. more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

As the data description below, our dataset can be regarded as two levels. The first level contains personal characteristics of clients such as Age, type of job, marital status, education level, has credit or not, if they has housing loan or personal loan. Note that the first level is not related to the second level, campaign. It contains the communication type that the campaign was conducted by, which month the client was contacted and which day of the week the client was contacted. Except Age, all other variables are categorical.

Bank Client Data Level 1							
Y(Response)	Age	Job	Marital	Education	Default	Housing	Loan
YES/NO	Numeric	Job type	Marital status	Academic Level	Has Credit	Has house loan	Has Personal Loan
Related with the last contact of the current campaign Level 2							
Y(Response)	Contact	Month	Day of Week				
YES/No	Communication Type	last contact month of year	last contact day of the week				

Our purpose is to illustrate how to build a multilevel model using this data set.

Building Multilevel Logistic Regression Model

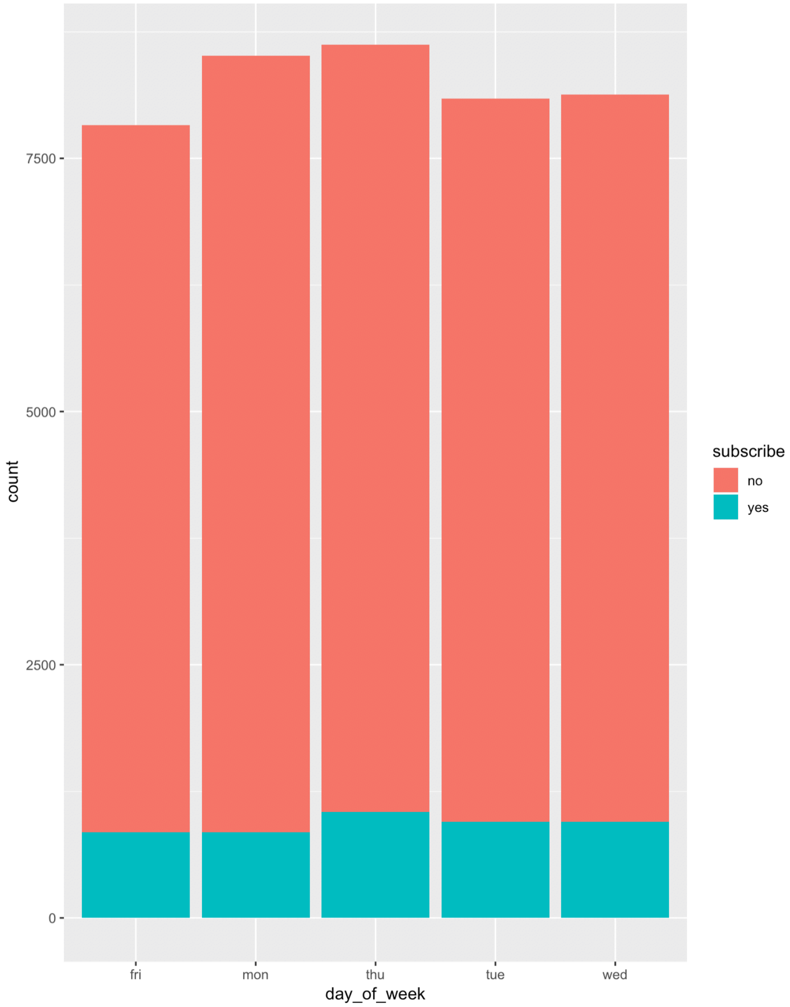
1. Variable Selection

We decided to initialize with a logistic regression model, shown as:

$$log\frac{P(\text{Subscribe})}{P(\text{Not subscribe})} = \beta_0 + \beta\mathbf{X} + \gamma\mathbf{U}$$

Where $\beta\boldsymbol{X}$ and $\gamma\boldsymbol{U}$ are correspondingly the linear combination of the variables in the first level, individual information, and the second level, campaign information. We started by performing stepwise variable selection on this initial model. For individual level, stepwise removed Housing and Loan. While for campaign, all the variables was kept.

However, we think that Day-of-Week should not be a significant variable in this model because whether you take phone calls in whichever weekdays should not affect if you are going to subscribe a term deposit. To validate this intuition, we started by plotting Day-of-Week:



We can see that there are no obvious differences between each weekdays against the proportion of subscription. Furthermore, we checked the significance of each dummy variables in Day-of-Week by looking at their significance of their Coefficients in the summary. The output shown that only Tuesday is resulted significant. So we want to further confirm our theory by using Likelihood Ratio Test. We set the reduced model without Day_of_Week and full model includes Day-of-Week. The test failed to reject the reduced model in the null hypothesis. Thus, we dropped the Day_of_Week variable.

2. Group-Level Data

After removing Day-of-Week, we had two variables left in Level 2, which were Contact(communication type) and Month(month of the year). There were 2 types of Contact and 10 different Months, based on which we generated 20 different types of campaign to be used in the latter multilevel regression model.

3. Multi-level Model

Now we have identified how many types of campaign there are, we can build our multilevel model. The model looks like this:

$$log\frac{P(\text{Subscribe})}{P(\text{Not subscribe})} = \alpha_j + \beta X$$

$$Where\ \ \alpha_j \sim N(\mu_\alpha, \sigma_\alpha), \text{j} = 1,2,\ldots,20$$

$\boldsymbol{\beta X}$ is the linear combination of the selected individual variables after stepwise variable selection. And the changeable intercept α_j follows a normal distribution, with mean and standard deviation calculated from complete pooling, putting all campaign category into a whole.

The first line is the model for the first level, the response variable is the same as our original model, beta is the coefficient of variables for individuals, which is constant and alpha is the coefficient based on twenty different types of campaign. The second line is the model for the second level in which alpha is normally distributed.

Conclusion

Based on the multilevel model, we can estimate how the probability of subscription changes across all the campaign type. Although the data we use is not nested data, that means individuals are not physically nested in each campaign type, multilevel model is still helpful. For one reason, the multilevel model we are using essentially reduced the number of parameters. If we used the classical logistic regression, a large number of dummy variables would be generated in the model leading to difficulties of interpretation and high variance. Secondly, assuming the effect of campaign to follow some distribution can help to make more stable or smooth estimation.

To improve the model, we might also try to build a new level for Job(Variable in Level 1), since there are many categories of job in the level 1 model. These levels and levels of campaign can mutually contribute to the individual level model. In this case, non-nested multilevel model can be used. But since job may be related to other individual variables, we might need to allow the slopes of individual variables to change with levels of Job.

Reference

Kreft, I., and De Leeuw, J. (1998), Introducing Multilevel Modeling, London: Sage.

Sommet, Nicolas, and Davide Morselli. "Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS." International Review of Social Psychology 30.1 (2017).

Browne, W. and Rasbash, J. (2004). ‘Multilevel Modelling’, in Hardy, M. and Bryman, A. (eds.), Handbook of data analysis, Sage Publications, pp 459-78.