MA678 homework 08

Name

November 10, 2017

presidential preference and income for the 1992 election

The folder `nes` contains the survey data of presidential preference and income for the 1992 election analyzed in Section 5.1, along with other variables including sex, ethnicity, education, party identification, political ideology, and state.

1. Fit a logistic regression predicting support for Bush given all these inputs except state. Consider how to include these as regression predictors and also consider possible interactions.
2. Now formulate a model predicting support for Bush given the same inputs but allowing the intercept to vary over state. Fit using `lmer()` and discuss your results.
3. Create graphs of the probability of choosing Bush given the linear predictor associated with your model separately for each of eight states as in Figure 14.2.

Three-level logistic regression:

the folder `rodents` contains data on rodents in a sample of New York City apartments.

1. Build a varying intercept logistic regression model (varying over buildings) to predict the presence of rodents (the variable rodent2 in the dataset) given indicators for the ethnic groups (race) as well as other potentially relevant predictors describing the apartment and building. Fit this model using lmer() and interpret the coefficients at both levels.
2. Now extend the model in (1) to allow variation across buildings within community district and then across community districts. Also include predictors describing the community districts. Fit this model using lmer() and interpret the coefficients at all levels.
3. Compare the fit of the models in (1) and (2).

Item-response model:

the folder `exam` contains data on students' success or failure (item correct or incorrect) on a number of test items. Write the notation for an item-response model for the ability of each student and level of difficulty of each item.

Multilevel logistic regression

The folder `speed.dating` contains data from an experiment on a few hundred students that randomly assigned each participant to 10 short dates with participants of the opposite sex (Fisman et al., 2006). For each date, each person recorded several subjective numerical ratings of the other person (attractiveness, compatibility, and some other characteristics) and also wrote down whether he or she would like to meet the other person again. Label $y_{ij} = 1$ if person i is interested in seeing person j again 0 otherwise. And $r_{ij1}, \ldots, r_{ij6}$ as person i's numerical ratings of person j on the dimensions of attractiveness, compatibility, and so forth. Please look at http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/Speed%20Dating%20Data%20Key.doc for details.

```r
dating<-fread("http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/Speed%20Dating%20Data.csv
```

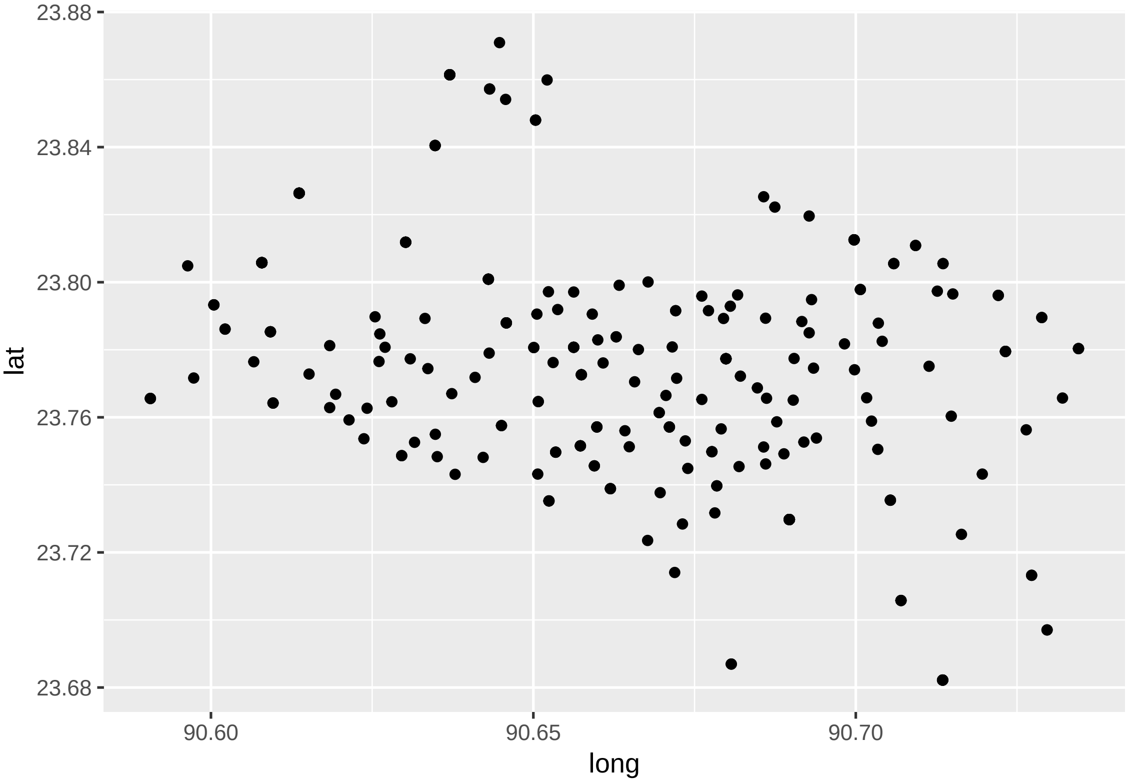
1. Fit a classical logistic regression predicting $Pr(y_{ij} = 1)$ given person i's 6 ratings of person j. Discuss the importance of attractiveness, compatibility, and so forth in this predictive model.
2. Expand this model to allow varying intercepts for the persons making the evaluation; that is, some people are more likely than others to want to meet someone again. Discuss the fitted model.
3. Expand further to allow varying intercepts for the persons being rated. Discuss the fitted model.
4. You will now fit some models that allow the coefficients for attractiveness, compatibility, and the other attributes to vary by person. Fit a no-pooling model: for each person i, fit a logistic regression to the data y_{ij} for the 10 persons j whom he or she rated, using as predictors the 6 ratings $r_{ij1}, \dots, r_{ij6}$. (Hint: with 10 data points and 6 predictors, this model is difficult to fit. You will need to simplify it in some way to get reasonable fits.)
5. Fit a multilevel model, allowing the intercept and the coefficients for the 6 ratings to vary by the rater i.
6. Compare the inferences from the multilevel model in (5) to the no-pooling model in (4) and the complete-pooling model from part (1) of the previous exercise.

The well-switching data described in Section 5.4 are in the folder arsenic.

1. Formulate a multilevel logistic regression model predicting the probability of switching using log distance (to nearest safe well) and arsenic level and allowing intercepts to vary across villages. Fit this model using `lmer()` and discuss the results.

```
##   [1] 90.65935 90.65597 90.65880 90.65495 90.65626       NA 90.65306
##   [8] 90.65233 90.64581 90.64639 90.63671 90.64581 90.64581 90.64302
##  [15] 90.64362 90.64302 90.64302 90.63558 90.64302 90.65269 90.66257
##  [22] 90.65241 90.65117 90.65009 90.64803 90.65639 90.72322 90.72322
##  [29] 90.72322 90.72322 90.72322 90.71482       NA       NA 90.70537
##  [36] 90.70537 90.66840 90.72209 90.72209       NA       NA       NA
##  [43] 90.71137 90.71962 90.71714 90.73207 90.58984 90.59430 90.58820
##  [50] 90.61552 90.59997 90.62609 90.63542 90.61346 90.63092 90.61186
##  [57] 90.59872 90.60423 90.62412 90.64314 90.61513 90.61180 90.61643
##  [64] 90.62983 90.61935 90.62513 90.63485 90.62372 90.60387       NA
##  [71] 90.61828 90.62245 90.60233 90.60045 90.60045       NA 90.61521
##  [78] 90.60789 90.60789 90.61098 90.60789 90.60789 90.60789       NA
##  [85] 90.60789 90.60789 90.60705 90.60789 90.60789 90.60657 90.60789
##  [92] 90.59640 90.62541 90.62878 90.63019 90.63019 90.60687 90.63019
##  [99] 90.63289 90.61128 90.60923 90.60923       NA 90.60093 90.60734
## [106] 90.67717 90.68167 90.67949 90.67207 90.67207 90.69164 90.65744
## [113] 90.65744 90.65701 90.64367 90.66587       NA       NA       NA
## [120] 90.68054 90.68602 90.66082 90.68008 90.66220       NA 90.65906
## [127] 90.66789 90.67159 90.66779 90.63025 90.64392 90.63984 90.65347
## [134] 90.62741 90.62104       NA       NA 90.63289 90.62896       NA
## [141] 90.64439 90.64953 90.64977 90.65729 90.65729 90.64505 90.65729
## [148]       NA 90.65984 90.65984 90.65364 90.64108 90.64336 90.67109
## [155] 90.62524 90.66380 90.66194 90.66194 90.66194 90.64562 90.64447
## [162] 90.65947 90.65947 90.65069 90.64039 90.65872 90.62859       NA
## [169] 90.71639 90.70702 90.70702 90.70702 90.72727 90.72727 90.72967
## [176] 90.72967 90.71347 90.71347 90.71347 90.71347 90.71347 90.71347
## [183] 90.68069 90.66670 90.66714       NA 90.68874 90.68969 90.68969
## [190] 90.69063 90.67349 90.68525 90.67970 90.67774 90.68969 90.68969
## [197] 90.68969 90.67331 90.68969 90.67396 90.67659 90.68969 90.68969
## [204] 90.67361 90.68969 90.69442 90.68613 90.68643 90.68969       NA
## [211] 90.70244 90.68212 90.70169 90.68474 90.68617 90.67588 90.67987
## [218] 90.67987 90.66844 90.67987 90.67987 90.70409 90.69344 90.69824
## [225] 90.69029       NA 90.67154 90.69982 90.69044 90.69277 90.71504
## [232] 90.71354 90.71354 90.69314 90.70589 90.70589 90.70589 90.69277
## [239] 90.68572 90.68744 90.70349 90.70072 90.70072 90.70927 90.70927
## [246] 90.71264 90.69974 90.69974 90.69974 90.69974 90.69974 90.67349
## [253] 90.63477 90.63477 90.63477 90.63477 90.63477 90.64476 90.64569
## [260] 90.64322 90.64322 90.61368 90.61368 90.61368 90.61368 90.61368
## [267] 90.61368 90.61368 90.61368 90.61368 90.61368 90.65035 90.65035
## [274] 90.65035 90.63703 90.63703 90.63703 90.63703 90.63703 90.63703
## [281] 90.63703 90.63703 90.65213 90.68773 90.67769 90.67769 90.67769
## [288] 90.66809 90.67436 90.70342 90.67494 90.68602 90.66967 90.67807
## [295] 90.67516 90.67329 90.68572 90.67321       NA 90.67102 90.67082
## [302] 90.69224

## Warning: Removed 7 rows containing missing values (geom_point).
```



2. Extend the model in (1) to allow the coefficient on arsenic to vary across village, as well. Fit this model using `lmer()` and discuss the results.
3. Create graphs of the probability of switching wells as a function of arsenic level for eight of the villages.
4. Compare the fit of the models in (1) and (2).