MA678 homework 09

Megha Pandit

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.

```
m1 <- glm(rvote ~ age + female + educ1 + income + occup1 + partyid7, data = data,
              family = binomial(link = logit))
summary(m1)
```

```
##
## Call:
## glm(formula = rvote ~ age + female + educ1 + income + occup1 +
##     partyid7, family = binomial(link = logit), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6496  -0.6646   0.2670   0.5216   2.3136
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.555579   0.153903 -23.103  < 2e-16 ***
## age          0.004588   0.001582   2.900  0.00373 **
## female      -0.163211   0.050326  -3.243  0.00118 **
## educ1       -0.066236   0.028829  -2.298  0.02158 *
## income       0.143308   0.024026   5.965 2.45e-09 ***
## occup1       0.075972   0.015259   4.979 6.40e-07 ***
## partyid7     0.851908   0.013464  63.271  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18528  on 13375  degrees of freedom
## Residual deviance: 11428  on 13369  degrees of freedom
##   (21532 observations deleted due to missingness)
## AIC: 11442
##
## Number of Fisher Scoring iterations: 5
```

The model seems to give coefficient estimates that are statistically significant at two standard errors.

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.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rvote ~ age + female + educ1 + income + occup1 + partyid7 + (1 |
##     state)
##    Data: data
##
## REML criterion at convergence: 11263.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3572 -0.6176 -0.1242  0.5039  2.6940
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.003528 0.0594
##  Residual             0.134398 0.3666
## Number of obs: 13376, groups:  state, 47
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) -0.1159856  0.0221301  -5.241
## age          0.0004111  0.0002109   1.949
## female      -0.0229373  0.0068132  -3.367
## educ1       -0.0109317  0.0038854  -2.814
## income       0.0248797  0.0032675   7.614
## occup1       0.0090030  0.0020296   4.436
## partyid7     0.1520337  0.0014820 102.586
##
## Correlation of Fixed Effects:
##          (Intr) age    female educ1  income occup1
## age      -0.615
## female   -0.092  0.027
## educ1    -0.430  0.160 -0.123
## income   -0.424  0.224  0.115 -0.274
## occup1   -0.392  0.040 -0.331  0.362  0.017
## partyid7 -0.094 -0.081  0.021 -0.133 -0.105 -0.045
```

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.

```
m3 <- lmer(rodent2 ~ race + stories + scale(totincom2) + housing + poverty + (1|bldg), data = apt_dt)
summary(m3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## rodent2 ~ race + stories + scale(totincom2) + housing + poverty +
##     (1 | bldg)
##    Data: apt_dt
```

```
##
## REML criterion at convergence: 1595.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.90801 -0.48386 -0.28457  0.02399  2.42730
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  bldg     (Intercept) 0.05523  0.2350
##  Residual             0.11735  0.3426
## Number of obs: 1522, groups:  bldg, 900
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)       0.080801   0.058620   1.378
## race              0.039235   0.007819   5.018
## stories           0.022730   0.006973   3.260
## scale(totincom2) -0.038732   0.010937  -3.541
## housing          -0.004070   0.014299  -0.285
## poverty           0.042121   0.027260   1.545
##
## Correlation of Fixed Effects:
##             (Intr) race   storis scl(2) housng
## race        -0.326
## stories     -0.660  0.011
## scl(ttncm2)  0.026  0.079 -0.071
## housing     -0.857  0.042  0.384 -0.071
## poverty     -0.050 -0.068 -0.090  0.285  0.013
```

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.

```r
model1 <- glm(match ~ attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o, data = dating,
              family = binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = match ~ attr_o + sinc_o + fun_o + amb_o + intel_o +
##     shar_o, family = binomial, data = dating)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5300  -0.6362  -0.4420  -0.2381   3.1808
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.62091    0.21859 -25.714  < 2e-16 ***
## attr_o       0.22047    0.02388   9.233  < 2e-16 ***
## sinc_o      -0.01996    0.03067  -0.651   0.5152
## fun_o        0.25315    0.02922   8.665  < 2e-16 ***
## amb_o       -0.12099    0.02838  -4.264 2.01e-05 ***
## intel_o      0.07176    0.03716   1.931   0.0535 .
## shar_o       0.21225    0.02209   9.608  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6466.6  on 7030  degrees of freedom
## Residual deviance: 5611.0  on 7024  degrees of freedom
##   (1347 observations deleted due to missingness)
## AIC: 5625
##
## Number of Fisher Scoring iterations: 5
```

From the fitted model, $Logodds(match = 1) = -5.62 + 0.22attr_o - 0.02sinc_o + 0.25fun_o - 0.12amb_o + 0.07intel_o + 0.21shar_o$ Therefore, a unit increase in attractiveness will lead to an increase of $\frac{0.22}{4} = 0.055$ or 5.5% in the willingness to have another date.

Similarly, a unit increase in sincerity decreases the willingness to have another date by $\frac{0.02}{4} = 0.005$ or 0.5%. But this coefficient is not statistically significant at two standard errors and hence may not be influential in switching the willingness for another date from 1 to 0.

One unit increase in humor increases the willingness for another date by $\frac{0.25}{4} = 0.0625$ or 6.25%.

One unit increase in ambition decreases the willingness for another date by $\frac{0.12}{4} = 0.03$ or 3%.

A unit increase in intelligence increases the willingness to have another date by $\frac{0.07}{4} = 0.0175$ or 1.75%.

One unit increase in shared interest increases the willingness to have anotehr date by $\frac{0.21}{4} = 0.0525$ or 5.25%.

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.

```
model2 <- glmer(match ~ scale(attr_o) + scale(sinc_o) + scale(fun_o) + scale(amb_o) + scale(intel_o) +
                family = binomial)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.119726 (tol =
## 0.001, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
summary(model2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## match ~ scale(attr_o) + scale(sinc_o) + scale(fun_o) + scale(amb_o) +
##     scale(intel_o) + scale(shar_o) + gender + (1 | iid)
##    Data: dating
##
##      AIC      BIC   logLik deviance df.resid
##   5543.2   5605.0  -2762.6   5525.2     7022
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7459 -0.4453 -0.2877 -0.1454 10.3718
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  iid    (Intercept) 0.4294   0.6553
## Number of obs: 7031, groups:  iid, 551
##
## Fixed effects:
##                  Estimate Std. Error  z value Pr(>|z|)
## (Intercept)    -2.1320649  0.0006220 -3427.62   <2e-16 ***
## scale(attr_o)   0.4605811  0.0006553   702.88   <2e-16 ***
## scale(sinc_o)  -0.0249443  0.0006216   -40.13   <2e-16 ***
## scale(fun_o)    0.5132471  0.0006217   825.56   <2e-16 ***
## scale(amb_o)   -0.2352678  0.0006420  -366.45   <2e-16 ***
## scale(intel_o)  0.1087984  0.0006420   169.46   <2e-16 ***
## scale(shar_o)   0.4845416  0.0006217   779.39   <2e-16 ***
## gender          0.1542607  0.0006554   235.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) scl(t_) scl(sn_) scl(f_) scl(m_) scl(n_) scl(sh_)
## scale(ttr_)  0.000
## scale(snc_)  0.000  0.000
## scale(fun_)  0.000  0.000   0.000
## scale(amb_)  0.000  0.000   0.000    0.000
## scale(ntl_)  0.000  0.000   0.000    0.000  -0.250
## scale(shr_)  0.000  0.000   0.000    0.000   0.000   0.000
## gender       0.000  0.200   0.000    0.000   0.000   0.000   0.000
## convergence code: 0
## Model failed to converge with max|grad| = 0.119726 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

Fixed Effects: $logodds(match = 1) = -2.13 + 0.46attr_o - 0.02sinc_o + 0.51fun_o - 0.23amb_o + 0.11intel_o + 0.48shar_o + 0.15gender$ one unit increase in attractiveness increases the willingness to have another date by $\frac{0.46}{4} = 0.115$ or 11.5%.

one unit increase in sincerity decreases the willingness for another date by $\frac{0.02}{4} = 0.005$ or 0.5%.

a unit increase in humor increases the willingness to have another date by $\frac{0.51}{4} = 0.1275$ or 12.75%.

one unit increase in ambition decreases the willingness to have another date by $\frac{0.23}{4} = 0.0575$ or 5.75%.

one unit increase in intelligence increases the willingness for another date by $\frac{0.11}{4} = 0.0275$ 0r 2.75%.

a unit increase in shared interest increases the willingness to have another date by $\frac{0.48}{4} = 0.12$ or 12%.

Compared to a female dating partner, a male partner is $\frac{0.15}{4} = 0.0375$ or 3.75% more likely to have another date.

Random Effects: For person 1: $logodds(match = 1) = -1.64 + 0.46attr_o - 0.02sinc_o + 0.51fun_o - 0.23amb_o + 0.11intel_o + 0.48shar_o + 0.15gender$

For person 2: $logodds(match = 1) = -2.13 + 0.46attr_o - 0.02sinc_o + 0.51fun_o - 0.23amb_o + 0.11intel_o + 0.48shar_o + 0.15gender$

For person 3: $logodds(match = 1) = -2.54 + 0.46attr_o - 0.02sinc_o + 0.51fun_o - 0.23amb_o + 0.11intel_o + 0.48shar_o + 0.15gender$

For person 4: $logodds(match = 1) = -2.24 + 0.46attr_o - 0.02sinc_o + 0.51fun_o - 0.23amb_o + 0.11intel_o + 0.48shar_o + 0.15gender$

3. Expand further to allow varying intercepts for the persons being rated. Discuss the fitted model.

```
model3 <- glmer(match ~ scale(attr_o) + scale(sinc_o) + scale(fun_o) + scale(amb_o) + scale(intel_o) +
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.185818 (tol =
## 0.001, component 1)
```

```
summary(model3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## match ~ scale(attr_o) + scale(sinc_o) + scale(fun_o) + scale(amb_o) +
##     scale(intel_o) + scale(shar_o) + gender + (1 | iid) + (1 |      pid)
##    Data: dating
##
##      AIC      BIC   logLik deviance df.resid
##   5257.6   5326.1  -2618.8   5237.6     7021
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7829 -0.3827 -0.2194 -0.0917  9.1667
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  iid    (Intercept) 0.5932   0.7702
##  pid    (Intercept) 1.2592   1.1222
## Number of obs: 7031, groups:  iid, 551; pid, 537
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.53935    0.11740 -21.629  < 2e-16 ***
## scale(attr_o)   0.63782    0.06373  10.008  < 2e-16 ***
## scale(sinc_o)   0.03540    0.06785   0.522   0.6019
## scale(fun_o)    0.57774    0.07099   8.138 4.02e-16 ***
## scale(amb_o)   -0.16654    0.06466  -2.576   0.0100 *
## scale(intel_o)  0.17128    0.07359   2.327   0.0199 *
## scale(shar_o)   0.58891    0.06158   9.564  < 2e-16 ***
## gender          0.17340    0.14943   1.160   0.2459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) scl(t_) scl(sn_) scl(f_) scl(m_) scl(n_) scl(sh_)
## scale(ttr_) -0.221
## scale(snc_) -0.049 -0.064
## scale(fun_) -0.140 -0.220  -0.123
## scale(amb_)  0.072 -0.051   0.011   -0.168
## scale(ntl_) -0.009 -0.024  -0.438   -0.098  -0.334
## scale(shr_) -0.139 -0.072  -0.057   -0.234  -0.159  -0.020
## gender      -0.647  0.093   0.037    0.009  -0.070  -0.044   0.004
## convergence code: 0
## Model failed to converge with max|grad| = 0.185818 (tol = 0.001, component 1)
```

All the coefficients estimates, except the ones for sincerity and gender, seem to be significant at two standard errors.

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 $r_{ij1}, \ldots, r_{ij6}$. (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.

```
model5 <- glm(match ~ (1 + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o | iid) +
              attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o, data = dating,
              family = binomial)
summary(model5)
```

```
##
## Call:
## glm(formula = match ~ (1 + attr_o + sinc_o + fun_o + amb_o +
##     intel_o + shar_o | iid) + attr_o + sinc_o + fun_o + amb_o +
##     intel_o + shar_o, family = binomial, data = dating)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5300  -0.6362  -0.4420  -0.2381   3.1808
##
## Coefficients: (1 not defined because of singularities)
##                                                                   Estimate
## (Intercept)                                                       -5.62091
## 1 + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o | iidTRUE       NA
## attr_o                                                             0.22047
## sinc_o                                                            -0.01996
## fun_o                                                              0.25315
## amb_o                                                             -0.12099
## intel_o                                                            0.07176
## shar_o                                                             0.21225
##                                                                   Std. Error
## (Intercept)                                                          0.21859
## 1 + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o | iidTRUE         NA
## attr_o                                                               0.02388
## sinc_o                                                               0.03067
## fun_o                                                                0.02922
## amb_o                                                                0.02838
## intel_o                                                              0.03716
## shar_o                                                               0.02209
##                                                                   z value
## (Intercept)                                                       -25.714
## 1 + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o | iidTRUE      NA
## attr_o                                                              9.233
## sinc_o                                                             -0.651
## fun_o                                                               8.665
## amb_o                                                              -4.264
## intel_o                                                             1.931
## shar_o                                                              9.608
##                                                                   Pr(>|z|)
## (Intercept)                                                        < 2e-16
## 1 + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o | iidTRUE       NA
## attr_o                                                             < 2e-16
## sinc_o                                                              0.5152
## fun_o                                                              < 2e-16
## amb_o                                                             2.01e-05
## intel_o                                                             0.0535
## shar_o                                                             < 2e-16
##
## (Intercept)                                                       ***
## 1 + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o | iidTRUE
## attr_o                                                            ***
## sinc_o
## fun_o                                                             ***
## amb_o                                                             ***
## intel_o                                                           .
## shar_o                                                            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6466.6  on 7030  degrees of freedom
## Residual deviance: 5611.0  on 7024  degrees of freedom
##   (1347 observations deleted due to missingness)
## AIC: 5625
```

```
##
## Number of Fisher Scoring iterations: 5
```

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.

```
anova(model1, model4, model5)
```

```
## Analysis of Deviance Table
##
## Model 1: match ~ attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o
## Model 2: match ~ attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o +
##     factor(iid) - 1
## Model 3: match ~ (1 + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o |
##     iid) + attr_o + sinc_o + fun_o + amb_o + intel_o + shar_o
##   Resid. Df Resid. Dev   Df Deviance
## 1      7024     5611.0
## 2      6474      779.8  550   4831.2
## 3      7024     5611.0 -550  -4831.2
```

The AICs of the three models do not differ much from each other. Model 4 seems to be slightly better than the otehr two.

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.
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).