HW5

Avinash Ramu

October 11, 2016

Q14.5

A

$$Pr(dec = 1) = logit^{-1}(\beta_0 + \beta_1 attr + \beta_2 sinc + \beta_3 intel + \beta_4 fun + \beta_5 amb + \beta_6 shar)$$

```
sd <- read.csv("speed_dating.csv", head = T)
sd[is.na(sd)] <- 0
attach(sd)
fit.1 <- glm (dec ~ attr + sinc + intel + fun + amb + shar, family=binomial(link="logit"))
summary(fit.1)
```

```
##
## Call:
## glm(formula = dec ~ attr + sinc + intel + fun + amb + shar, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5487  -0.8485  -0.3222   0.8750   3.1869
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.071814   0.158020 -32.096  < 2e-16 ***
## attr         0.565207   0.018947  29.831  < 2e-16 ***
## sinc        -0.046328   0.019253  -2.406   0.0161 *
## intel        0.005835   0.021296   0.274   0.7841
## fun          0.225431   0.017645  12.776  < 2e-16 ***
## amb         -0.110656   0.013840  -7.995 1.29e-15 ***
## shar         0.144584   0.011895  12.155  < 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: 11398  on 8377  degrees of freedom
## Residual deviance:  8579  on 8371  degrees of freedom
## AIC: 8593
##
## Number of Fisher Scoring iterations: 5
```

Attr, fun and shar have a positive coefficient and increase the chance of a repeat-date. The coefficients are significant. sinc and amb also have significant coefficients and reduce the chance of a repeat date.

A 1 point increase in attractiveness increases the probability of a repeat date by 14% A 1 point increase in fun increases probability of repeat date by about 5.5% A 1 point increase in shar increases probability of a repeat date by about 3% A 1 point increase in sinc reduces probability of a repeat date by about 1% A 1 point increase in amb reduces probability of a repeat date by about 2.7%

B

Add a explanatory variable `iid`

$$Pr(dec = 1) = logit^{-1}(\beta_0 + \beta_1 attr + \beta_2 sinc + \beta_3 intel + \beta_4 fun + \beta_5 amb + \beta_6 shar + \beta_7 iid)$$

```
fit.2 <- glm (dec ~ attr + sinc + intel + fun + amb + shar + iid, family=binomial(link="logit"))
summary(fit.2)
```

```
##
## Call:
## glm(formula = dec ~ attr + sinc + intel + fun + amb + shar +
##     iid, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5504  -0.8488  -0.3204   0.8742   3.2019
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.1291579  0.1669896 -30.715  < 2e-16 ***
## attr         0.5657412  0.0189604  29.838  < 2e-16 ***
## sinc        -0.0459110  0.0192569  -2.384   0.0171 *
## intel        0.0064327  0.0213075   0.302   0.7627
## fun          0.2247983  0.0176562  12.732  < 2e-16 ***
## amb         -0.1106814  0.0138426  -7.996 1.29e-15 ***
## shar         0.1445464  0.0118950  12.152  < 2e-16 ***
## iid          0.0001811  0.0001682   1.077   0.2814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11398.5  on 8377  degrees of freedom
## Residual deviance:  8577.9  on 8370  degrees of freedom
## AIC: 8593.9
##
## Number of Fisher Scoring iterations: 5
```

The coefficients are similar to the previous model. The rater does not have a significant coefficient.

C

Add a explanatory variable `pid`

$$Pr(dec = 1) = logit^{-1}(\beta_0 + \beta_1 attr + \beta_2 sinc + \beta_3 intel + \beta_4 fun + \beta_5 amb + \beta_6 shar + \beta_7 iid + \beta_8 pid)$$

```
fit.3 <- glm (dec ~ attr + sinc + intel + fun + amb + shar + iid + pid, family=binomial(link="logit"))
summary(fit.3)
```

```
##
## Call:
## glm(formula = dec ~ attr + sinc + intel + fun + amb + shar +
##     iid + pid, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5277  -0.8486  -0.3144   0.8701   3.2397
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.145558   0.167654 -30.691  < 2e-16 ***
## attr         0.560513   0.019008  29.488  < 2e-16 ***
## sinc        -0.048402   0.019296  -2.508   0.0121 *
## intel        0.012965   0.021390   0.606   0.5444
## fun          0.225513   0.017718  12.728  < 2e-16 ***
## amb         -0.106710   0.013879  -7.689 1.49e-14 ***
## shar         0.143274   0.011927  12.013  < 2e-16 ***
## iid          0.006571   0.001383   4.752 2.01e-06 ***
## pid         -0.006422   0.001379  -4.656 3.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11398.5  on 8377  degrees of freedom
## Residual deviance:  8556.1  on 8369  degrees of freedom
## AIC: 8574.1
##
## Number of Fisher Scoring iterations: 5
```

Both the rater and the person being rated now have significant coefficients. This indicates that certain raters have preferences for certain kinds of dates.

Q14.6

A

I used the average of the six predictors as a explanatory variable. I set the NA's to zero.

$$Pr(dec_i = 1) = logit^{-1}(\beta_{i0} + \beta_{i1} newscore)$$

```
sd$newscore <- (sd$attr + sd$sinc +  sd$intel + sd$fun + sd$amb +  sd$shar)/6
sd$iid <- as.factor(sd$iid)
some_iids <- sample(sd$iid, 5)
for (iid1 in some_iids) {
  sd_subset <- sd[sd$iid == iid1, ]
  fit.21 <- glm (sd_subset$dec ~ sd_subset$newscore, family=binomial(link="logit"))
  print(summary(fit.21))
}
```

```
##
## Call:
```

```
## glm(formula = sd_subset$dec ~ sd_subset$newscore, family = binomial(link = "logit"))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.30962  -0.41785  -0.19584   0.06111   2.44387
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -19.134      8.129  -2.354   0.0186 *
## sd_subset$newscore    3.240      1.407   2.303   0.0213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25.782  on 21  degrees of freedom
## Residual deviance: 11.922  on 20  degrees of freedom
## AIC: 15.922
##
## Number of Fisher Scoring iterations: 7
##
##
## Call:
## glm(formula = sd_subset$dec ~ sd_subset$newscore, family = binomial(link = "logit"))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.94994  -0.61411  -0.06046   0.42666   1.87676
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -27.405     13.137  -2.086   0.0370 *
## sd_subset$newscore    3.974      1.930   2.059   0.0395 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25.782  on 21  degrees of freedom
## Residual deviance: 15.201  on 20  degrees of freedom
## AIC: 19.201
##
## Number of Fisher Scoring iterations: 7
##
##
## Call:
## glm(formula = sd_subset$dec ~ sd_subset$newscore, family = binomial(link = "logit"))
##
## Deviance Residuals:
##       1        2        3        4        5        6
## -0.4872  -0.5366   1.5458  -0.8493   1.0950  -1.3527
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)          -9.701     10.133  -0.957    0.338
## sd_subset$newscore    1.237      1.365   0.906    0.365
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7.6382  on 5  degrees of freedom
## Residual deviance: 6.6650  on 4  degrees of freedom
## AIC: 10.665
##
## Number of Fisher Scoring iterations: 4
##
##
## Call:
## glm(formula = sd_subset$dec ~ sd_subset$newscore, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6637  -1.0621  -0.6137   1.1981   1.4669
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.0318     1.4795  -1.373    0.170
## sd_subset$newscore   0.4577     0.3328   1.375    0.169
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 29.065  on 20  degrees of freedom
## Residual deviance: 26.849  on 19  degrees of freedom
## AIC: 30.849
##
## Number of Fisher Scoring iterations: 4

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call:
## glm(formula = sd_subset$dec ~ sd_subset$newscore, family = binomial(link = "logit"))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.177   0.000   0.000   0.000   1.177
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -678.2   335334.1  -0.002    0.998
## sd_subset$newscore    123.3    60969.8   0.002    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 29.0645  on 20  degrees of freedom
## Residual deviance:  2.7726  on 19  degrees of freedom
## AIC: 6.7726
##
## Number of Fisher Scoring iterations: 24
```

The average predictor does not come out to be significant for either of the raters, we appear to lose information by averaging over all the predictors.

B

Vary the intercept and coefficients by group of the rater.

$$Pr(dec = 1) = logit^{-1}(\beta_{j0} + \beta_{j1}attr + \beta_{j2}sinc + \beta_{j3}intel + \beta_{j4}fun + \beta_{j5}amb + \beta_{j6}shar + \beta_{j7}iid + \beta_{j8}pid)$$

```
fit.lmer1 <- lmer(dec ~ 1 + (1 + attr + sinc + intel + fun + amb + shar | iid), family=binomial(link="l
```

```
## Warning in lmer(dec ~ 1 + (1 + attr + sinc + intel + fun + amb + shar | :
## calling lmer with 'family' is deprecated; please use glmer() instead

## Warning in (function (fn, par, lower = rep.int(-Inf, n), upper =
## rep.int(Inf, : failure to converge in 10000 evaluations

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 15.3976 (tol =
## 0.001, component 1)
```

```
display(fit.lmer1)
```

```
## lme4::glmer(formula = dec ~ 1 + (1 + attr + sinc + intel + fun +
##     amb + shar | iid), data = sd, family = binomial(link = "logit"),
##     control = list(optimizer = c("bobyqa", "Nelder_Mead"), calc.derivs = TRUE,
##         use.last.params = FALSE, restart_edge = FALSE, boundary.tol = 1e-05,
##         tolPwrss = 1e-07, compDev = TRUE, nAGQ0initStep = TRUE,
##         checkControl = list(check.nobs.vs.rankZ = "ignore", check.nobs.vs.nlev = "stop",
##             check.nlev.gtreq.5 = "ignore", check.nlev.gtr.1 = "stop",
##             check.nobs.vs.nRE = "stop", check.rankX = "message+drop.cols",
##             check.scaleX = "warning", check.formula.LHS = "stop",
##             check.response.not.const = "stop"), checkConv = list(
##             check.conv.grad = list(action = "warning", tol = 0.001,
##                 relTol = NULL), check.conv.singular = list(action = "ignore",
##                 tol = 1e-04), check.conv.hess = list(action = "warning",
##                 tol = 1e-06)), optCtrl = list()))
## coef.est  coef.se
##    -1.87     0.30
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  iid      (Intercept) 11.66
##           attr         1.16    -0.96
##           sinc         0.16    -0.43  0.37
##           intel        0.21    -0.77  0.69  0.69
##           fun          0.51    -0.90  0.85  0.15  0.59
##           amb          0.08     0.71 -0.71 -0.83 -0.90 -0.45
##           shar         0.40    -0.79  0.67  0.13  0.66  0.83 -0.38
##  Residual              1.00
## ---
## number of obs: 8378, groups: iid, 551
## AIC = 7694.8, DIC = 1217.2
## deviance = 4427.0
```

Attractiveness seems to vary by the rater the most. Other significant predictors that are helpful at the group level are sinc, intel, fun and share. Amb could probably be removed from the grouping based on the SD explained. The raters also have significant intercepts indicating that grouping by rater is a good idea.

C

I have included comments under each model. The no-pooling model cannot be interpreted since the average score does not seem to be a significant predictor. The residual deviance is lowest under the model in 14.6 (B), the multilevel model grouped by the rater. The residual deviance in the model in 14.5 (A) is twice as much as the multilevel model. Grouping by rater seems to give us the best model in this example.

15.1

$$presvote\_intent = \beta_0 + \beta_1 gender + \beta_2 race + \beta_3 age + \beta_4 educ1 + \beta_5 income + \beta_6 religion + \alpha_{j[i]}$$

15.2

A

$$presvote\_intent = \beta_0 + \beta_1 gender + \beta_2 race + \beta_3 age + \beta_4 educ1 + \beta_5 income + \beta_6 religion + \alpha_{j[i]} + \beta_{7j[i]} ideology$$

B

```
nes <- read.dta("nes5200_processed_voters_realideo.dta")
attach(nes)
```

```
## The following objects are masked from sd:
##
##     age, gender, income, race
```

```
nes$presvote_intent <- as.numeric(nes$presvote_intent)
fit.lmer1 <- lmer( presvote_intent ~ 1 + gender + race + age + educ1 + income + religion + (1 + real_id
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge: degenerate Hessian with 1 negative
## eigenvalues
```

```
summary(fit.lmer1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: presvote_intent ~ 1 + gender + race + age + educ1 + income +
##     religion + (1 + real_ideo | state)
##    Data: nes
##
## REML criterion at convergence: 13487
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1033 -0.8214  0.0847  0.5387  4.0550
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state    (Intercept) 0.00000  0.0000
##           real_ideo   0.01651  0.1285    NaN
##  Residual             0.30288  0.5503
## Number of obs: 8023, groups:  state, 48
##
## Fixed effects:
##                                                  Estimate Std. Error
## (Intercept)                                     2.1673951  0.0443371
## gender2. female                                 0.0123553  0.0125204
## race2. black                                   -0.4131829  0.0229532
## race3. asian                                    0.0440938  0.0623164
## race4. native american                         -0.1119647  0.0420101
## race5. hispanic                                -0.1350478  0.0342384
## age                                            -0.0010139  0.0004003
## educ12. high school (12 grades or fewer, incl  -0.0011529  0.0286476
## educ13. some college(13 grades or more,but no  -0.0021672  0.0302734
## educ14. college or advanced degree (no cases   -0.0416001  0.0305119
## income2. 17 to 33 percentile                    0.0145059  0.0254252
## income3. 34 to 67 percentile                    0.0483680  0.0224331
## income4. 68 to 95 percentile                    0.0622777  0.0230662
## income5. 96 to 100 percentile                   0.1599957  0.0319422
## religion2. catholic (roman catholic)           -0.0699287  0.0157068
## religion3. jewish                              -0.2269536  0.0390425
## religion4. other and none (also includes dk pref -0.1333591  0.0222269
##                                                t value
## (Intercept)                                      48.88
## gender2. female                                   0.99
## race2. black                                    -18.00
## race3. asian                                      0.71
## race4. native american                           -2.67
## race5. hispanic                                  -3.94
## age                                              -2.53
## educ12. high school (12 grades or fewer, incl    -0.04
## educ13. some college(13 grades or more,but no    -0.07
## educ14. college or advanced degree (no cases     -1.36
## income2. 17 to 33 percentile                      0.57
## income3. 34 to 67 percentile                      2.16
## income4. 68 to 95 percentile                      2.70
## income5. 96 to 100 percentile                     5.01
## religion2. catholic (roman catholic)             -4.45
## religion3. jewish                                -5.81
## religion4. other and none (also includes dk pref   -6.00

##
## Correlation matrix not shown by default, as p = 17 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```

```
## convergence code: 0
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

The model failed to converge. The real_ideo variable does not seem to show a strong effect of grouping by state. The SD explained by this variable at the group level is a fraction of the residual SD.