Generalized Linear Mixed Model on US smoking teenagers

Timothy Lee

11/04/2020

Context

Data from the 2014 American National Youth Tobacco Survey is available on `http://pbrown.ca/teaching/303/data`, where there is an R version of the 2014 dataset `smoke.RData`, a pdf documentation file `2014-Codebook.pdf`, and the code used to create the R version of the data `smokingData.R`. We aim to investigate the hypothesis that state-level differences in chewing tobacco usage amongst high school students are much larger than differences between schools within a state. In other words, if one was interested in identifying locations with many tobacco chewers (in order to sell chewing tobacco to children, or if one prefer to implement programs to reduce tobacco chewing), would it be important to find individual schools with high chewing rates or would targeting those states where chewing is most common be sufficient?

```
smokeFile = "smokeDownload.RData"
if (!file.exists(smokeFile)) {
  download.file("http://pbrown.ca/teaching/303/data/smoke.RData",smokeFile)
}
(load(smokeFile))
```

```
## [1] "smoke"        "smokeFormats"
```

The smoke object is a `data.frame` containing the data, the `smokeFormats` gives some explanation of the variables. The `colName` and `label` columns of smokeFormats contain variable names in smoke and descriptions respectively. The model and set of results is shown below.

```
smokeFormats[smokeFormats[, "colName"] == "chewing_tobacco_snuff_or",
c("colName", "label")]
```

```
##                       colName
## 151 chewing_tobacco_snuff_or
##                                                                                  label
## 151 RECODE: Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
```

```
# get rid of 9, 10 year olds and missing age and race
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)),]
smokeSub$ageC = smokeSub$Age - 16
library("glmmTMB")
smokeModelT = glmmTMB(chewing_tobacco_snuff_or ~ ageC * Sex +
RuralUrban + Race + (1 | state/school), data = smokeSub,
family = binomial(link = "logit"))

knitr::kable(summary(smokeModelT)$coef$cond, digits = 2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.08	0.17	-17.91	0.00
ageC	0.36	0.03	11.97	0.00

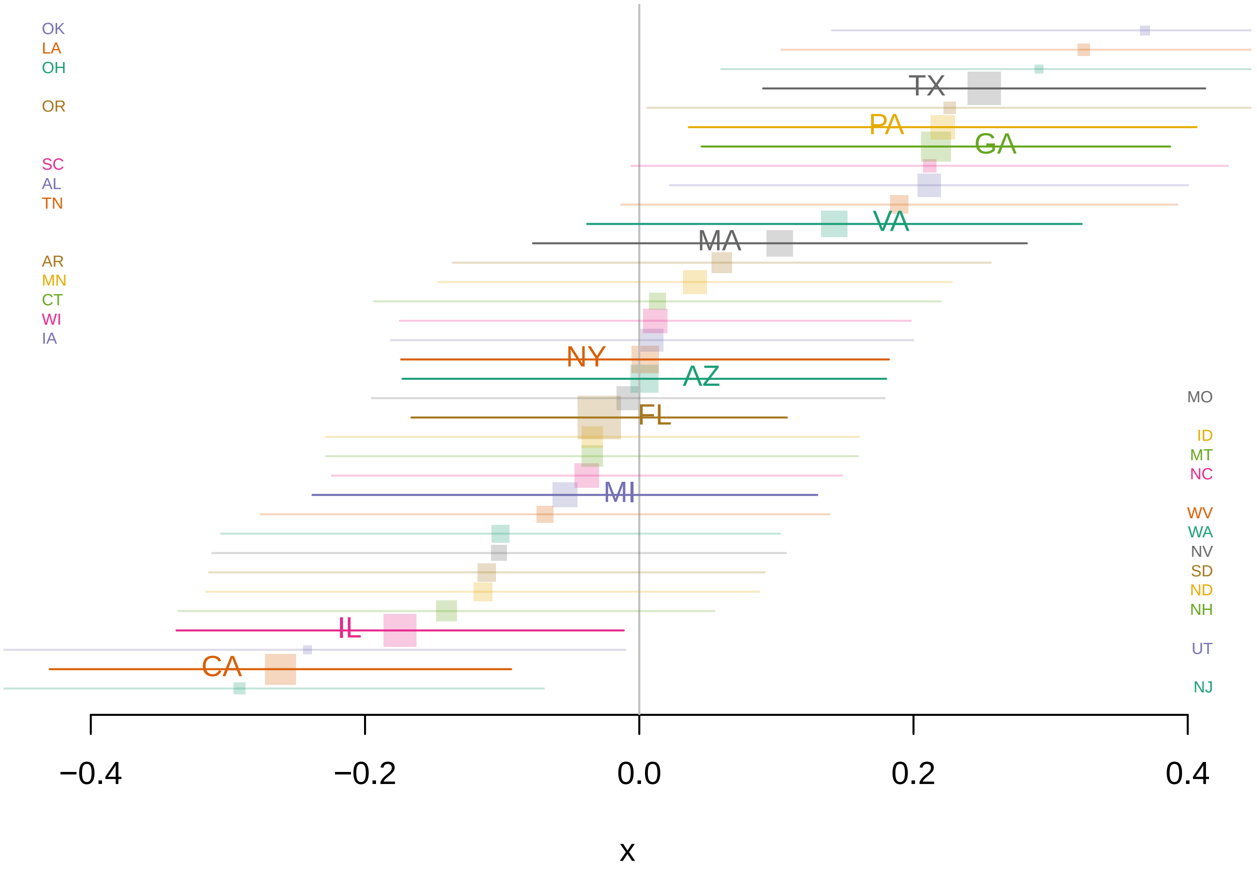
	Estimate	Std. Error	z value	Pr(>|z|)
SexF	-2.04	0.13	-16.21	0.00
RuralUrbanRural	1.00	0.19	5.28	0.00
Raceblack	-1.53	0.19	-8.17	0.00
Racehispanic	-0.51	0.12	-4.29	0.00
Raceasian	-1.12	0.35	-3.16	0.00
Racenative	0.03	0.29	0.10	0.92
Racepacific	1.12	0.39	2.87	0.00
ageC:SexF	-0.33	0.06	-5.91	0.00

```
Pmisc::coefTable(smokeModelT)
```

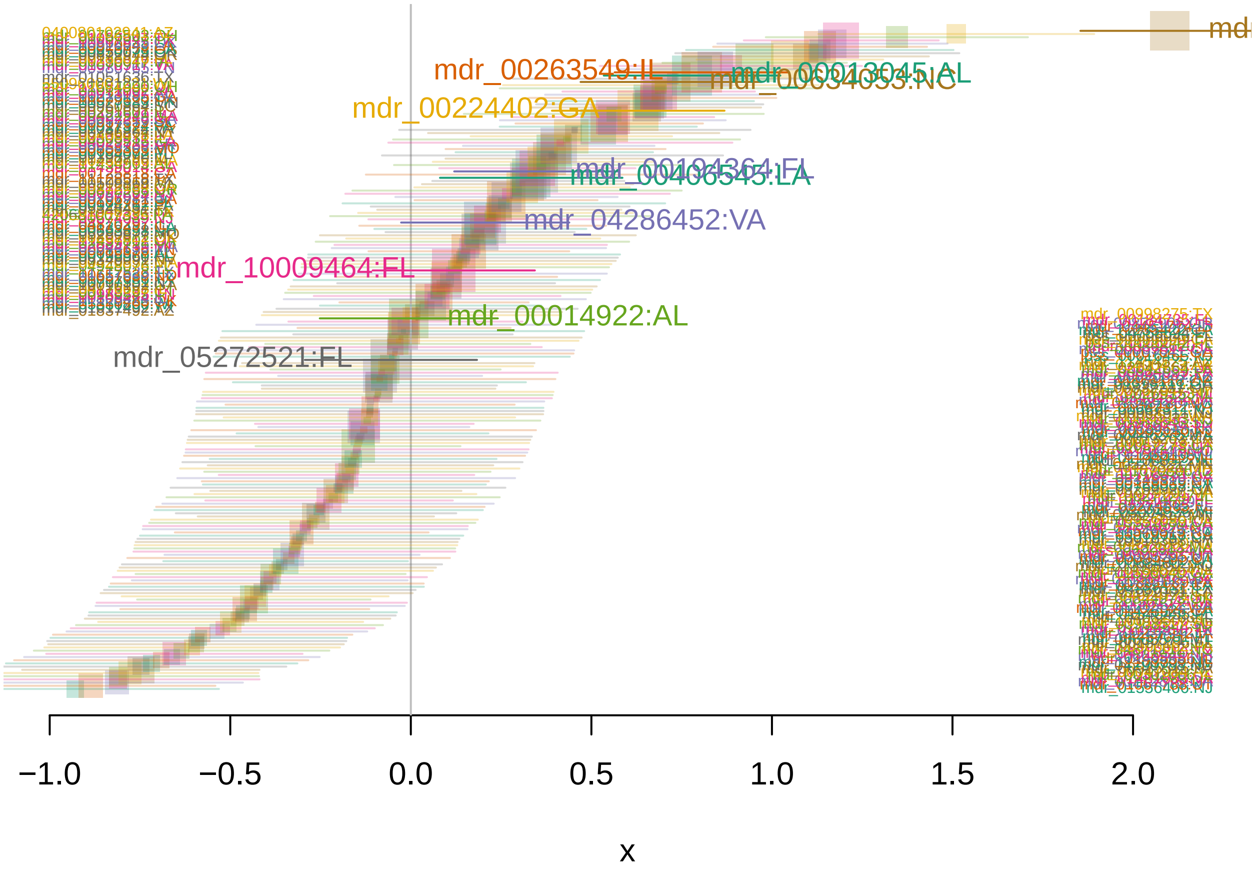
```
## $confint
##                                            2.5 %     97.5 %   Estimate
## cond.(Intercept)                      -3.4115164 -2.7386150 -3.0750657
## cond.ageC                              0.2998825  0.4172636  0.3585731
## cond.SexF                             -2.2857722 -1.7926850 -2.0392286
## cond.RuralUrbanRural                   0.6298634  1.3744609  1.0021622
## cond.Raceblack                        -1.8920774 -1.1598957 -1.5259866
## cond.Racehispanic                     -0.7453189 -0.2775739 -0.5114464
## cond.Raceasian                        -1.8087463 -0.4239125 -1.1163294
## cond.Racenative                       -0.5404718  0.5993118  0.0294200
## cond.Racepacific                       0.3559952  1.8877982  1.1218967
## cond.ageC:SexF                        -0.4374384 -0.2196560 -0.3285472
## school:state.cond.Std.Dev.(Intercept)  0.5935253  0.9486883  0.7503802
## state.cond.Std.Dev.(Intercept)         0.1338326  0.7409208  0.3148958
##
## $tableRaw
## $tableRaw$cond
##                   Estimate Std. Error     z value     Pr(>|z|)
## (Intercept)     -3.0750657 0.17166167 -17.9135249 9.248860e-72
## ageC             0.3585731 0.02994471  11.9745053 4.833118e-33
## SexF            -2.0392286 0.12578986 -16.2113909 4.189782e-59
## RuralUrbanRural  1.0021622 0.18995183   5.2758754 1.321238e-07
## Raceblack       -1.5259866 0.18678449  -8.1697712 3.089750e-16
## Racehispanic    -0.5114464 0.11932490  -4.2861665 1.817828e-05
## Raceasian       -1.1163294 0.35328043  -3.1598959 1.578255e-03
## Racenative       0.0294200 0.29076645   0.1011809 9.194069e-01
## Racepacific      1.1218967 0.39077323   2.8709661 4.092194e-03
## ageC:SexF       -0.3285472 0.05555774  -5.9136166 3.346763e-09
##
## $tableRaw$zi
## NULL
##
## $tableRaw$disp
## NULL
##
##
## $table
##                   variable         level        est      2.5 %     97.5 %
## (Intercept)       ref prob M:Urban:white 0.04414757 0.03193748 0.06073286
## ageC                  ageC               1.43128563 1.34970025 1.51780261
## SexF                   Sex             F 0.13012905 0.10169550 0.16651248
## RuralUrbanRural RuralUrban         Rural 2.72416560 1.87735419 3.95294520
## Raceblack              Race         black 0.21740647 0.15075829 0.31351889
## Racehispanic           Race      hispanic 0.59962765 0.47458292 0.75761958
## Raceasian              Race         asian 0.32747964 0.16385944 0.65448117
## Racenative             Race        native 1.02985704 0.58247340 1.82086518
## Racepacific            Race       pacific 3.07067281 1.42760074 6.60480990
## ageC:SexF          ageC:Sex             F 0.71996894 0.64568831 0.80279489
## school:state.SD          sd  school:state 0.75038024 0.59352531 0.94868828
## state.SD                 sd         state 0.31489584 0.13383264 0.74092084
```

Plots of both state level and school within state level random effects

```
Pmisc::ranefPlot(smokeModelT, grpvar = "state", level = 0.5,
maxNames = 12)
```



```
Pmisc::ranefPlot(smokeModelT, grpvar = "school:state", level = 0.5,
maxNames = 12, xlim = c(-1, 2.2))
```



Statistical Model of `smokeModelT`

$$
\begin{aligned}
Y_{ijk} &\sim Bernoulli(\rho_{ijk}) \\
logit(\rho_{ijk}) &= X_{ij}\beta + A_i + B_{ij} \\
A_i &\sim N(0, \sigma_A^2) \\
B_{ij} &\sim N(0, \sigma_B^2)
\end{aligned}
$$

, where ρ_{ijk} is the predicted (binary) response for the kth American youth from the jth school of the ith state.

Model Choice

The difference between this GLMM and GLM is that this model accounts for the fact that although the individual subjects (American youth) between different schools are considered as independendent, the observations of these subjects within the groups of individual schools within states are dependent (i.e., there is a school-level affect on a student's test scores). GLM doesn't account for this potential dependent relationship (hence there will be no random effect added to the model).

Hence, a `logit` link is more appropriate as we are now working with a bernoulli distribution (of response variables) for each observation, i.e., our responses are now binary responses of the youth having chewed tobacco, snuff, or dip (which is modelled by either 0 or 1, representing yes or no). Hence, the `logit` function captures this notion of our response variable better than a linear mixed model (which only works with a continous response variable).

Conclusions

Based on the summary table and plots, we can say that our data and results are not consistent/supporting the hypothesis that "state-level differences in chewing tobacco usage amongst high school students are much larger than differences between schools within a state", since more of the variation of the response is explained by school differences within states rather than state-level differences. In other words, I think it is more important to find individual schools with high chewing rates rather than targeting those states where chewing is most common.

This conclusion could be seen from the higher standard deviation from table 3 of schools nested within states which has a point estimate of s.d. 0.75 or a point estimate of a variance of 0.5625, while the point estimate of the standard deviation for states effect only is only 0.31 or a point estimate of the variance of 0.0961. This means that more variation is explained by individual schools nested within states. However, we could see that the 2 CI does indeed overlap a little, however both end points or range for the schools nested within states effect are larger than that of just states. Hence, I am not so worried about this overlap (perhaps due to limited sample sizes), but there might still be a (small) chance that there is indeed no difference between state-level and school within state-level differences. In conclusion, if I have to choose one, I will still choose the program that targets individual schools with high chewing rates (within states) as it has a higher point estimate for the variation explained.