# A2 Soln

Xiangyu Kong

27/02/2020

## Question 1

```
school_data = read_csv("school.csv")
```

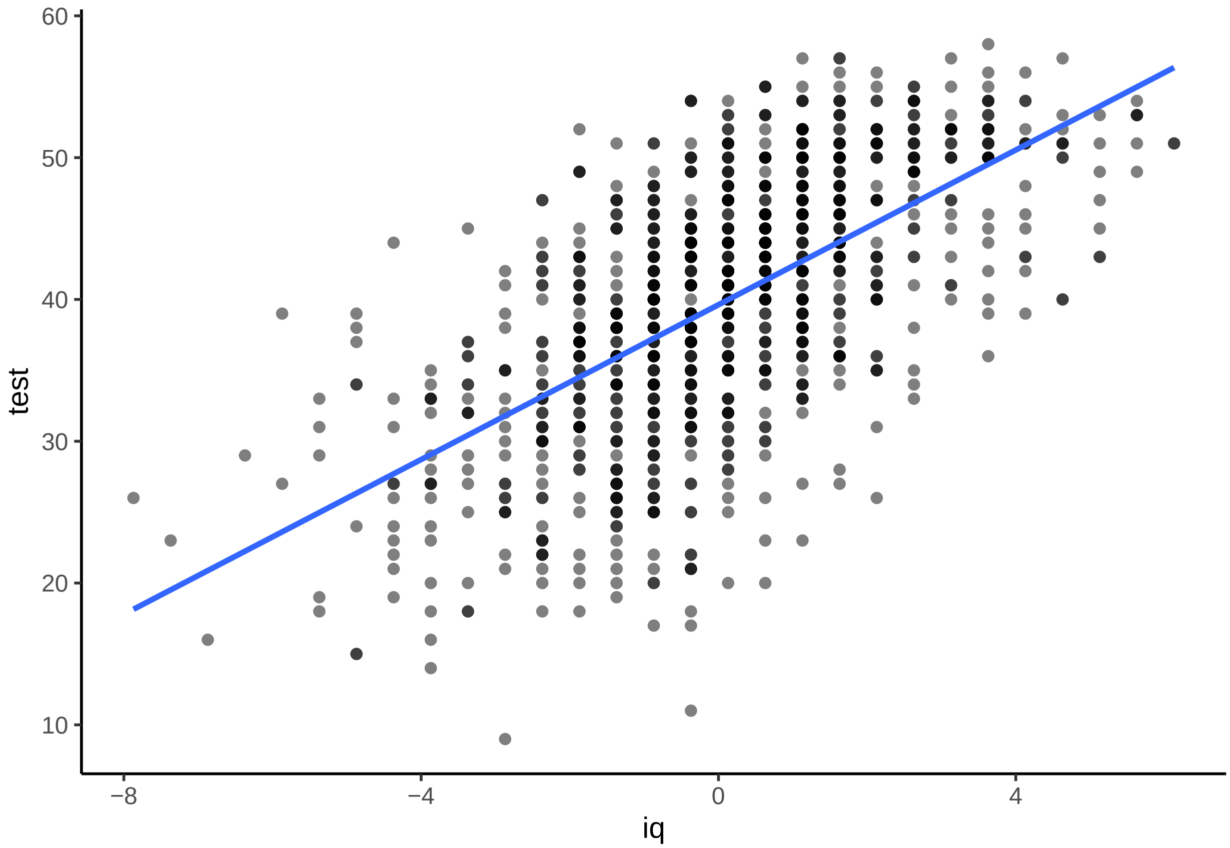
```
## Warning: Missing column names filled in: 'X1' [1]
```

### Question 1.a

The independence assumption may be violated. A school with better teaching resources or better environment may be more likely to have students with better end-of-year language scores. Since there will be multiple observations taken from the same school, they may be dependent and correlated.

### Question 1.b

```
ggplot(school_data, aes(x = iq, y = test)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_classic()
```




Students' iqs and end-of-year language scores are positively related. Students with higher iq tend to achieve a better score in the end-of-year language test.

### Question 1.c

```
school_data = school_data %>%
  group_by(school) %>%
  mutate(mean_ses = mean(ses),
         mean_iq = mean(iq))
```

## Question 1.d

```r
school_lm = lm(test ~ iq + sex + ses + minority_status + mean_ses + mean_iq,
               data = school_data)

summary(school_lm)
```

```
##
## Call:
## lm(formula = test ~ iq + sex + ses + minority_status + mean_ses +
##     mean_iq, data = school_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4126  -4.5967   0.5543   4.9639  18.6042
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     38.45808    0.31251 123.061  < 2e-16 ***
## iq               2.28556    0.11979  19.079  < 2e-16 ***
## sex              2.34325    0.43385   5.401 8.30e-08 ***
## ses              0.19332    0.02641   7.319 5.19e-13 ***
## minority_status -0.17083    0.97592  -0.175    0.861
## mean_ses        -0.21555    0.04641  -4.644 3.88e-06 ***
## mean_iq          1.42674    0.30264   4.714 2.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.818 on 985 degrees of freedom
## Multiple R-squared:  0.4511, Adjusted R-squared:  0.4477
## F-statistic: 134.9 on 6 and 985 DF,  p-value: < 2.2e-16
```

```r
knitr::kable(confint(school_lm), digits = 4)
```

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 37.8448 | 39.0714 |
| iq | 2.0505 | 2.5206 |
| sex | 1.4919 | 3.1946 |
| ses | 0.1415 | 0.2452 |
| minority_status | -2.0860 | 1.7443 |
| mean_ses | -0.3066 | -0.1245 |
| mean_iq | 0.8329 | 2.0206 |

**Estimates:**

- The intercept shows that the average end-of-year language scores for the baseline subgroup is 38.46. This baseline subgroup consists of male, white (non-minority ethnics) students with a verbal IQ score of 0, who live in a family with socioeconomic status of 0, and study in a school with students' mean socioeconomic status of 0 and mean verbal IQ score of 0.
- An increase in student's iq level by 1 tends to make a student's end-of-year language score increase by 2.29.
- A female student tend to have an end-of-year language score 2.34 higher than a male student.
- An increase in student's ses level by 1 tends to make a student's end-of-year language score increase by 0.19.
- A minority student tend to have an end-of-year language score 0.17 lower than a non-minority student.
- An increase in the school's mean ses level by 1 tends to make a student's end of year language score decrease by 0.22.
- An increase in the school's mean iq level by 1 tends to make a student's end of year language score increase by 1.43.

**Confidence Intervals:**

- The 95% confidence interval for the model's intercept is between 37.84 and 39.07.
- For `iq`, `sex`, `ses` and `mean_iq`, the 95% confidence intervals are positive. This indicates that they are likely to have a positive relationship with the student's end-of-year language scores.
- For `mean_ses`, the 95% confidence interval is negative. This means that there is likely to be a negative relationship between `mean_ses` the students' end-of-year language scores.
- The confidence interval for `minority_status` includes 0. It is possible that it does not have a strong effect on the students' test scores.

### Question 1.e

```r
school_lmm <-
  lme4::lmer(test ~ iq + sex + ses + minority_status + mean_ses +
               mean_iq + (1 | school),
             data = school_data)

summary(school_lmm)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: test ~ iq + sex + ses + minority_status + mean_ses + mean_iq +
##     (1 | school)
##    Data: school_data
##
## REML criterion at convergence: 6518.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9926 -0.6304  0.0757  0.6945  2.6361
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept)  8.177   2.859
##  Residual             38.240   6.184
## Number of obs: 992, groups:  school, 58
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)     38.37951    0.48384  79.323
## iq               2.27784    0.10881  20.935
## sex              2.29199    0.40260   5.693
## ses              0.19283    0.02396   8.047
## minority_status -0.65259    0.96943  -0.673
## mean_ses        -0.20131    0.08000  -2.517
## mean_iq          1.62512    0.52017   3.124
##
## Correlation of Fixed Effects:
##             (Intr) iq     sex    ses    mnrty_ men_ss
## iq          -0.035
## sex         -0.408  0.045
## ses          0.013 -0.284 -0.048
## minrty_stts -0.129  0.131  0.001  0.053
## mean_ses    -0.140  0.092  0.003 -0.296  0.039
## mean_iq      0.089 -0.199 -0.007  0.064  0.052 -0.494
```

```r
knitr::kable(confint(school_lmm), digits = 4)
```

| | 2.5 % | 97.5 % |
|---|---|---|
| .sig01 | 2.1819 | 3.5182 |
| .sigma | 5.9011 | 6.4604 |
| (Intercept) | 37.4412 | 39.3176 |
| iq | 2.0649 | 2.4909 |
| sex | 1.5045 | 3.0801 |
| ses | 0.1459 | 0.2398 |
| minority_status | -2.5424 | 1.2493 |
| mean_ses | -0.3564 | -0.0461 |
| mean_iq | 0.6166 | 2.6352 |

**Random Effects:**

- By looking at the Random effects section in the summary, we can see that `school` effects explain about $\frac{8.177}{8.177+38.240} \times 100 = 17.62\%$ of the residual variance in the model.

**Estimates:**

- The intercept shows that the average end-of-year language scores for the baseline subgroup is 38.38. This baseline subgroup consists of male, white (non-minority ethnics) students with a verbal IQ score of 0, who live in a family with socioeconomic status of 0, and study in a school with students' mean socioeconomic status of 0 and mean verbal IQ score of 0.
- An increase in student's iq level by 1 tends to make a student's end-of-year language score increase by 2.28.
- A female student tend to have an end-of-year language score 2.29 higher than a male student.
- An increase in student's ses level by 1 tends to make a student's end-of-year language score increase by 0.19.
- A minority student tend to have an end-of-year language score 0.65 lower than a non-minority student.
- An increase in the school's mean ses level by 1 tends to make a student's end of year language score decrease by 0.20.
- An increase in the school's mean iq level by 1 tends to make a student's end of year language score increase by 1.64.

**Confidence Intervals:** - The first line in the linear mixed model's confidence interval is `.sig01`. This is the confidence interval for the standard deviation for the random effect for school. The standard deviation for random effect for school is not small, so this means that individual schools will have an impact on their students' end-of-year language scores. - The second line in the linear mixed model's confidence interval is `.sigma`. This is the confidence interval for the residuals' standard deviation. This represents the variability of students' end-of-year verbal test scores that is not caused by being in different schools. - The 95% confidence interval for the model's intercept is between 37.44 and 39.32. - For `iq`, `sex`, `ses` and `mean_iq`, the 95% confidence intervals are positive. This indicates that they are likely to have a positive relationship with the student's end-of-year language scores. - For `mean_ses`, the 95% confidence interval is negative. This means that there is likely to be a negative relationship between `mean_ses` the students' end-of-year language scores. - The confidence interval for `minority_status` includes 0. It is possible that it does not have a strong effect on the students' test scores.

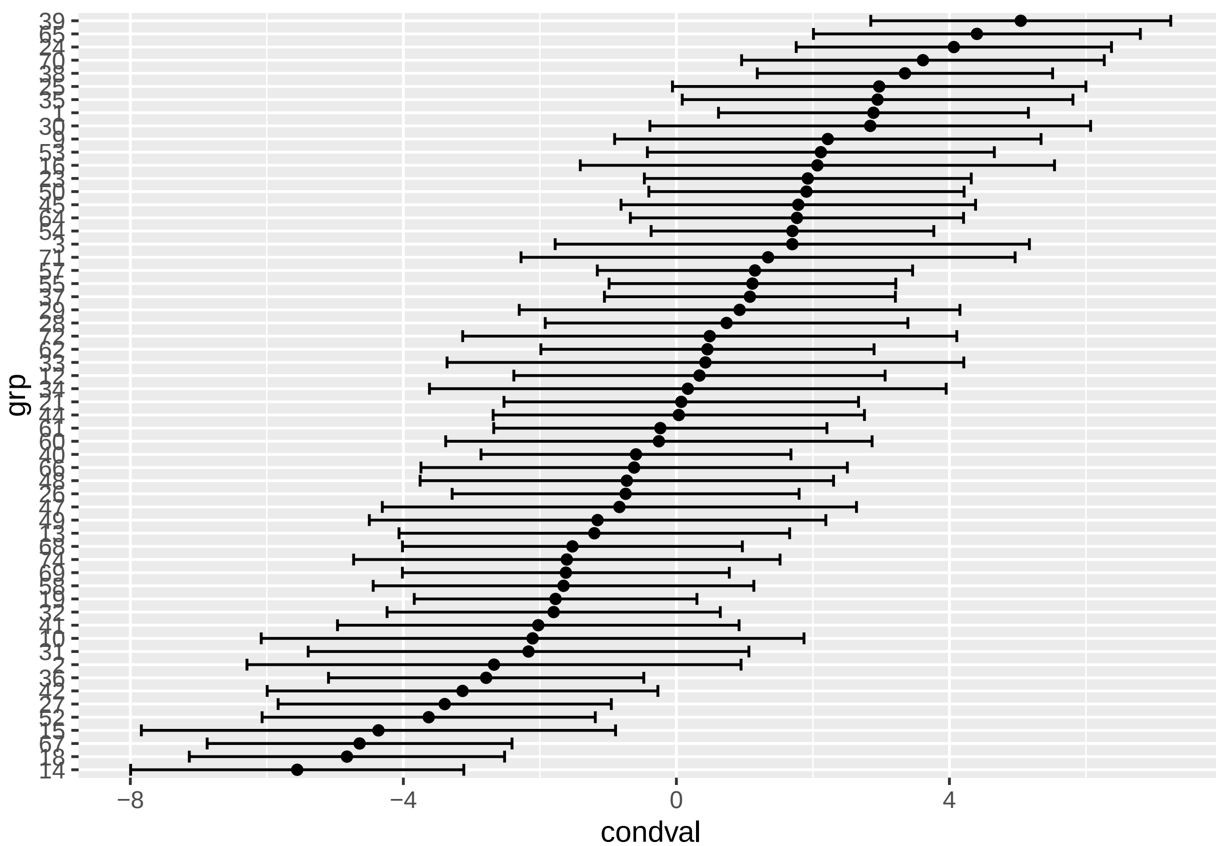
### Question 1.f

The estimated fixed effects in the mixed linear model are very similar to the estimates given by the simple linear regression model. Most estimated values are slightly decreased, meaning they have a weaker effect when considering `school` as a random effect. `minority_status` and `mean_iq` have slightly higher absolute values comparing to the linear regression model, meaning when considering `school` as a random effect, they have more impact on the students' end-of-year language scores.

### Question 1.g

```r
set.seed(1234)

rand_effects <- lme4::ranef(school_lmm, condVar = TRUE)
ranef_df <- as.data.frame(rand_effects)
ranef_df %>%
  ggplot(aes(
    x = grp,
    y = condval,
    ymin = condval - 2 * condsd,
    ymax = condval + 2 * condsd
  )) +
  geom_point() +
  geom_errorbar() +
  coord_flip()
```



The plot shows that for different schools, there conditional mean values differ from the grand. The confidence interval does not completely overlap with each other. The plot forms a visible trend, and this indicates that adding the random effect is appropriate in this case because it explains the variation in mean to a certain degree.

## Question 1.h

It is inappropriate to directly compare students' iq with their end-of-year language score. This is because the schools that the students are studying in also affects the student's language score. This can be visually shown in the conditional mean and confidence interval plot above, and can also be statistically shown after fitting a linear mixed model with `school` as the random effect. The `school` effects explain about $\frac{8.177}{8.177+38.240} \times 100 = 17.62\%$ of the residual variance in the model.

After considering the `school` effect, by looking at the confidence interval for iq, we can see that the 95% confidence interval is greater than 0. This shows that considering the effect explained by students studying in different schools, in general, students with higher iq will achieve a better end-of-year language score.

As for the other factors, by looking at their confidence intervals, female students are more likely to achieve a higher score than male students. Students in families with higher socioeconomic status is more likely to achieve a higher score. Students' minority statuses may have less of an impact because the p-value for the estimate is insignificant, and the confidence intervals include 0. Students in a school with higher mean iq are also likely to achieve a higher score, but students in a school with higher average socioeconomic status is likely to achieve a lower score.

# Question 2

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

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

```r
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
```

```r
# 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
```

```r
library("glmmTMB")

smokeModelT = glmmTMB(
  chewing_tobacco_snuff_or ~ ageC * Sex +
    RuralUrban + Race + (1 | state / school),
  data = smokeSub,
  family = binomial(link = "logit")
)

summary(smokeModelT)
```

```
##  Family: binomial  ( logit )
## Formula:
## chewing_tobacco_snuff_or ~ ageC * Sex + RuralUrban + Race + (1 |
##     state/school)
## Data: smokeSub
##
##      AIC      BIC   logLik deviance df.resid
##   4973.4   5068.5  -2474.7   4949.4    20382
##
## Random effects:
##
## Conditional model:
##  Groups       Name        Variance Std.Dev.
##  school:state (Intercept) 0.56307  0.7504
##  state        (Intercept) 0.09916  0.3149
## Number of obs: 20394, groups:  school:state, 207; state, 35
```

```
##
## Conditional model:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.07507    0.17166 -17.914  < 2e-16 ***
## ageC              0.35857    0.02994  11.975  < 2e-16 ***
## SexF             -2.03923    0.12579 -16.211  < 2e-16 ***
## RuralUrbanRural   1.00216    0.18995   5.276 1.32e-07 ***
## Raceblack        -1.52599    0.18678  -8.170 3.09e-16 ***
## Racehispanic     -0.51145    0.11932  -4.286 1.82e-05 ***
## Raceasian        -1.11633    0.35328  -3.160  0.00158 **
## Racenative        0.02942    0.29077   0.101  0.91941
## Racepacific       1.12190    0.39077   2.871  0.00409 **
## ageC:SexF        -0.32855    0.05556  -5.914 3.35e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

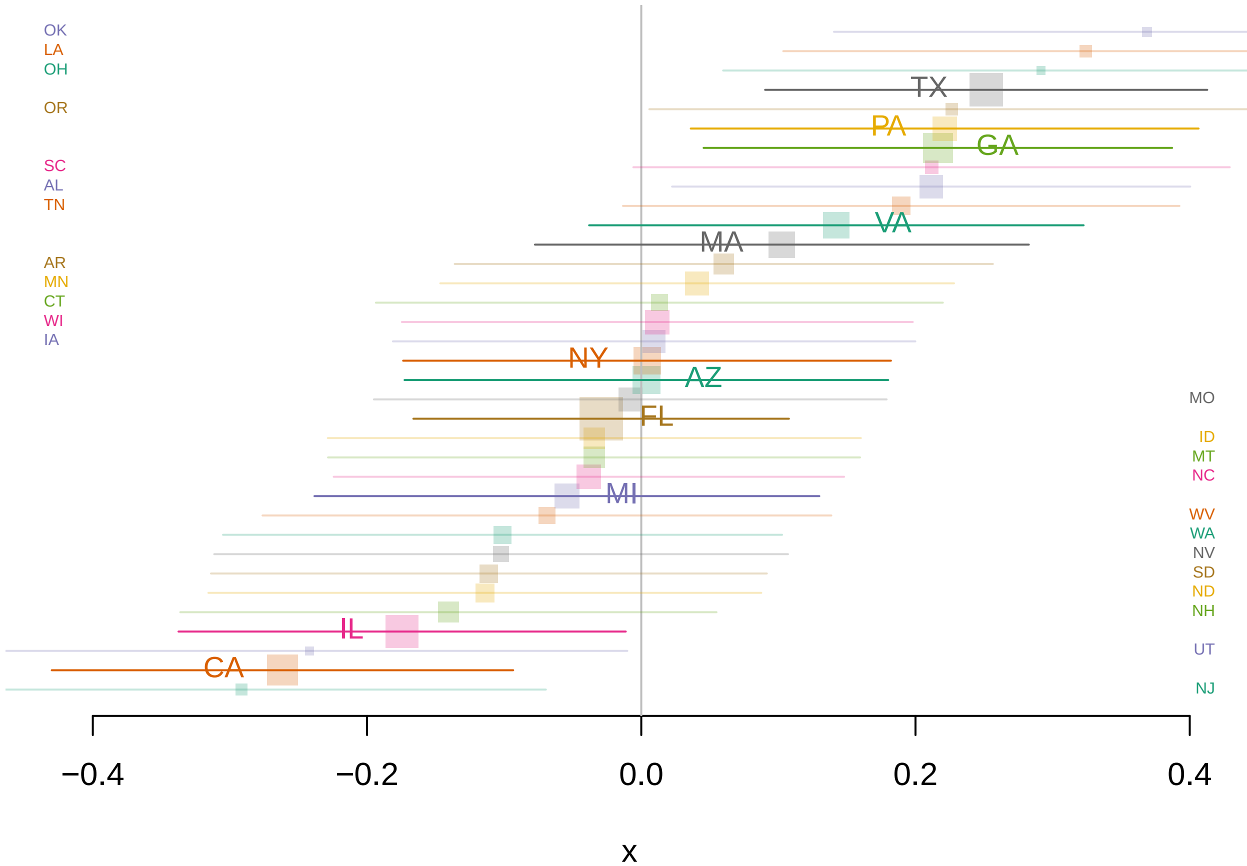
| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -3.08 | 0.17 | -17.91 | 0.00 |
| ageC | 0.36 | 0.03 | 11.97 | 0.00 |
| SexF | -2.04 | 0.13 | -16.21 | 0.00 |
| RuralUrbanRural | 1.00 | 0.19 | 5.28 | 0.00 |
| Raceblack | -1.53 | 0.19 | -8.17 | 0.00 |
| Racehispanic | -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 9.248859e-72
```

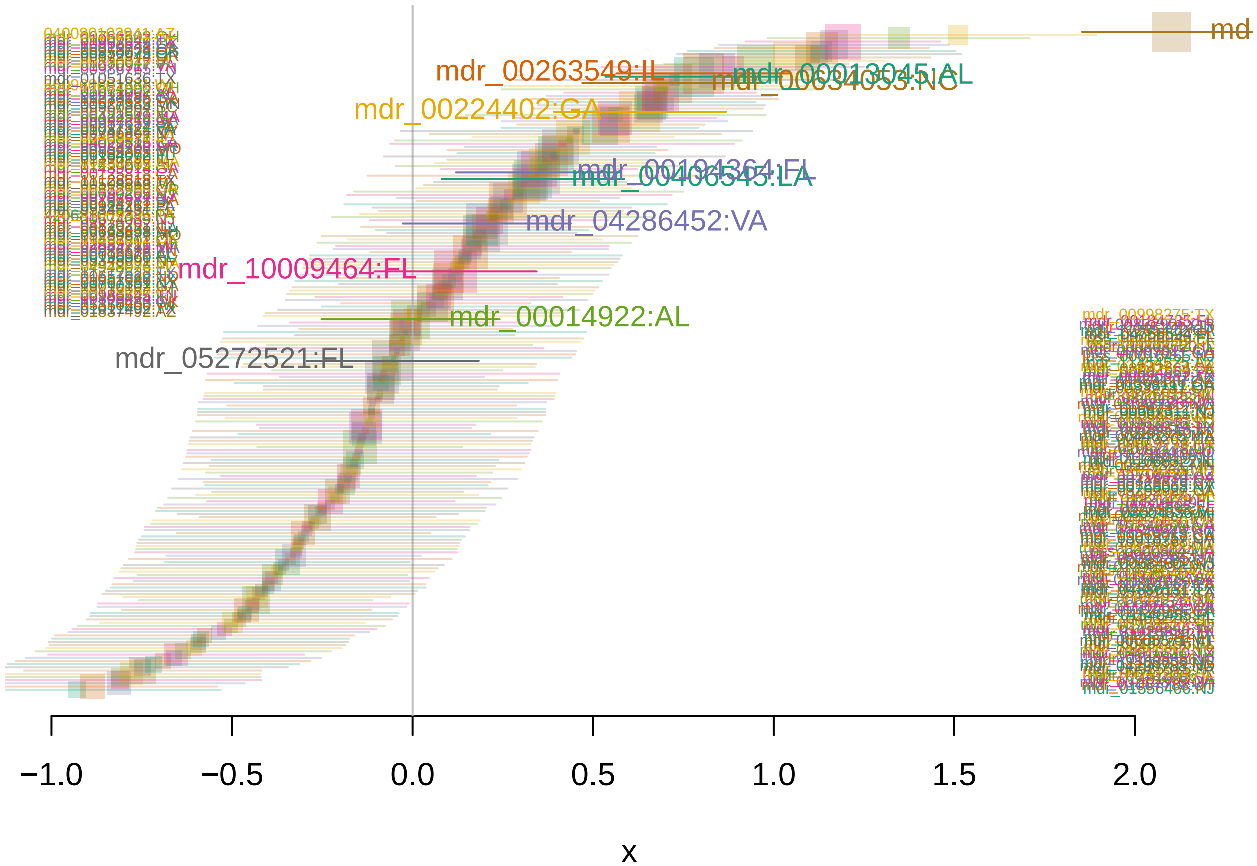
```
## ageC             0.3585731 0.02994471  11.9745053 4.833118e-33
## SexF            -2.0392286 0.12578986 -16.2113909 4.189782e-59
## RuralUrbanRural  1.0021622 0.18995183   5.2758754 1.321238e-07
## Raceblack       -1.5259866 0.18678449  -8.1697712 3.089750e-16
## Racehispanic    -0.5114464 0.11932490  -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
```

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



OK
LA
OH
OR
SC
AL
TN
AR
MN
CT
WI
IA
TX
PA
GA
VA
MA
NY
AZ
FL
MI
IL
CA
MO
ID
MT
NC
WV
WA
NV
SD
ND
NH
UT
NJ
−0.4
−0.2
0.0
0.2
0.4
x

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



## Question 2.a

The model can be represented by

$$
\begin{gathered}
Y_{ij} \mid A, B \sim Binomial(N_{ij}, \mu_{ij}) \\
h(\mu_{ij}) = logit(\mu_{ij}) = \frac{\mu_{ij}}{1-\mu_{ij}} = X_{ij}\beta + A_i + B_{ij} \\
A_i \sim N(0, \sigma_A^2) \\
B_{ij} \sim N(0, \sigma_B^2)
\end{gathered}
$$

where

- $Y_{ij}$ is the number of people who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days, from the $j$th school of the $i$th state.
- $N_{ij}$ is the number of people in the $j$th school of the $i$th state.
- $\mu_{ij}$ is the proportion of the people who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days, from the $j$th school of the $i$th state.
- $A_i$ is the state $i$'s deviation from the population average
- $B_{ij}$ is the $i$th state's $j$'s school's deviation from the population average.
- $X_{ij}$ is the covariate matrix for the $j$th school of the $i$th state.
- $h(\mu_{ij})$ is the logit function.

`smokeModelT` is a Generalized Linear Mixed Model. The difference between `smokeModelT` and the Generalized Linear Model is that `smokeModelT` contains nested random effects from `state` and `school`. It assumes that the state and the school that the students are from affects the number of people who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days.

## Question 2.b

The generalized linear mixed model with a logit link is more appropriate for this dataset than a linear mixed model because the responses are binary, and we are interested in the number of success (number of people who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days) out of a fixed number of trials (the total number of people within the specific subgroup).

## Question 2.c

From the table output using `Pmisc::coefTable(smokeModelT)`, the last two lines of the table represents the standard deviation explained by schools nested within states and states. We can see that one standard deviation difference in school nested in state level variation increases weight by $100 \times [\exp(0.75) - 1] = 111.70\%$, where as one standard deviation difference in state level variation increases weight by $100 \times [\exp(0.31) - 1] = 36.34\%$. This means that the difference between schools within a state is much greater than the state-level differences.

Also, according to the plot generated using `state` as the group variable, there is a trend and very little overlap between the confidence intervals, meaning that using `state` as a random variable is appropriate. However, the plot generated using `school:state` indicates that there is a greater trend than the plot generated using just `state`. The plot is right skewed (there are more schools with value greater than 0), and the confidence intervals also do not overlap.

The two evidence above indicate that differences between schools within a state in chewing tobacco usage amongst high school students are much larger than state-level differences, contradicting to the hypothesis. This suggests that if one was interested in identifying locations with many tobacco chewers, it would be more important to find individual schools with high chewing rates, rather than just targeting those states where chewing is most common.

# Question 3

```
pedestrainFile = Pmisc::downloadIfOld('http://pbrown.ca/teaching/303/data/pedestrians.rds')
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time), ]
pedestrians$y = pedestrians$Casualty_Severity == 'Fatal'
```

```
theGlm = glm(
  y ~ sex + age + Light_Conditions + Weather_Conditions,
  data = pedestrians,
  family = binomial(link = "logit")
)
knitr::kable(summary(theGlm)$coef, digits = 3)
```

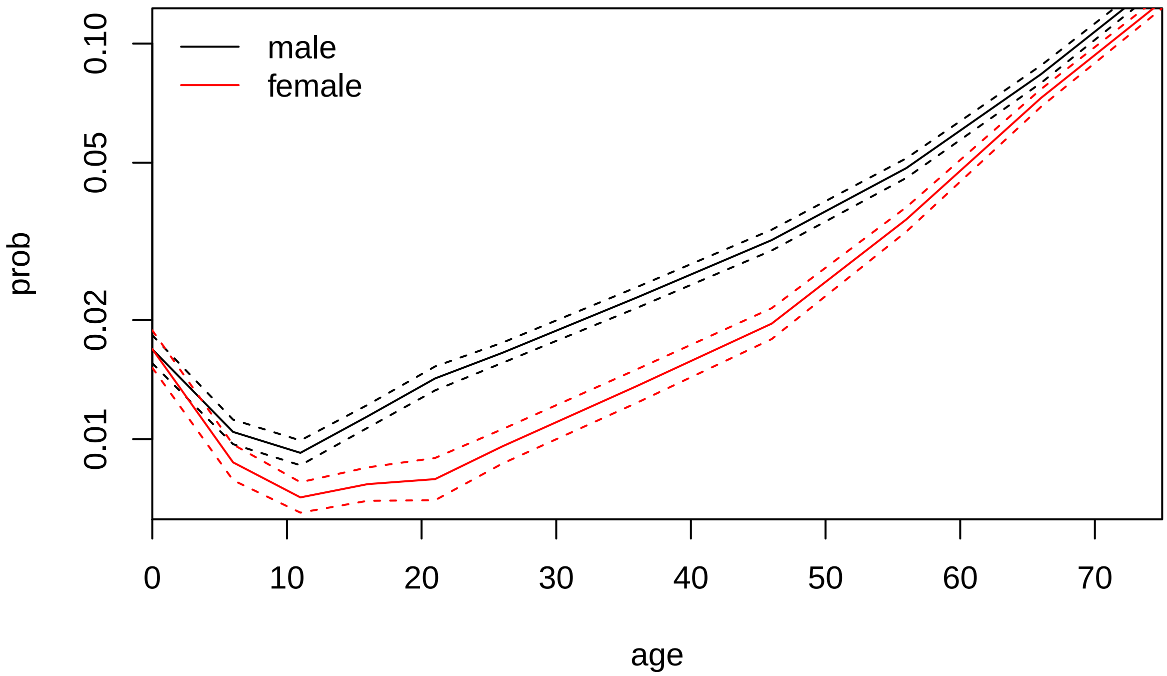
| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.177 | 0.020 | -203.929 | 0.000 |
| sexFemale | -0.275 | 0.011 | -24.665 | 0.000 |
| age0 - 5 | 0.186 | 0.032 | 5.831 | 0.000 |
| age6 - 10 | -0.357 | 0.030 | -12.030 | 0.000 |
| age11 - 15 | -0.504 | 0.029 | -17.668 | 0.000 |
| age16 - 20 | -0.338 | 0.027 | -12.298 | 0.000 |
| age21 - 25 | -0.159 | 0.029 | -5.457 | 0.000 |
| age36 - 45 | 0.324 | 0.027 | 12.213 | 0.000 |
| age46 - 55 | 0.660 | 0.026 | 25.030 | 0.000 |
| age56 - 65 | 1.138 | 0.025 | 45.355 | 0.000 |
| age66 - 75 | 1.760 | 0.023 | 75.234 | 0.000 |
| ageOver 75 | 2.328 | 0.022 | 104.302 | 0.000 |
| Light_ConditionsDarkness - lights lit | 0.995 | 0.012 | 81.220 | 0.000 |
| Light_ConditionsDarkness - lights unlit | 1.176 | 0.052 | 22.415 | 0.000 |
| Light_ConditionsDarkness - no lighting | 2.765 | 0.021 | 131.303 | 0.000 |
| Light_ConditionsDarkness - lighting unknown | 0.259 | 0.068 | 3.788 | 0.000 |
| Weather_ConditionsRaining no high winds | -0.214 | 0.017 | -12.957 | 0.000 |
| Weather_ConditionsSnowing no high winds | -0.751 | 0.092 | -8.136 | 0.000 |
| Weather_ConditionsFine + high winds | 0.175 | 0.037 | 4.774 | 0.000 |
| Weather_ConditionsRaining + high winds | -0.066 | 0.040 | -1.648 | 0.099 |
| Weather_ConditionsSnowing + high winds | -0.550 | 0.172 | -3.193 | 0.001 |
| Weather_ConditionsFog or mist | 0.069 | 0.069 | 0.989 | 0.323 |

```
theGlmInt = glm(
  y ~ sex * age + Light_Conditions + Weather_Conditions,
  data = pedestrians,
  family = binomial(link = "logit")
)
knitr::kable(summary(theGlmInt)$coef, digits = 3)
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.103 | 0.023 | -179.887 | 0.000 |
| sexFemale | -0.545 | 0.044 | -12.425 | 0.000 |
| age0 - 5 | 0.021 | 0.039 | 0.544 | 0.587 |
| age6 - 10 | -0.460 | 0.035 | -13.105 | 0.000 |
| age11 - 15 | -0.582 | 0.035 | -16.625 | 0.000 |
| age16 - 20 | -0.369 | 0.032 | -11.461 | 0.000 |
| age21 - 25 | -0.149 | 0.033 | -4.501 | 0.000 |
| age36 - 45 | 0.322 | 0.031 | 10.508 | 0.000 |
| age46 - 55 | 0.656 | 0.031 | 21.281 | 0.000 |
| age56 - 65 | 1.075 | 0.030 | 35.727 | 0.000 |
| age66 - 75 | 1.622 | 0.029 | 56.315 | 0.000 |
| ageOver 75 | 2.180 | 0.027 | 79.597 | 0.000 |
| Light_ConditionsDarkness - lights lit | 0.990 | 0.012 | 80.676 | 0.000 |
| Light_ConditionsDarkness - lights unlit | 1.174 | 0.052 | 22.399 | 0.000 |
| Light_ConditionsDarkness - no lighting | 2.746 | 0.021 | 130.165 | 0.000 |
| Light_ConditionsDarkness - lighting unknown | 0.257 | 0.068 | 3.759 | 0.000 |
| Weather_ConditionsRaining no high winds | -0.211 | 0.017 | -12.764 | 0.000 |
| Weather_ConditionsSnowing no high winds | -0.746 | 0.092 | -8.075 | 0.000 |
| Weather_ConditionsFine + high winds | 0.176 | 0.037 | 4.803 | 0.000 |
| Weather_ConditionsRaining + high winds | -0.062 | 0.040 | -1.545 | 0.122 |
| Weather_ConditionsSnowing + high winds | -0.548 | 0.172 | -3.189 | 0.001 |
| Weather_ConditionsFog or mist | 0.065 | 0.069 | 0.943 | 0.346 |
| sexFemale:age0 - 5 | 0.546 | 0.068 | 7.970 | 0.000 |
| sexFemale:age6 - 10 | 0.367 | 0.066 | 5.606 | 0.000 |
| sexFemale:age11 - 15 | 0.285 | 0.062 | 4.603 | 0.000 |
| sexFemale:age16 - 20 | 0.150 | 0.062 | 2.408 | 0.016 |
| sexFemale:age21 - 25 | -0.041 | 0.069 | -0.596 | 0.551 |
| sexFemale:age36 - 45 | 0.029 | 0.062 | 0.475 | 0.635 |
| sexFemale:age46 - 55 | 0.059 | 0.060 | 0.976 | 0.329 |
| sexFemale:age56 - 65 | 0.246 | 0.056 | 4.417 | 0.000 |
| sexFemale:age66 - 75 | 0.406 | 0.052 | 7.877 | 0.000 |
| sexFemale:ageOver 75 | 0.411 | 0.049 | 8.348 | 0.000 |

```r
newData = expand.grid(
  age = levels(pedestrians$age),
  sex = c('Male', 'Female'),
  Light_Conditions = levels(pedestrians$Light_Conditions)[1],
  Weather_Conditions = levels(pedestrians$Weather_Conditions)[1]
)
thePred = as.matrix(as.data.frame(
  predict(theGlmInt, newData, se.fit = TRUE)[1:2])) %*% Pmisc::ciMat(0.99)
thePred = as.data.frame(thePred)
thePred$sex = newData$sex
thePred$age = as.numeric(gsub("[[:punct:]].*|[[:alpha:]]", "", newData$age))
toPlot2 = reshape2::melt(thePred, id.vars = c('age', 'sex'))
toPlot3 = reshape2::dcast(toPlot2, age ~ sex + variable)

matplot(
  toPlot3$age,
  exp(toPlot3[, -1]),
  type = 'l',
  log = 'y',
  col = rep(c('black', 'red'), each = 3),
  lty = rep(c(1, 2, 2), 2),
  ylim = c(0.007, 0.11),
  xaxs = 'i',
  xlab = 'age',
  ylab = 'prob'
)
legend(
  'topleft',
  lty = 1,
  col = c('black', 'red'),
  legend = c('male', 'female'),
  bty = 'n'
)
```



male
female
0.10
0.05
0.02
0.01
prob
0
10
20
30
40
50
60
70
age

## Question 3.a

A case-control model that correspond to `theGlm` and `theGlmInt` models can be sampled from a pool of patients who are injured in motor vehicle accidents. The case in this study is the group of people who have been fatally injured in motor vehicle accidents. The control in this study is the group of the people who are slightly injured in motor vehichles accidents, under similar age, sex, lighting and weather conditions. The covariates are the lighting and weather conditions.

## Question 3.b

To address this research question, we need to look at a model that contains intersection between `sex` and `age` in order to compare the probability of male and female getting fatal injuries when they are teenagers and in early adulthood. Thus `theGlmInt` is a better option.

By looking at the intersection estimated log odds for `sexFemale:age16 - 20`, we see that the estimated log odds is 0.150 with a p-value of 0.016. Assuming using a standard significance level of $\alpha = 0.05$, this result is significant. By exponentiating this value, we obtain an odds value of $\exp(0.150) = 1.16$. This indicates that within the age group between 16 to 20-year-olds, the odds of a female getting fatal injury is greater than a male getting fatal injury. This suggests that the hypothesis that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood is false.

By looking at the probability against age graph, within the age range around 15-20 years old, males have a higher probability of getting fatal injuries in motor accidents than female do, and the 99% confidence intervals do not overlap. This supports the hypothesis that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood is false.

## Question 3.c

The control group is not a valid one for assessing whether women are on average better at road safety than man. This is because in the control group, people are still injured and not completely healthy. As stated in the question, men may be less likely than women to report minor injuries caused by road accidents, which leads to skewed data. A valid control group could be sampled from police reports of motor accedents where pedastrians involved were not injured at all. However, this is very difficult to sample, so we can only rely on the given data for now.