303A2

ChenxinranShen

24/02/2020

Question 1

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tid
yverse 1.3.0 ──
```

```
## ✔ ggplot2 3.2.1     ✔ purrr   0.3.3
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   1.0.0     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse
_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
school_data <- read.csv("school.csv")
head(school_data)
```

```
##   X school    ses test    iq sex minority_status denomination
## 1 1      1  -4.73   46  3.13   0               0            1
## 2 2      1 -17.73   45  2.63   0               1            1
## 3 3      1 -12.73   33 -2.37   0               0            1
## 4 4      1  -4.73   46 -0.87   0               0            1
## 5 5      1 -17.73   20 -3.87   0               0            1
## 6 6      1 -17.73   30 -2.37   0               1            1
```

```
glimpse(school_data)
```

```
## Observations: 992
## Variables: 8
## $ X              <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
, 14, 1…
## $ school         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, …
## $ ses            <dbl> -4.73, -17.73, -12.73, -4.73, -17.73, -17
.73, -4…
## $ test           <int> 46, 45, 33, 46, 20, 30, 30, 57, 36, 36, 2
9, 40, …
## $ iq             <dbl> 3.13, 2.63, -2.37, -0.87, -3.87, -2.37, -
2.37, 1…
## $ sex            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, …
## $ minority_status <int> 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1,
1, 0, …
## $ denomination   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, …
```

Question 1a

Why you would have a concern about one of the assumptions of linear regression?

The independence assumption for linear regression is doubted since all observations are from the same city. Students from each school are related to each other.

Question 1b

Create a scatter plot to examine the relationship between verbal IQ scores and end of year language scores.

```
attach(school_data)
plot(iq, test, main = "Scatter plot of verbal IQ scores and language
scores")
abline(lm(test~iq),col = "red")
```

From the scatter plot, we can see there's a positive linear relationship between verbal iq score and language test score. Since there's no otliers, we can say there's a strong relationship.

Question 1c

Create 2 new variables, mean_ses and mean_iq for each school

```
school_data2<-school_data %>%
  group_by(school) %>%
  #summarise(mean_ses = mean(ses),mean_iq = mean(iq))
  mutate(mean_ses = mean(ses),mean_iq = mean(iq))
head(school_data2)
```

```
## # A tibble: 6 x 10
## # Groups:   school [1]
##       X school    ses  test    iq   sex minority_status denominat
ion
##   <int>  <int>  <dbl> <int> <dbl> <int>           <int>        <i
nt>
## 1     1      1      1  -4.73    46  3.13     0               0
1
## 2     2      1      1 -17.7     45  2.63     0               1
1
## 3     3      1      1 -12.7     33 -2.37     0               0
1
## 4     4      1      1  -4.73    46 -0.87     0               0
1
## 5     5      1      1 -17.7     20 -3.87     0               0
1
## 6     6      1      1 -17.7     30 -2.37     0               1
1
## # … with 2 more variables: mean_ses <dbl>, mean_iq <dbl>
```

Question 1d

Fit a linear model, briefly interpret the results.

```
head(school_data2)
```

```
## # A tibble: 6 x 10
## # Groups:   school [1]
##       X school    ses  test    iq   sex minority_status denominat
ion
##   <int>  <int>  <dbl> <int> <dbl> <int>           <int>        <i
nt>
## 1     1      1  -4.73    46  3.13     0               0
1
## 2     2      1 -17.7     45  2.63     0               1
1
## 3     3      1 -12.7     33 -2.37     0               0
1
## 4     4      1  -4.73    46 -0.87     0               0
1
## 5     5      1 -17.7     20 -3.87     0               0
1
## 6     6      1 -17.7     30 -2.37     0               1
1
## # … with 2 more variables: mean_ses <dbl>, mean_iq <dbl>
```

```
school_data2$sex <- as.factor(school_data2$sex)
school_data2$minority_status <- as.factor(school_data2$minority_stat
us)
#school_data2$mean_ses <- as.factor(school_data2$mean_ses)
lm1d<-lm(test~iq+sex+ses+minority_status+mean_ses+mean_iq,data = sch
ool_data2)

summary(lm1d)
```

```
##
## Call:
## lm(formula = test ~ iq + sex + ses + minority_status + mean_ses +
##     mean_iq, data = school_data2)
##
## 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 ***
## sex1              2.34325    0.43385   5.401 8.30e-08 ***
## ses               0.19332    0.02641   7.319 5.19e-13 ***
## minority_status1 -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
```

```
confint(lm1d)
```

```
##                      2.5 %     97.5 %
## (Intercept)      37.8448162 39.0713519
## iq                2.0504849  2.5206429
## sex1              1.4918849  3.1946222
## ses               0.1414857  0.2451566
## minority_status1 -2.0859568  1.7442963
## mean_ses         -0.3066319 -0.1244709
## mean_iq           0.8328516  2.0206247
```

What the intercept means?For which subgroup of students it applies?

For one unit increase of variable in reference group (sex = 0, minority_status = 0, iq = 0, ses = 0, mean_ses = 0, mean_iq = 0), the response variable test score increase by 38.45808.

The location of the confidence intervals for each covariate below 0, include0, or above 0?

below 0: mean_ses

include 0: subgroup of student with minority status (value = 1).

above 0: iq, mean_iq, subgroup of student with sex value = 1.

Question 1e

Fit a linear mixed model with the same fixed effect as 1c, with a random intercept for school https://strengejacke.wordpress.com/2014/10/26/visualizing-generalized-linear-mixed-effects-models-with-ggplot-rstats-lme4/ (https://strengejacke.wordpress.com/2014/10/26/visualizing-generalized-linear-mixed-effects-models-with-ggplot-rstats-lme4/)

```
lmm1e<-lme4::lmer(test~iq+sex+ses+minority_status+mean_ses+mean_iq+(
1|school),data = school_data2)
summary(lmm1e)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: test ~ iq + sex + ses + minority_status + mean_ses + mea
n_iq +
##     (1 | school)
##    Data: school_data2
##
## 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
## sex1              2.29199    0.40260   5.693
## ses               0.19283    0.02396   8.047
## minority_status1 -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     sex1   ses    mnrt_1 men_ss
## iq          -0.035
## sex1        -0.408  0.045
## ses          0.013 -0.284 -0.048
## mnrty_stts1 -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
```

```
confint(lmm1e)
```

```
## Computing profile confidence intervals ...
```

```
##                        2.5 %      97.5 %
## .sig01             2.1818595  3.51821014
## .sigma             5.9011373  6.46042873
## (Intercept)       37.4412106 39.31755070
## iq                 2.0649432  2.49094360
## sex1               1.5044771  3.08014874
## ses                0.1459275  0.23975452
## minority_status1  -2.5423935  1.24925972
## mean_ses          -0.3564217 -0.04606047
## mean_iq            0.6166461  2.63522563
```

Variance of random effect is 8.177, variance of mixed effect is 38.24.The subject effect is explained by 17.62%.

Question 1f

Describe similarities and differences betwen the coefficients of the fixed effect in the results from 1d, 1e and what causes the difference.

The coefficient of 2 models are similar, which means the slopes of 2 models are close. 95% confidence interval of reference group is [37.4412106, 39.31755070] for linear mixed model, which has a larger range than CI of linear model, which is [37.8448162, 39.0713519].

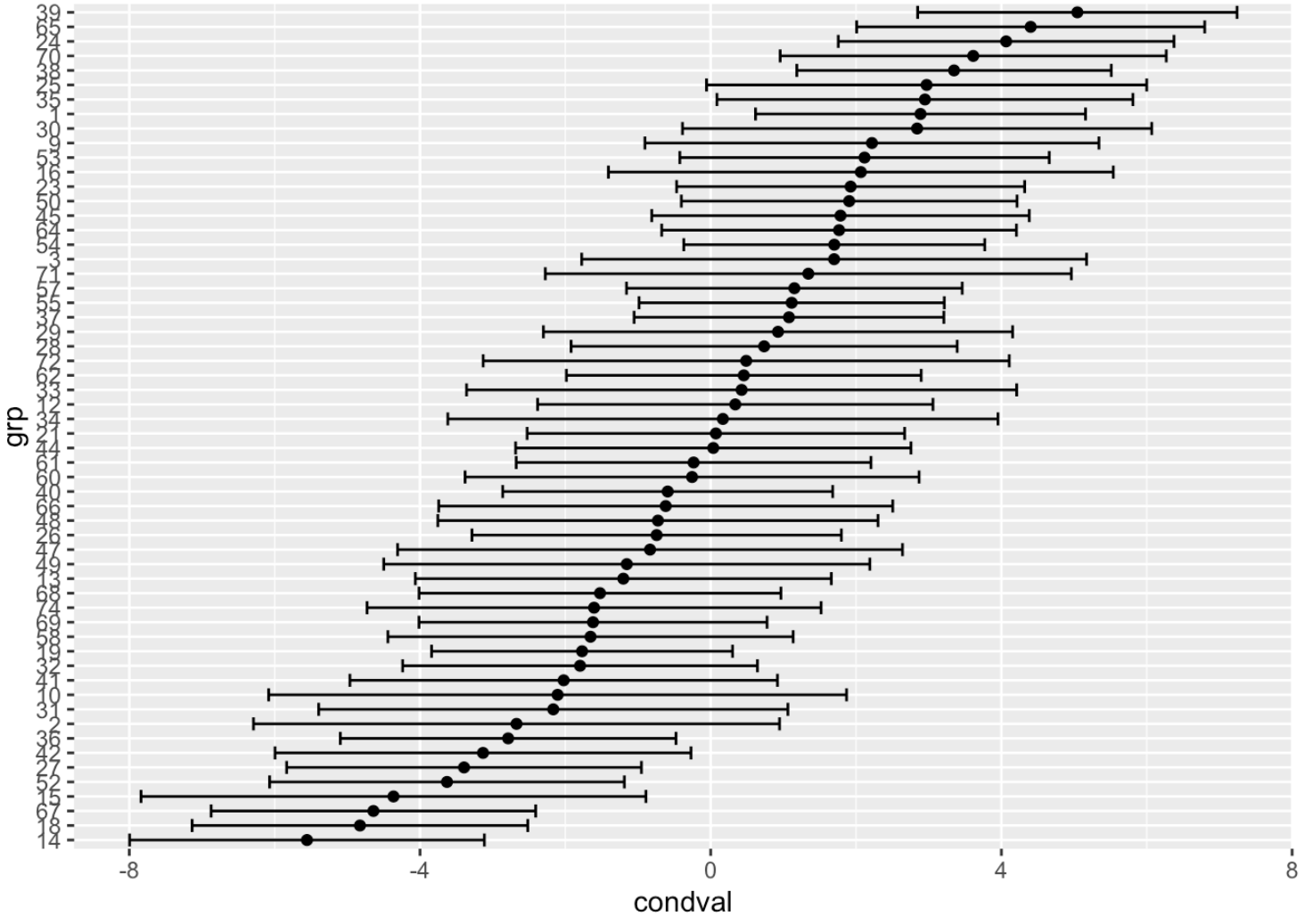
The confidence interval of linear mixed model is wider than linear model, this is because of the random effect of different schools.

Question 1g

Plot the random effects for the different schools. Does it seems reasonable to have included these random effects.

```
random_effects <- lme4::ranef(lmm1e, condVar=TRUE)
ranef_df <- as.data.frame(random_effects)

ranef_df %>%
ggplot(aes(x = grp, y = condval, ymin = condval - 2*condsd, ymax = c
ondval + 2*condsd)) +
geom_point() +
geom_errorbar() +
coord_flip()
```



It is reasonable to include different schools as random effects.

Different schools' intercepts have differences to each other. The highest intercept is around 5 above the average, and the lowest is around -4.8 below the average.

However, since the CI of most of schools have a wide range and almost two third of the schools' CI include 0, which means the differences between most of the individual school are not extreme.

Question 1h

Write a short paragraph summarising. Focus on question of interest: Which variables are associated with Grade 8 students' score on an end of year language test?

The two most significant variables associate with the score is IQ and socioeconomic status, both of them have a very small p-value. p-value< 2e-16 for IQ and p-value=5.19e-13 IQ have a 95%CI [2.0649432, 2.49094360]

The test score will increase by 38.45808 for each unit increasement in reference group. The model is suppose to have different school as random effect, since the proportion of residual variation i 17.62%.

As a conclusion, the language scores of students from same school are strongly related, and the individual IQ, status of their family are key variables associated with their test score.

Question 2

```
smokeFile <-'smoke.RData'
#if(!file.exists(smokeFile)){
 # download.file('http://pbrown.ca/teaching/303/data/smoke.RData', s
mokeFile)}
#(load(smokeFile))
load(smokeFile)
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
```

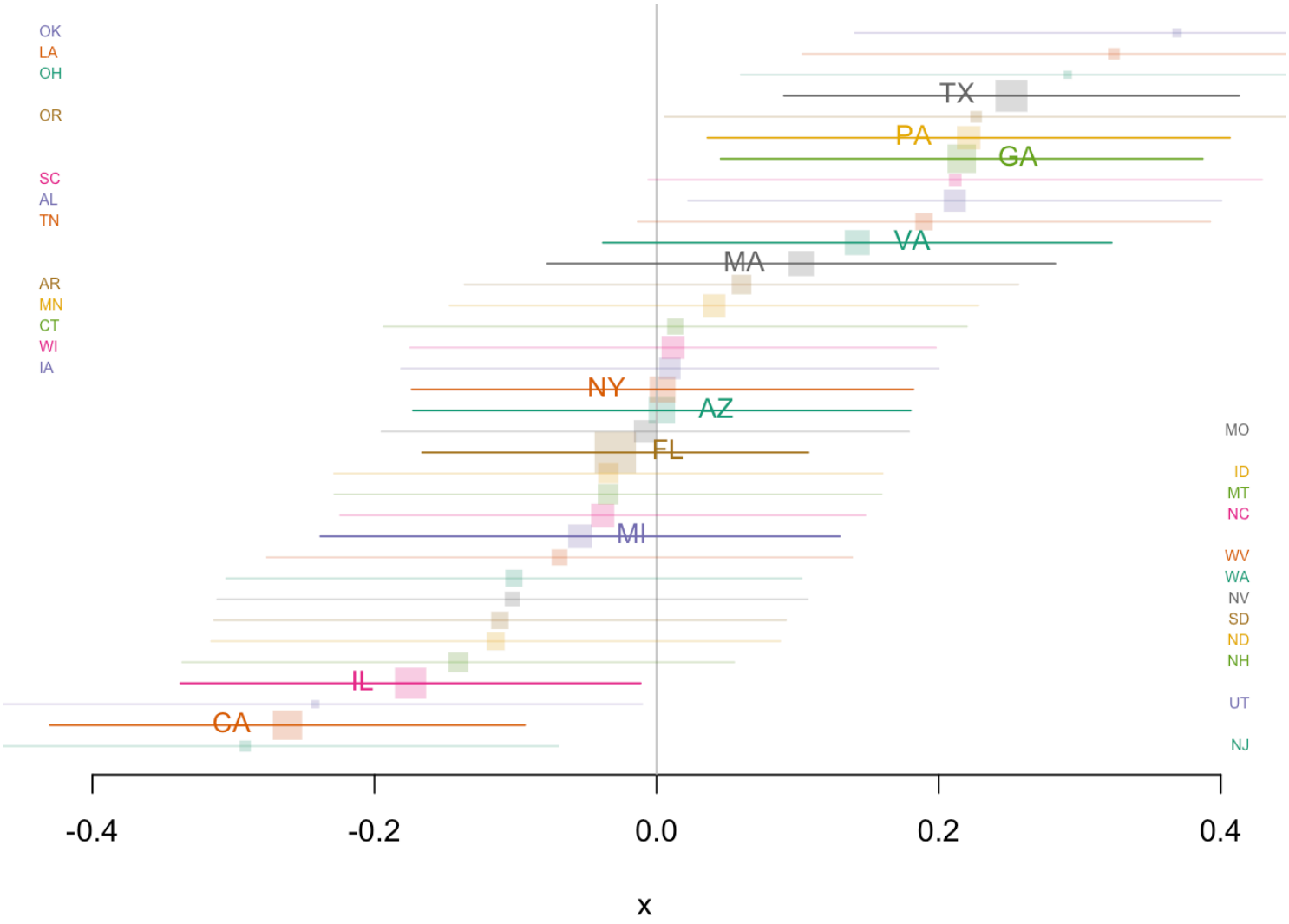
```
# get rid of 9, 10 year olds and missing age and race
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)),]
smokeSub$ageC = smokeSub$Age - 16
library("glmmTMB")
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsist
ency detected.
## TMB was built with Matrix version 1.2.18
## Current Matrix version is 1.2.17
## Please re-install 'TMB' from source using install.packages('TMB',
type = 'source') or ask CRAN for a binary version of 'TMB' matching
CRAN's 'Matrix' package
```

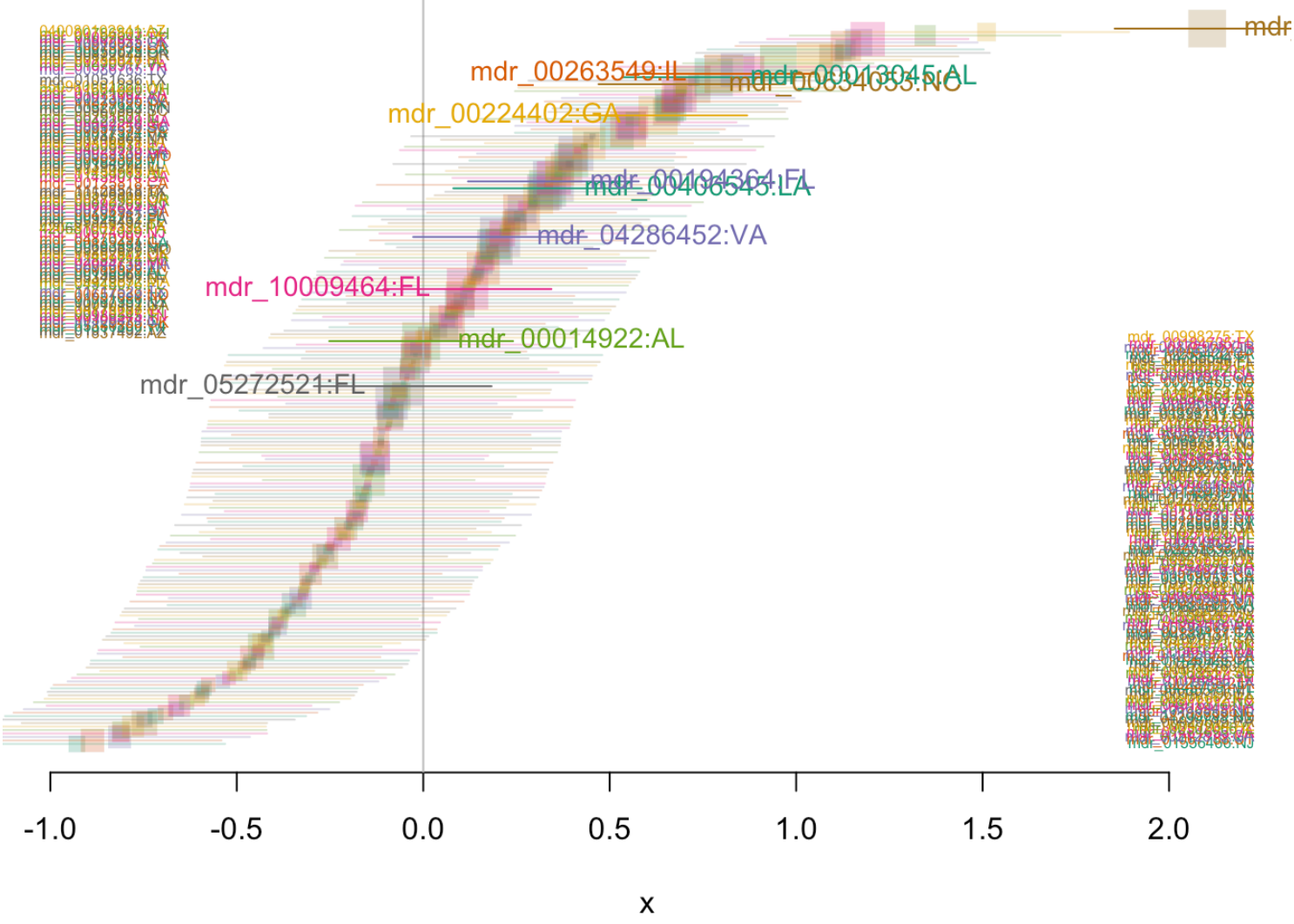
```
smokeModelT = glmmTMB(chewing_tobacco_snuff_or ~ ageC * Sex + RuralU
rban + Race + (1 | state/school), data = smokeSub, family = binomial
(link = "logit"))
knitr::kable(summary(smokeModelT)$coef$cond, digits = 2)
```

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




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



Question 2a

Write down a statistical model corresponding to smokeModelT. Briefly explain the difference between this model and a generalized linear model.

$$Y_{ijk} \sim Gamma(\alpha, \beta)$$

$$logit(p_{ijk}) = X_{ij}\beta + A_i + B_{ij}$$

$$A_{ij} \sim N(0, \sigma_A^2)$$

$$B_{ij} \sim N(0, \sigma_B^2)$$

The generalized linear model assume that all observations are independent to each other, which glmm doesn't.

Question 2b

Briefly explain why this generalized linear mixed model with a logit link is more appropriate for this dataset than a linear mixed model.

Because the response variable isn't noramlly distributed. By using a link function, we can make it become normal.

Question 2c

Q. Write a paragraph assessing the hypothesis that [state-level differences in chewing tabacoo usage amongst high school students], are much larger than [difference between schools within a state]

The hypothesis doesn't hold. Since the estimates of standard diviation for student smoke between states is 0.31. It's much lower than the CI for student smoke between schools in a state, which is 0.75.

Q. If one was interested in identifying locations with many tobacco chewers (in order to sell chewing tobacco to children, or if you 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?

It's more important to find individual schools with high chewing rate than finding state, since there's higher standard deviation between different schools in a state.

There's higher profit and efficiency to tartget school with higher chewing rate. The top rated school has >1.0 deriviation above the average, which the top rated states has about 0.2 deviation above the average.

The CIs of interceptions on the plot of schools in one state are narrower, which means there is higher accurecy for finding customers in the target schools than in the targed state.

Question 3

```
pedestrainFile = Pmisc::downloadIfOld('http://pbrown.ca/teaching/303
/data/pedestrians.rds')
```

```
## Loading required namespace: R.utils
```

```
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time), ]
pedestrians$y = pedestrians$Casualty_Severity == 'Fatal'
dim(pedestrians)
```

```
## [1] 1159371       7
```

```
pedestrians[1:3, ]
```

```
##                   time     age  sex Casualty_Severity
## 54 1979-01-01 22:40:00 26 - 35 Male            Slight
## 65 1979-01-02 10:40:00 26 - 35 Male            Slight
## 79 1979-01-02 14:25:00 46 - 55 Male            Slight
##         Light_Conditions    Weather_Conditions     y
## 54 Darkness - lights lit Snowing no high winds FALSE
## 65              Daylight Raining no high winds FALSE
## 79              Daylight Raining no high winds FALSE
```

```
table(pedestrians$Casualty_Severity, pedestrians$sex)
```

```
##
##             Male Female
##   Slight 637919 481811
##   Fatal   24429  15212
```

```
range(pedestrians$time)
```

```
## [1] "1979-01-01 01:00:00 EST" "2015-12-31 23:35:00 EST"
```

```
theGlm = glm(y ~ sex + age + Light_Conditions + Weather_Conditions,d
ata = 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

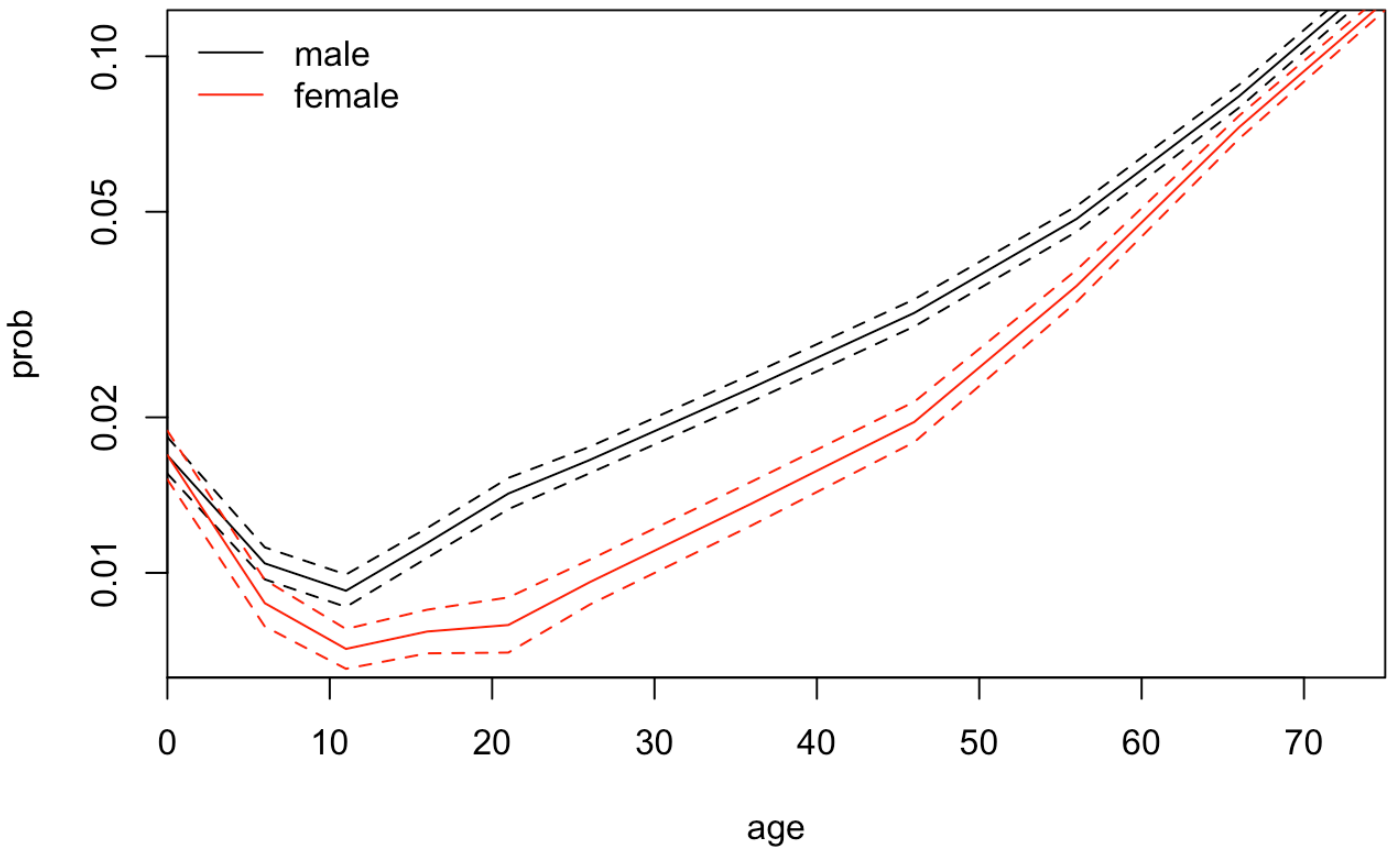
```
theGlmInt = glm(y ~ sex * age + Light_Conditions + Weather_Condition
s,data = pedestrians, family = binomial(link = "logit"))
knitr::kable(summary(theGlmInt)$coef, digits = 3)
```

	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

```
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:]]", "", newDa
ta$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), x
axs='i',xlab= 'age', ylab='prob')
legend('topleft', lty=1, col=c('black','red'), legend = c('male','fe
male'), bty='n')
```



Question 3a

Write a short paragraph describing a case/control model (not the results) corresponding the theGlm and theGlmInt objects. Be sure to specify the case definition and the control group,and what the covariates are.

The model is about if the patient has a fatal or a slight injury. case: fatal injury control: slight injury The covariates are gender, age, light conditions and weather conditions. The only difference between two models is, there is a interaction of female and age in theGlmint model.

Question 3b

Write a short report assessing whether the UK road accident data are consistent with the hypothesis that [women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood]. Explain which of the two models fit is more appropriate for addressing this research question

In theGLMInt model, the estimate of Female is 0.58 with CI[0.53, 0.63], which is much lower than male. It's reasonable to say that women tend to be safer than men.

However, teenage female (age 11-15, 16-20) tend to be in more danger than adult female. The estimates of teenage female around age 11-15 and 16-20 are are 1.33 and 1.16 respectively.

Therefore, the second model is better. As we are exploring the relationship between the fitted variable and gender and age at the same time, we need to study on the intraction of this two covariates.

Question 3c

It is well established that [women are generally more willing to seek medical attention for health problems than men], and it is hypothesized that [men are less likely than women to report minor injuries caused by road accidents]. Write a critical assessment of [whether or not the is a valid one for assessing whether women are on average better at road safety than man].

From figue2, the probability of female being a fatal case in baseline conditons is lower than the one of male. It is reasonable to say that female are more likely to be slight injuried than male . Since women more willing to seek medical attention and report accidents, female can be over represented in control group. By in overall, the case/control model is valid.