STA303 - ASSIGNMENT 2

Manyi Luo - 1003799419

3/6/2020

```
## ── Attaching packages ─────────────────────────────────────────────── tidyver
se 1.3.0 ──
```

```
## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0
```

```
## ── Conflicts ─────────────────────────────────────────────────── tidyverse_con
flicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Question 1

Question 1a

Briefly describe why, without even looking at these data, you would have a concern about one of the assumptions of linear regression.

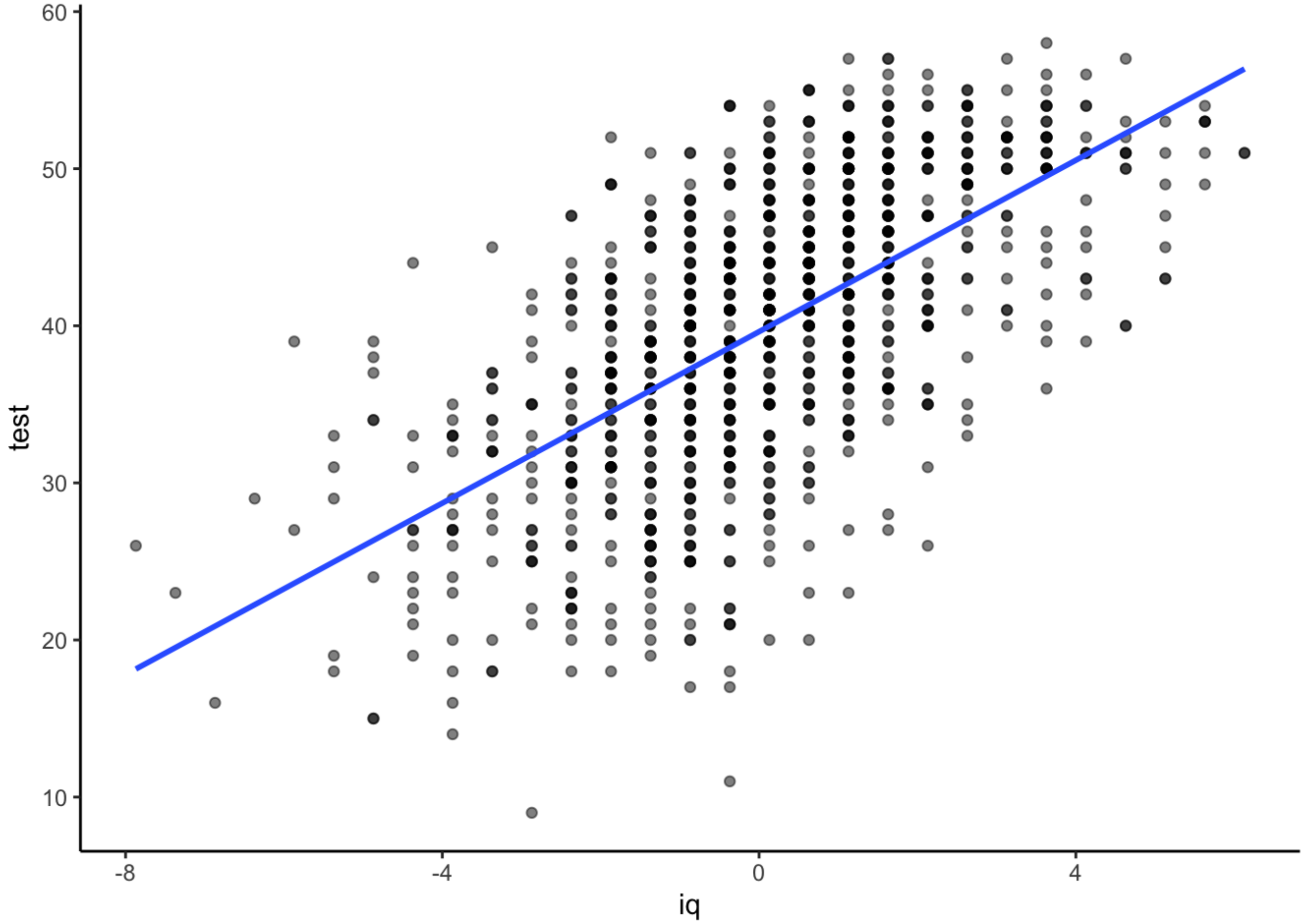
A concern toward the assumption of linear regression might be: each individual student's score on the end-of-year language test might be affected by different schools (for example: private school and public schools etc.), so students' scores within a particular school might be correlated.

Question 1b

Create a scatter plot to examine the relationship between verbal IQ scores and end-of-year language scores. Include a line of best fit. Briefly describe what you see in the plot in the context of the question of interest.

```
schooldata <- read.csv("/Users/mandy/Desktop/school.csv")

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



From the blue fitted line presented above, we can see that there is an approximate positive linear relationship between verbal IQ scores and end-of-year language scores, meaning that student with higher verbal IQ scores might have higher end-of-year language scores.

Also we can see that the variance of data points are not constant, as individual black dots are not randomly distributed visually. Points located in the middle have greater varibility, so linear model cannot be applied in this case.

Question 1c

Create two new variables in the data set, mean_ses that is the mean of ses for each school, and mean_iq that is mean of iq for each school.

```
school1 <- schooldata %>%
    group_by(school) %>%
    mutate(mean_ses = mean(ses), mean_iq = mean(iq))
school1
```

```
## # A tibble: 992 x 10
## # Groups:   school [58]
##        X school    ses  test    iq   sex minority_status denomination mean_ses
##    <int>  <int>  <dbl> <int> <dbl> <int>           <int>        <int>    <dbl>
##  1     1      1  -4.73    46  3.13     0               0            1    -13.9
##  2     2      1 -17.7     45  2.63     0               1            1    -13.9
##  3     3      1 -12.7     33 -2.37     0               0            1    -13.9
##  4     4      1  -4.73    46 -0.87     0               0            1    -13.9
##  5     5      1 -17.7     20 -3.87     0               0            1    -13.9
##  6     6      1 -17.7     30 -2.37     0               1            1    -13.9
##  7     7      1  -4.73    30 -2.37     0               1            1    -13.9
##  8     8      1 -17.7     57  1.13     0               0            1    -13.9
##  9     9      1 -14.7     36 -2.37     0               1            1    -13.9
## 10    10      1 -12.7     36 -0.87     0               1            1    -13.9
## # … with 982 more rows, and 1 more variable: mean_iq <dbl>
```

Question 1d

Fit a linear model with test as the response and use iq, sex, ses, minority_status, mean_ses and mean_iq as the covariates. Show the code for the model you fit and the results of running summary() and confint() on the model you fit and briefly interpret the results. (A complete interpretation here should discuss what the intercept means, and for which subgroup of students it applies, as well as the location of the confidence intervals for each covariate, i.e. below 0, includes 0 or above zero. Address the question of interest.)

```
line1 <- lm(test ~ iq + sex + ses + minority_status + mean_ses + mean_iq, data=school1)
summary(line1)
```

```
## 
## Call:
## lm(formula = test ~ iq + sex + ses + minority_status + mean_ses + 
##     mean_iq, data = school1)
## 
## 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
```

```
confint(line1)
```

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

The intercept means the baseline score of a student's end-of-year language test. This indicates the subgroup of students who have zero verbal IQ score (at average level), are male, have socioeconomic status of zero (at average level), and not belong to ethnic minority groups.

We can see that only the confidence interval of minority_status's coefficient (-2.0859568, 1.7442963) includes zero, meaning that the association between minority_status and student's score on an end-of-year language test is not significant.

The confidence interval of iq's coefficient (2.0504849, 2.5206429), sex's coefficient (1.4918849, 3.1946222), ses's coefficient (0.1414857, 0.2451566) and mean_iq's coefficient (0.8328516, 2.0206247) are above zero, meaning that iq, sex, ses and mean_iq have a significant positive effect toward student's score on an end-of-year language test.

Vice versa, the confidence interval of mean_ses's coefficient (-0.3066319, -0.1244709) is below zero, meaning that mean_ses has a significant negative effect toward student's score on an end-of-year language test.

Note: all confidence intervals mentioned in this question are 95%.

Question 1e

Fit a linear mixed model with the same fixed effects as 1c and with a random intercept for school. Show the code for the model you fit and the results of running summary() and confint() on the model you fit and briefly interpret the results. (Hint 1: Consider the estimated standard deviations in the summary to make sure you understand the first two rows of the confint output. Hint 2: If you want to suppress the 'Computing profile confidence intervals …' message you can use message=FALSE in the chunk.)

```
line2 <- lme4::lmer(test ~ iq + sex + ses + minority_status + mean_ses + mean_iq + (1|sc
hool), data = school1)
summary(line2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: test ~ iq + sex + ses + minority_status + mean_ses + mean_iq +
##     (1 | school)
##    Data: school1
##
## 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
```

```
confint(line2)
```

```
## 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
## sex              1.5044771  3.08014874
## ses              0.1459275  0.23975452
## minority_status -2.5423935  1.24925972
## mean_ses        -0.3564217 -0.04606047
## mean_iq          0.6166461  2.63522563
```

The estimated standard deviations in the summary can be used to compute confidence interval using point estimate. In general:

$$CI = (\hat{beta} - 1.96 * se(\hat{beta}), \hat{beta} + 1.96 * se(\hat{beta}))$$

Intercepts in this case still represent the baseline scores of student's end-of-year language test, but are classified into 58 different schools (random intercept effects).

Still, we can see that only the confidence interval of minority_status's coefficient (-2.5423935, 1.24925972) includes zero, meaning that there might not be significant association between minority_status and student's score on an end-of-year language test.

The confidence interval of iq's coefficient (2.0649432, 2.49094360), sex's coefficient (1.5044771, 3.08014874), ses's coefficient (0.1459275, 0.23975452) and mean_iq's coefficient (0.6166461, 2.63522563) are above zero, meaning that iq, sex, ses and mean_iq have a significant positive effect toward student's score on an end-of-year language test.

Vice versa, the confidence interval of mean_ses's coefficient (-0.3564217, -0.04606047) is below zero, meaning that mean_ses has a significant negative effect toward student's score on an end-of-year language test.

Also according to Hint 1, we can see that the confidence interval for random intercept's coefficient (2.1818595, 3.51821014) and residual's coefficient (5.9011373, 6.46042873) both have a significant positive effect toward student's score on an end-of-year language test. The standard deviation of random intercept is 2.859 and the standard deviation of residual is 6.184 (porportion of random effects variance among total variance is calculated and explained in 1g), which can be used to calculate confidence intervals using the formula above.

Note: all confidence intervals mentioned in this question are 95%.

Question 1f

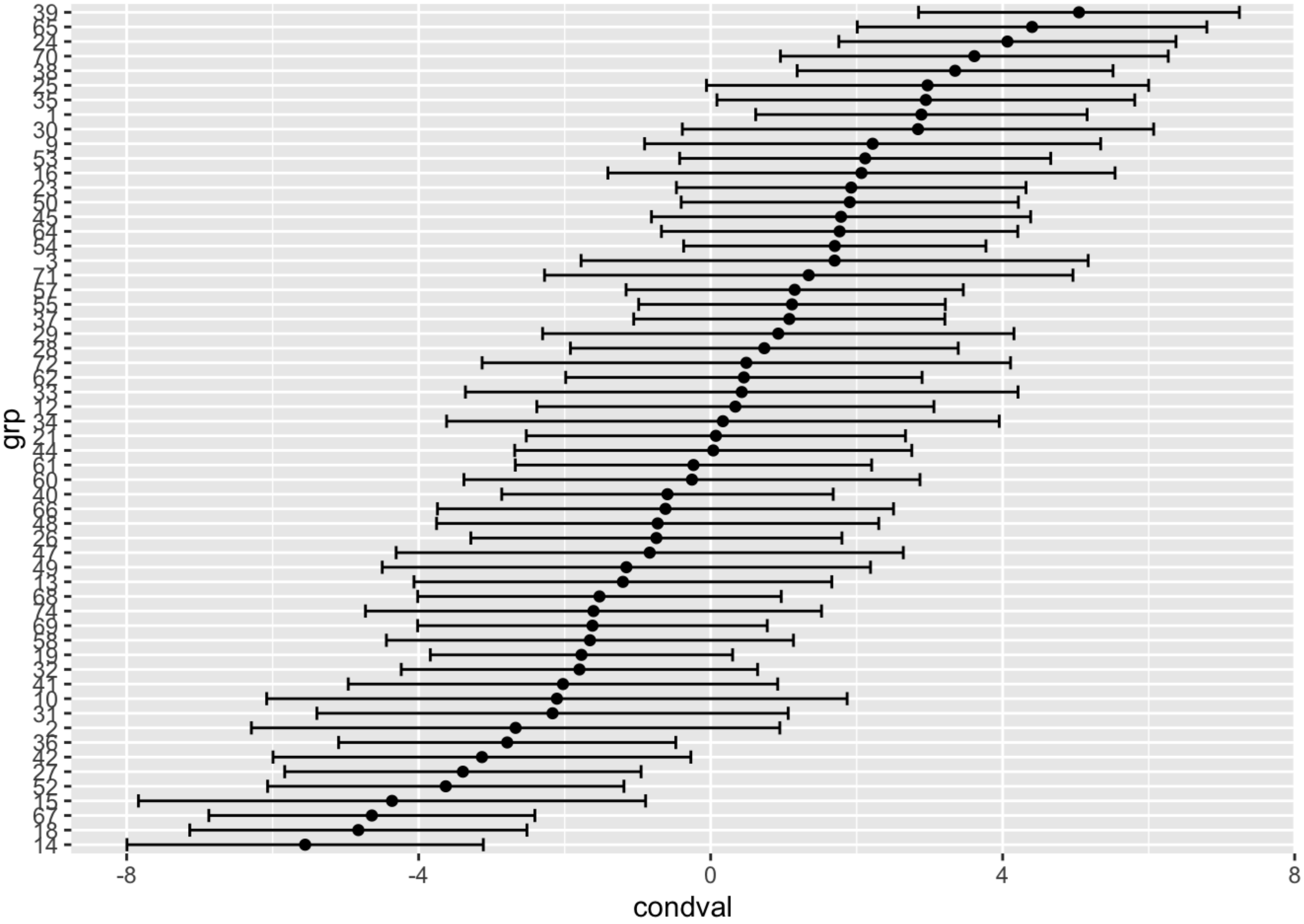
Briefly describe similarities and differences between the coefficients of the fixed effects in the results from 1d and 1e and what causes the differences. You may wish to use the use summaries of the data to help you. See the example code document.

Similarities between coefficients derived from 1d and 1e are slopes and differences between coefficients derived from 1d and 1e are intercepts. This is because we are only adding random intercepts effects and the model we fit in 1e are consists of multiple parallel lines with same slopes but different intercepts. What's more, the confidence interval derived in 1e are larger comparing to 1d due to the addition of random effects.

Question 1g

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

```
random_effects <- lme4::ranef(line2, condVar = TRUE)
ranef_df <- as.data.frame(random_effects)
ranef_df %>%
ggplot(aes(x = grp, y = condval, ymin = condval - 2*condsd, ymax = condval + 2*condsd))
 + geom_point() + geom_errorbar() + coord_flip()
```



It seems reasonable to have included these random effects. From the summary table above in 1e, the variance of random effects is 8.177 and the variance of residual is 38.240. The porportion of random effects variance among total variance of the model is: 8.177/(8.177 + 38.240) = 17.616%. This means adding random intercepts can explain 17.616% of the total variance, so the random effect should be included in the model due to its helpfulness.

Also, from the plot of confidence intervals for estimated random effects, we can see that the confidence intervals are not completely overlapping with each other (both upper bound and lower bound are not equal to each other) and the point estimates for coefficients are different as well (the shape is skewed). This also means that random effects are required and necessary to be included. Including this will also help to explain our data more precisly.

Question 1h

Write a short paragraph summarising, what you have learned from this analysis. Focus on answering the question of interest. Remember that interpreting confidence intervals is preferred to point estimates and make sure any discussion of p-values and confidence intervals are statistically correct. Also mention what proportion of the residual variation, after fitting the fixed effects, the differences between schools accounts for.

From the analysis of confidence interval above, we can see that the model with random intercepts effect should be adopted (1g).

From the analysis of confidence interval, student's verbal IQ score (iq), sex, socioeconomic status (ses), mean of student's verbal IQ score (mean_iq) and mean of student's socioeconomic status (mean_ses) are all significant and are associated with student's score on an end-of-year language test. However, ethic minority (minority_status) is not associated with student's score. This answers our question of interest.

To ensure statistical correctness, the conclusion of significance derived from confidence interval should also coincide with the conclusion derived from p-values, as p-value smaller than 0.05 (95% confidence interval) are considered as significant and the null hypothesis of coefficient equals to zero is rejected. Vice versa for p-value greater than 0.05. The proportion of residual variation over total after fitting the fixed effects is 1 - 17.616% = 82.384%. And the differences between schools accounts for differences in intercepts.

Question 2

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

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

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

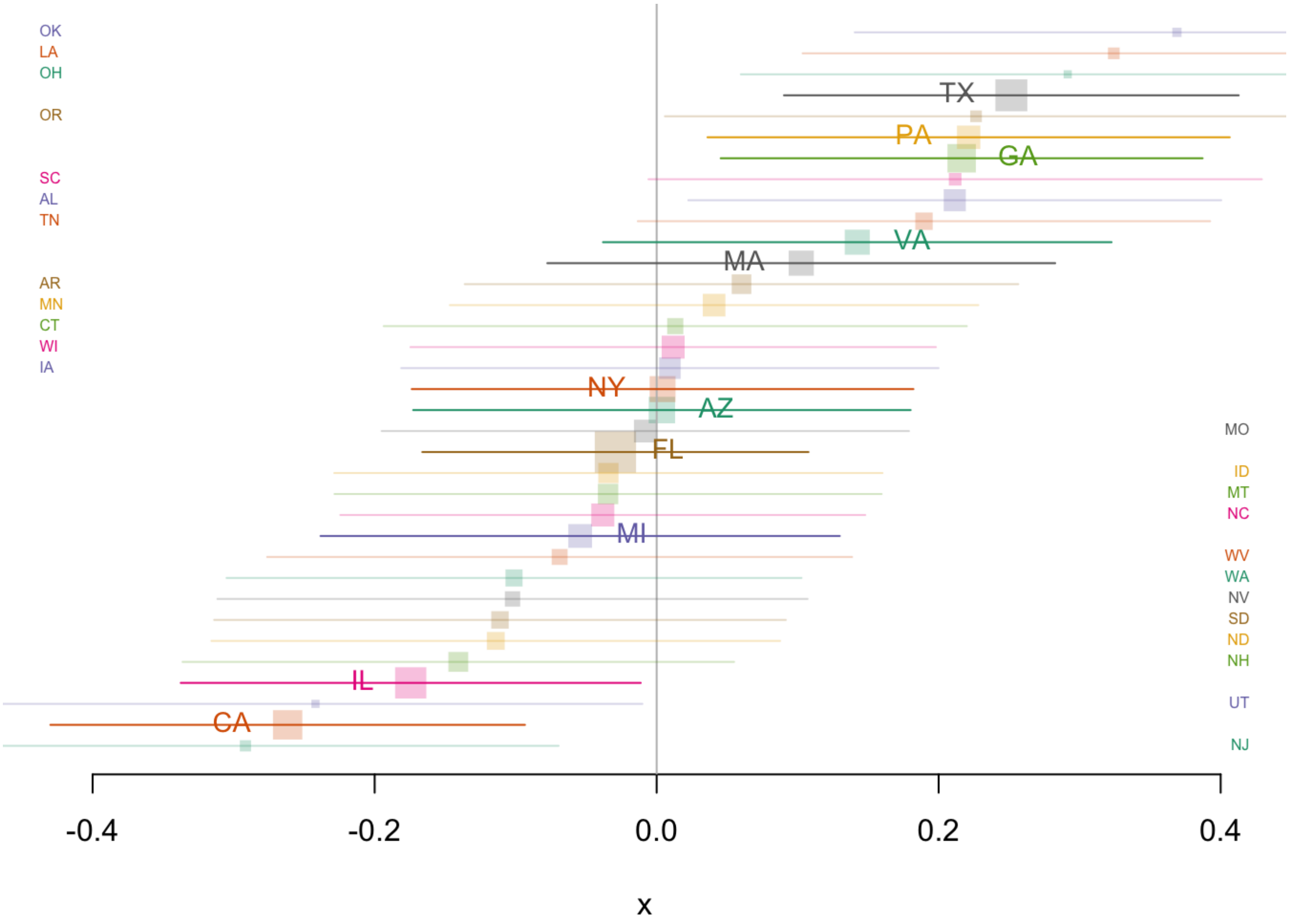
```
##                      colName
## 151 chewing_tobacco_snuff_or
##                                                                                    label
## 151 RECODE: Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
```

```
#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)
```

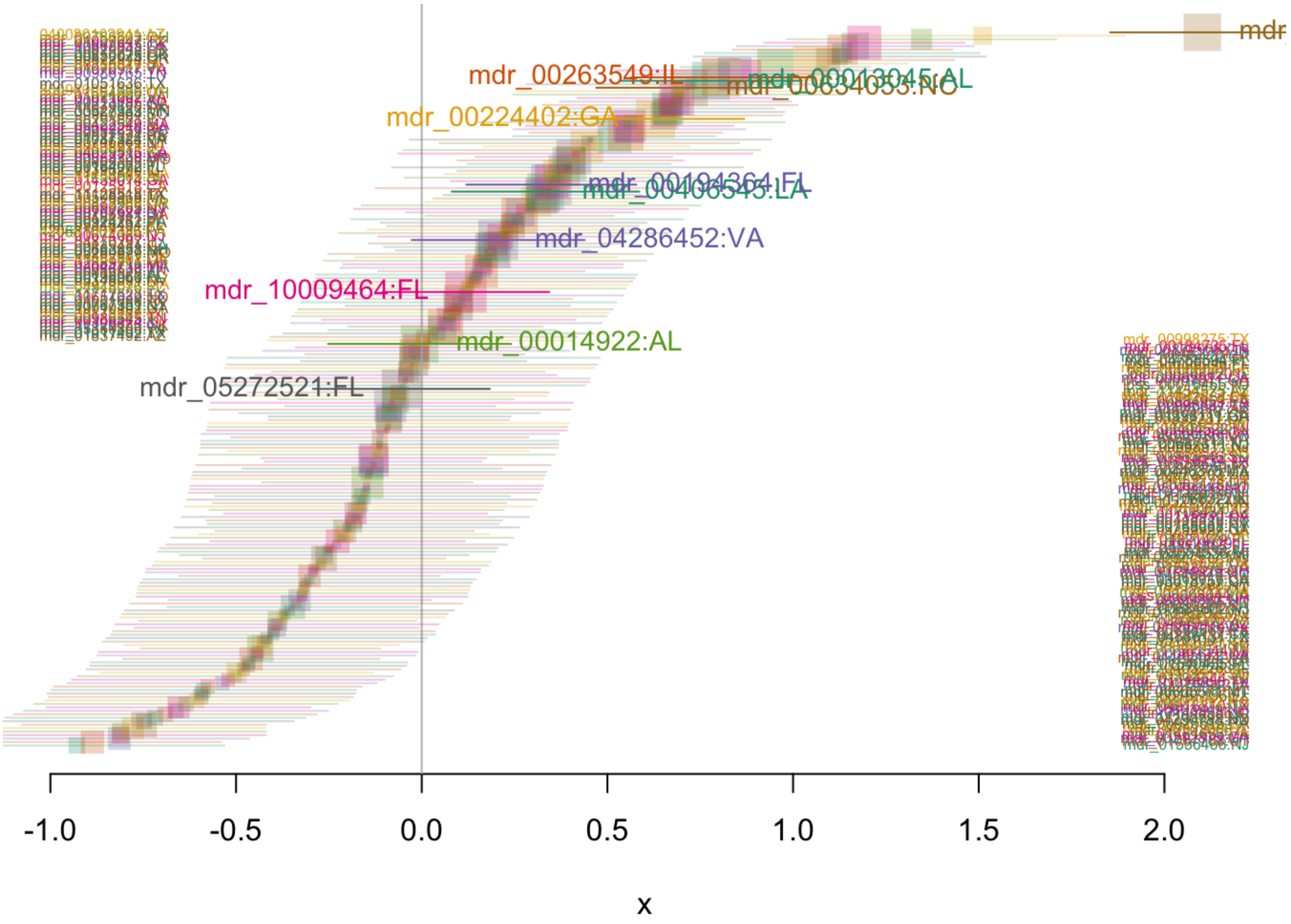
	Estimate	Std. Error	z value	Pr(>|z|)

	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,
maxNames = 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.

A statistical model corresponding to smokeModelT can be expressed as below:

$$Y_{ijk} \sim Binomial(N, \mu_{ijk})$$

$$logit(\frac{\mu_{ijk}}{1-\mu_{ijk}}) = x_{ij}\beta + A_i + B_{ij}$$

Here, A_i represents the random effect of different states (i), whereas

$$A_i \sim N(0, \sigma_A^2)$$

B_{ij} represents the random effect of different schools (j) under state, whereas

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

μ_{ijk} is the mean tobacco consumption for the k^{th} individual student from the j^{th} school of the i^{th} state.

The difference can be mainly described as: The generalized linear model (GLM) includes only fixed effects, whereas the generalized linear mixed model (GLMM) includes both fixed effects and random effects. What's more, GLMM also indicates the variance within a group; whereas GLM only indicates variance between groups.

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.

The reason why generalized linear mixed model (GLMM) with a logit link is more appropriate is because the response variable (chewing_tobacco_snuff_or) in this case is dummy, meaning that its output can only be either 0 or 1 (bernouli). And summing these responses up constructs binomial.

Question 2c

Write a paragraph assessing the hypothesis that state-level differences in chewing tobacco usage among high school students are much larger than differences between schools within a state. 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?

Based on the summary output presented in Table 3, we can see that within a particular state, the standard deviation of chewing tobacco for a school is 0.75. While among states, the standard deviation of chewing tobacco for a particular state is 0.31. We can see that the standard deviation (variance) of schools within states are larger. Therefore, we have strong evidence against the hypothesis that state-level differences in chewing tobacco usage among high school students are much larger than differences between schools within a state.

Also we can analysis through plots. The first plot derived in this question indicates the differences of tobacco usage among high school students under each state. The second plot indicates the differences of tobacco usage among high school students among different schools. We can identify that the second plot is more skewed (larger variation) comparing to the first plot, which supports the same conclusion that state-level differences in chewing tobacco usage among high school students are not larger than differences between schools within a state. Therefore, from these two aspects, we can see that targeting on schools with high tobacco using rates is more significant in determining locations of high tobacco chewing.

Question 3

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

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

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

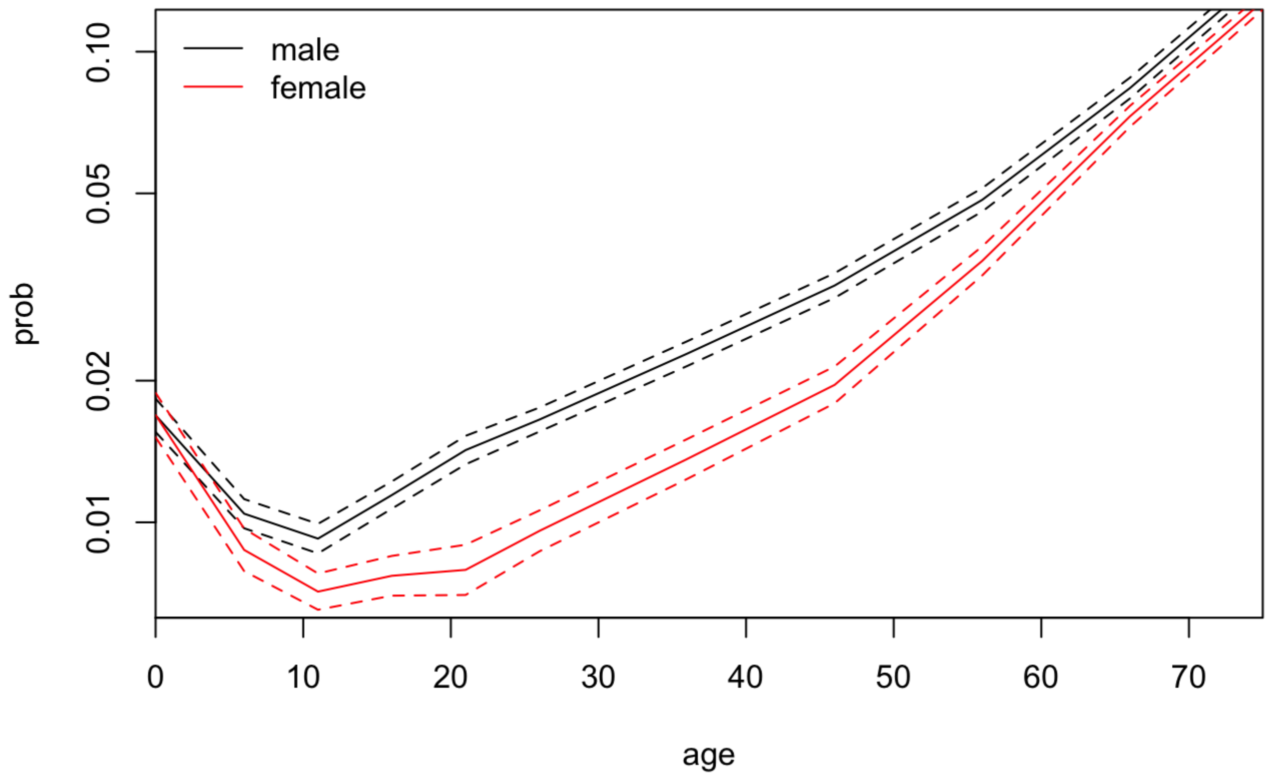
	Estimate	Std. Error	z value	Pr(>|z|)
Light_ConditionsDarkness - lights lit	0.995	0.012	81.220	0.000
Light_ConditionsDarkness - lights unlit	1.176	0.052	22.415	0.000
Light_ConditionsDarkness - no lighting	2.765	0.021	131.303	0.000
Light_ConditionsDarkness - lighting unknown	0.259	0.068	3.788	0.000
Weather_ConditionsRaining 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

	Estimate	Std. Error	z value	Pr(>|z|)
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

```
#Code for fig.2
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')
```



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.

Since the experiment in this case involves long time scope, we consider to use case/control model. The normal way is to wait and collect every fatal severity injuries and slight severity injuries data. While in the case/control model, we select people who once had fatal severity injuries and select the same amount of people who once had slight severity injuries. We then conduct researches on the situation of accidents through different aspects, which are the covariates listed below.

For both theGlm and theGlmInt objects, the case group is all pedestrians involved in motor vehicle accidents with fatal severity injuries and the control group is all pedestrians involved in motor vehicle accidents with slight severity injuries.

For theGlm objects, the covariates are sex (indicator variable, which equals 1 for female and 0 for male), age (numeric variable, which is partitioned by 5 years, for example: 0-5, 6-10…), light condition (catagorical variable, which is catagorized by level of darkness) and weather condition (catagorical variable, including rain, wind, snow…).

For theGlmInt objects, the covariates are similar to theGlm objects, but it also include interaction terms. Sex, age, light condition and weather condition (same as theGlm) are included in theGlmInt as well and theGlmInt also include the interaction term between age and sex (also range from 0-5, 6-10…).

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.

As the hypothsis suggested in the prompt, we focus on both sex (women) and age (as teenagers and in early adulthood), so theGlmInt model (model 2) which includes the interaction of sex and age is selected. First, we consider sex. From Table 6: Odds ratios for theGlm and theGlmInt, we can discover that the reference group in this case is male with age 26 - 35 years old. Under the row “sex”, we can see that the odd ratio for female is 0.58, which is smaller than 1 (reference group). So we can conclude that there is lower probability for women to have fatal accident comparing to men, which means they are safer.

Next, we considr age. Our answer depends on how we define “teenagers” (21 - 25 years old or 16 - 20 years old). Since we focus on early adulthood, we can find the estimated odds ratio for women with age 21 - 25 years old under the row “sex:age”. We can see that the point estimate for women with age 21 - 25 years old’s odd ratio is 0.96 and it’s confidence interval is (0.84, 1.10). This confidence interval includes 1 (reference group), meaning that it is not significant enough to conclude, eventhough 0.96 (smaller than 1) indicates there is lower probability for women with age 21 - 25 years old to have fatal accident.

So we further check women with age 16 - 20 years old. We can see that the point estimate for women with age 16 - 20 years old’s odd ratio is 1.16 and it’s confidence interval is (1.03, 1.31). Therefore it’s significant (since 1 is not included in the confidence interval) to say that there is higher probability for women with age 16 - 20 years old to have fatal accident, since 1.16 is larger than 1 (reference group).

Therefore, we can conclude that both sex and age are determining factors to make conclusions toward research question in this case, which suggests theGlmInt model should be used. Women tend to be, on average, safer as pedestrians than men, but we are not sure women as teenagers and in early adulthood are also safer.

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 control group is a valid one for assessing whether women are on average better at road safety than man.

In order to make an assessment, we can use figure 2: Predicted probability of being a case in baseline conditions (daylight, fine no wind) with 99% CI using theGlmInt to discuss. We can see that the probability of fatal in males (indicated by the black line) are absolutely higher than the probability of fatal in females (indicated by the red line) among all age groups. So there is evidence to support the hypothesis that men are less likely than women to report minor injuries caused by road accidents, since all accidents are either classified as fatal or slight.

This experiment design is good, but not good enough. There might exists a bias in data, as people with different sex may have different definition toward injuries, which may affect both classification of injuries (fatal or slight) and the decision of report injuries (for example: female may more likely to report). Also, the model might be a little bit inconclusive because it only includes two levels for injuries, so adding more levels may be helpful (for example: adding moderate injuries).

```
schooldata <- read.csv("/Users/mandy/Desktop/school.csv")
ggplot(schooldata, aes(x = iq, y = test)) + geom_point(alpha = 0.5) + geom_smooth(method
= "lm", se = FALSE) + theme_classic()
school1 <- schooldata %>%
    group_by(school) %>%
    mutate(mean_ses = mean(ses), mean_iq = mean(iq))
school1
line1 <- lm(test ~ iq + sex + ses + minority_status + mean_ses + mean_iq, data=school1)
summary(line1)
confint(line1)
line2 <- lme4::lmer(test ~ iq + sex + ses + minority_status + mean_ses + mean_iq + (1|sc
hool), data = school1)
summary(line2)
confint(line2)
random_effects <- lme4::ranef(line2, condVar = TRUE)
ranef_df <- as.data.frame(random_effects)
ranef_df %>%
ggplot(aes(x = grp, y = condval, ymin = condval - 2*condsd, ymax = condval + 2*condsd))
+ geom_point() + geom_errorbar() + coord_flip()

smokeFile = "smokeDownload.RData"
if (!file.exists(smokeFile)) {
download.file("http://pbrown.ca/teaching/303/data/smoke.RData", smokeFile)} (load(smokeF
ile))
smokeFormats[smokeFormats[, "colName"] == "chewing_tobacco_snuff_or",
c("colName", "label")]
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 | s
tate/school), data = smokeSub, family = binomial(link = "logit"))
knitr::kable(summary(smokeModelT)$coef$cond, digits = 2)
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))

pedestrainFile = Pmisc::downloadIfOld(
'http://pbrown.ca/teaching/303/data/pedestrians.rds')
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time), ]
pedestrians$y = pedestrians$Casualty_Severity == 'Fatal'
dim(pedestrians)
pedestrians[1:3, ]
table(pedestrians$Casualty_Severity, pedestrians$sex)
range(pedestrians$time)
theGlm = glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
data = pedestrians, family = binomial(link = "logit"))
knitr::kable(summary(theGlm)$coef, digits = 3)
theGlmInt = glm(y ~ sex * age + Light_Conditions + Weather_Conditions,
data = pedestrians, family = binomial(link = "logit"))
knitr::kable(summary(theGlmInt)$coef, digits = 3)
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')
```