STA303_A2

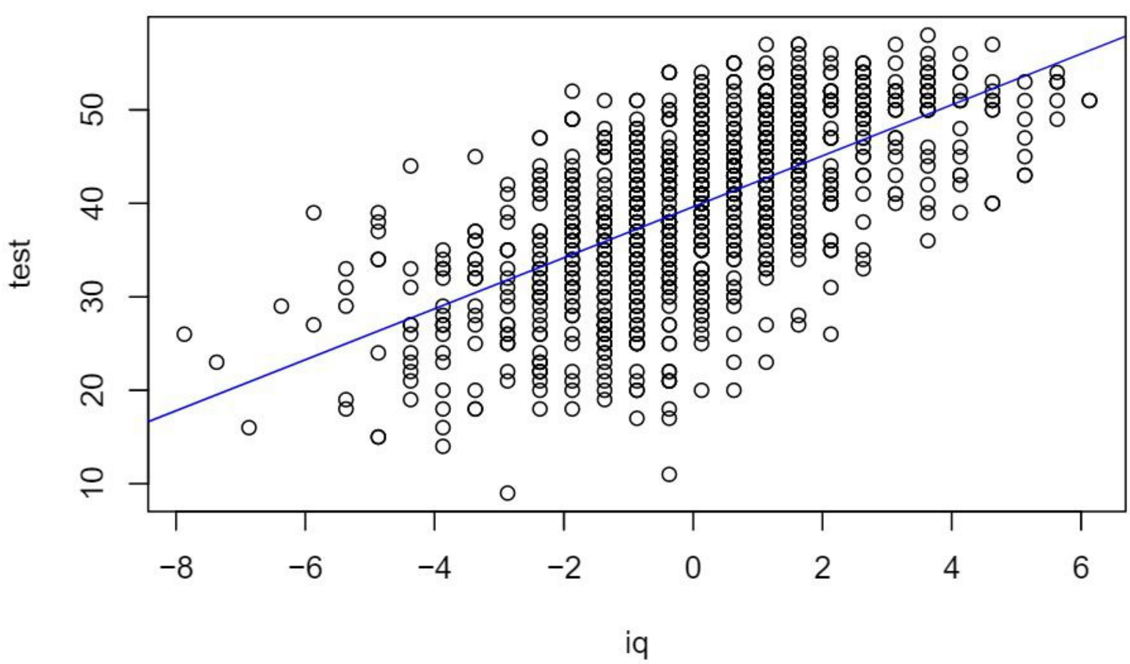
Question 1

1a

From the question, the data are collected from schools. However, althought the language test is standardize, but the learning and teaching quality in each school are different, but similar within each school. Therefore, the assumption of independence of errors(observations are independent).

1b

```
lmod <- lm(test ~ iq)
plot(iq, test)
abline(lm(test~iq), col='blue')
```



The scatter plot shows a relatively strong positive linear relationship bewteen iq and test score. THe best fit line shows a relative not constant variance of data.

1c

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

```r
school
```

```
## # A tibble: 992 x 10
## # Groups:   school [58]
##        X school    ses  test    iq   sex minority_status denomination
##    <int>  <int>  <dbl> <int> <dbl> <int>           <int>        <int>
##  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
##  7     7      1  -4.73    30 -2.37     0               1            1
##  8     8      1 -17.7     57  1.13     0               0            1
##  9     9      1 -14.7     36 -2.37     0               1            1
## 10    10      1 -12.7     36 -0.87     0               1            1
## # ... with 982 more rows, and 2 more variables: mean_iq <dbl>,
## #   mean_ses <dbl>
```

1d

```r
lmod <- lm(test ~ iq + sex + ses + minority_status + mean_ses + mean_iq, data = school)

summary(lmod)
```

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

```
##                      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
```

According to summary output, the data are poorly explained by this linear model, and the covariate minortiy_status was not able to show its statistically significance in this linear model.

From confidence interval table, interval for minority_status contain 0, which means 0 is a reasonable possibility for the true value of the difference for minority status within 95% confidence interval. In another words, there is no evidence to reject H_0 which $\beta_{minority_status} = 0$. The t test for minority_status in summary also support this conclusion.

However, covariates of iq, sex, ses, mean_ses, and mean_iq 's confidence intervals are below or above zero, which have sufficient evidence to reject $H_0 : \beta_i = 0$.

1e

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

summary(lmmod)
```

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

```
## 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
```

Sig01, 95% confidence interval for standard deviation of schools, does not contains 0. In another word, this means there is sufficient evidence to reject the H_0 that there are no difference between each school's intercept.

As mentioned in 1d, minority_status interval contain 0, which have no evidence to reject its true difference is 0 (H_0). Other than minority_status, all other covariates seems to be statistcally siginicant to this model, since they have interval that does not contain 0.

1f

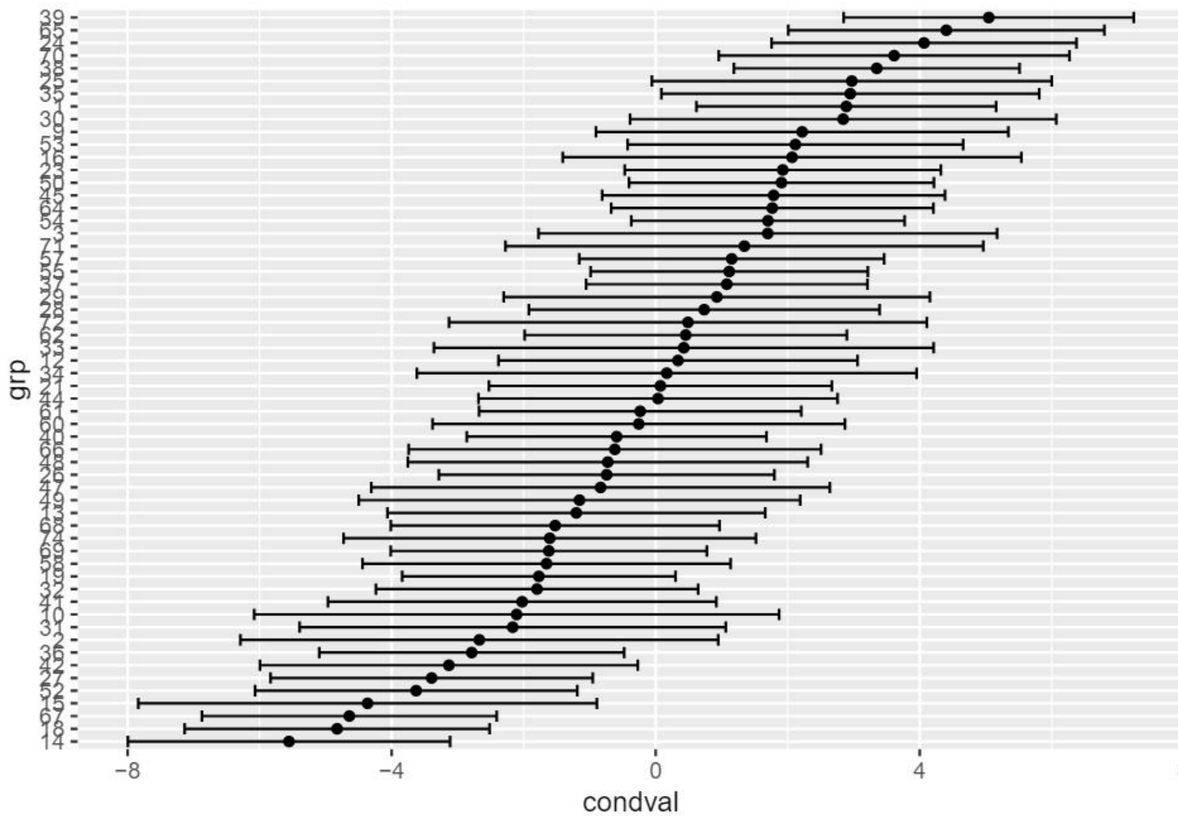
According to the result of 1d and 1e, the fixed effects from 1e has same statistically signicance on covariates with 1d. However, their is small difference on coefficients, which possibly is because of the random effects that is added in 1e.

1g

```r
random_eff <- ranef(lmmod, condVar=TRUE)

ranef_df <- as.data.frame(random_eff)

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



Yes, it seems reasonable to have a random effects, since the variation between schools are relative different.

1h

In this anaylsis, covariates :iq , sex, ses, mean_ses and mean_iq are statistcally significant in predicting a students language test score. From confidence interval table, both of these covariates' intervals are whether above or below 0. Addtionally, the t test in summary output also supports that those covariates rejects the null hypothesis. Minority Status has a 95% confidence interval which contains zero, so 0 is a reasonable possibility for the true value of the difference for minority status. Its t test also failed tp reject null hypothesis in both Linear Model and Linear Mixed Effect Model. Besides, random effects, school, also have a confidence interval that above zero, which means there is enough evidence that difference between schools are significant. Overall, variables which associated with Grade 8 students'scores on an end-of-year language test are iq , sex, ses, mean_ses, mean_iq and schools.

Question 2

set up

```
smokeFile = "smokeDownload.RData"
if (!file.exists(smokeFile)) {
download.file("http://pbrown.ca/teaching/303/data/smoke.RData",
smokeFile)
3
}
(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
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)),
]
smokeSub$ageC = smokeSub$Age - 16
library("glmmTMB")
```

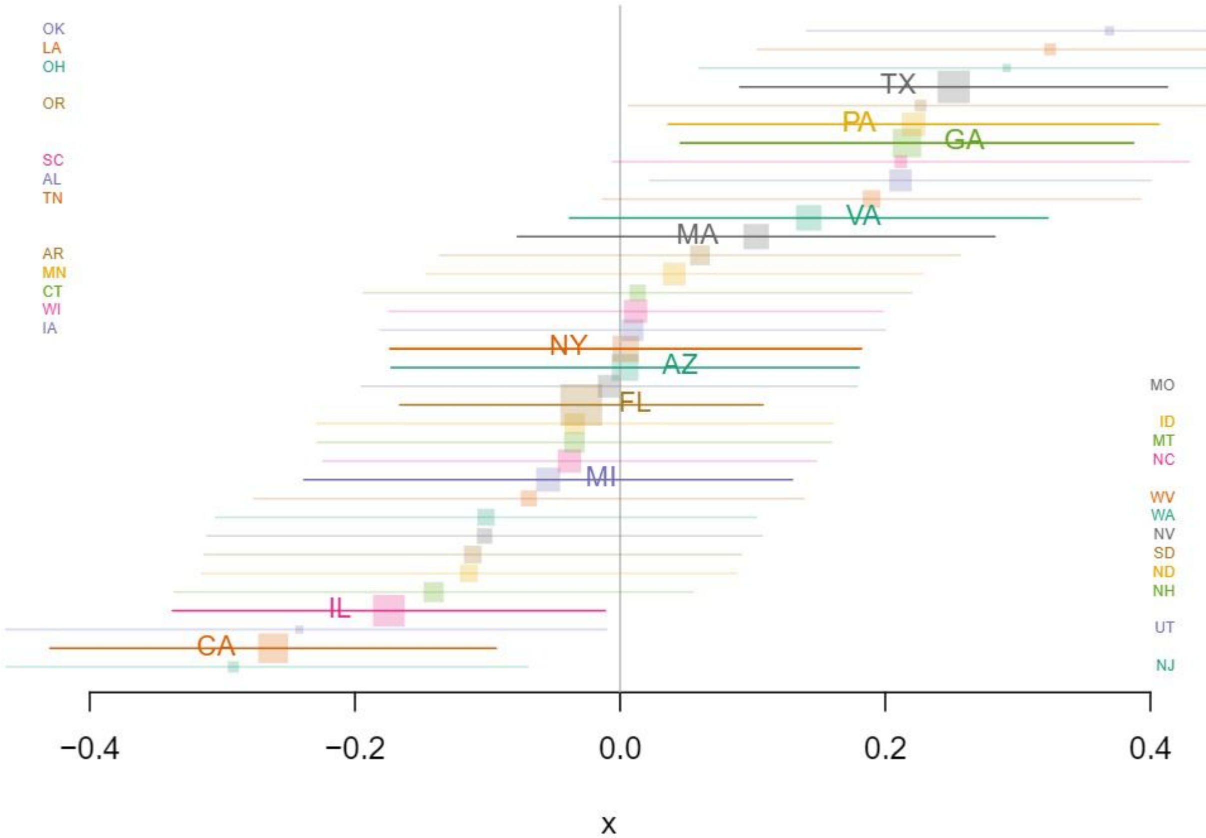
```
## Warning: package 'glmmTMB' was built under R version 3.6.2
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency 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
```

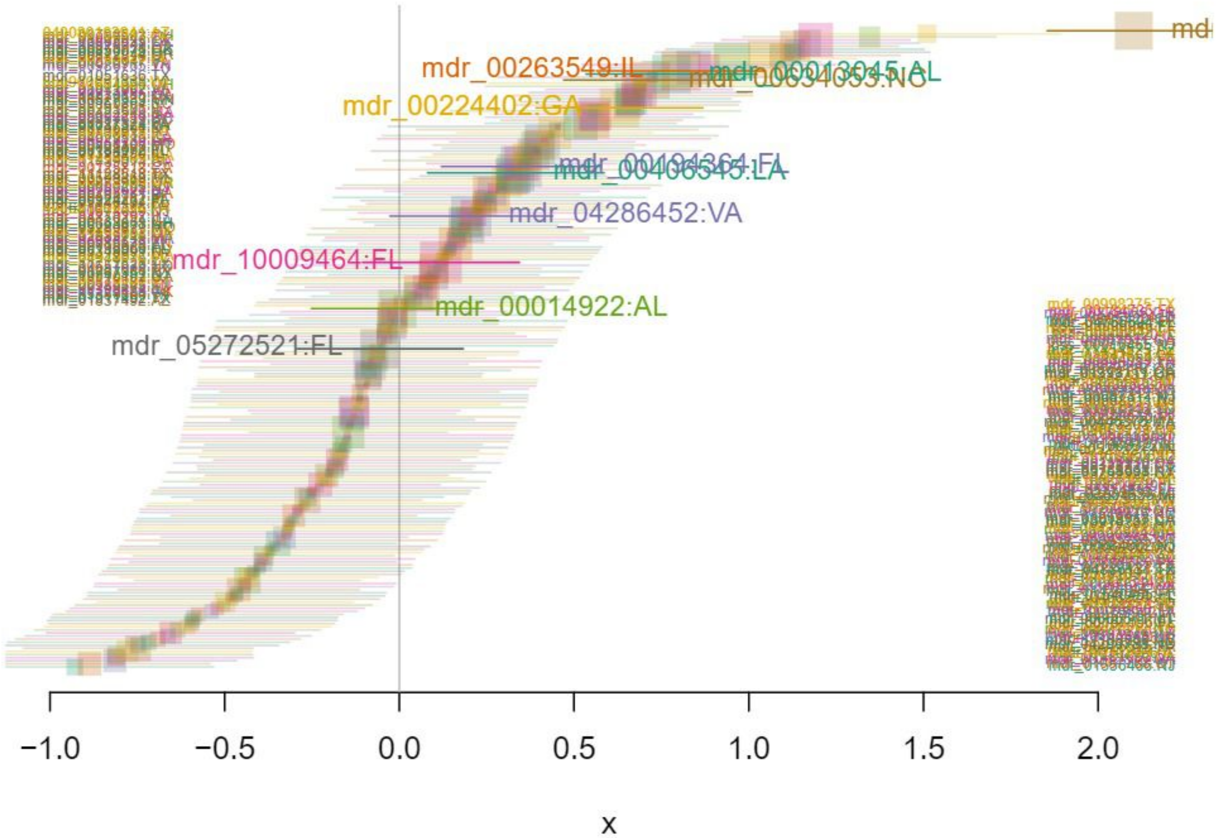
```r
smokeModelT = glmmTMB(chewing_tobacco_snuff_or ~ ageC * Sex +
RuralUrban + Race + (1 | state/school), data = smokeSub,
family = binomial(link = "logit"))
knitr::kable(summary(smokeModelT)$coef$cond, digits = 2)
```

	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
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

```r
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))
```



2a

$log(\frac{p_{it}}{1-p_{it}}) = \mu + X_{it}\beta + U_i$

X_{it} are covariates such as Race, Sex ,RuralUrbanRural, and β are coresponding coefficients. U_i is the random effect for each different school. μ is the intercept or baseline of this model. p_{it} is proportion of person chewing_tobacco_snuff_or_not.

2b

Generalized Linear Mixed Model with logit link gives a binary output which is model needed. However, linear mixed model produce continous output. Therefore, Generialized Linear Mixed Model is more suitiable for this dataset.

2c

The hypothesis that "state-level differences in chewing tobacco usage amongst high school students are much larger than differences between schools within a state" is a resonable assumptiopn. To be specific, the bias terms between each states are varied, but they are relatively smaller difference within a states. For instance, some states have larger proportion of Urban Rural population, while RuralUrbanRural is proven to be significant in model from table. As a result, the difference between states are larger than schools within a state.

To implement a program of reducing chewing tobacco usage is more efficient to identify those states where chewing is most common. First, it is much more costly to implement to all schools with high rates among states compare to simply implement to states. Secondly, those high chewing tobacco usages' lower bound is much larger than some loer chewing tobacco usages' upper bound, so it is sufficient enough to have programs in those states that chewing is common.

Question 3

```r
pedestrians = readRDS('C:/Users/maich/Desktop/pedestrians.rds')
pedestrians = pedestrians[!is.na(pedestrians$time), ]
pedestrians$y = pedestrians$Casualty_Severity == 'Fatal'
```

```r
theGlm = glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
data = pedestrians, family = binomial(link = "logit"))

theGlmInt = glm(y ~ sex * age + Light_Conditions + Weather_Conditions,
data = pedestrians, family = binomial(link = "logit"))
```

3a

Randomly select 1000 pedestrians who have experience fatal injuries to a motor vehicle accidents, and select 1000 pedestrians who have experienced slight injuries to a motor vehicle acciedent.

Cases are 1000 pedestrians who have experienced fatal injuries to a motor vehicle accidents. Control group are 1000 pedestrians who have experienced slight injuries to a motor vehicle acciedent.

Covariates are different from theGlm to theGlmInt. In theGlm, covarites are Sex, age(levels), Light Conditions, and Weather Conditions.

In theGlmInt, covarites are Sex, age(levels), Light Conditions, Weather Conditions, and Sex:Age interactions.

However, inclusion in case/control model doesn not depend on coariates but only Casualty Severity.

3b

```r
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

	Estimate	Std. Error	z value	Pr(>|z|)
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
In this case, we assume teenagers means age 11	-15 and ear	ly adulthood	means age 1	6-20.

Because of the comparsion are between different sex's age group, so GlmInt will be more appropriate for this question. Specifically, GlmInt contain the covariates of interaction between sex and ageC, and "sexFemale:age16 - 20" and "sexFemale:age11 - 15" are statistcally significant with p-values of 0.000 and 0.016 respectively. From their p-value, there is sufficient evidence to reject null hypothesis.

From Table above, the SexFemale has a negative estiamtion, which mean Female have a overall lower odds to experience a fatal motor accident. Moreover, with the interaction of sexFemale interaction terms, the estimation of Female age group is still negative compare to males. As a result, we have enough evidence to say that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood.

3c

If women are more willing to seek mdeical attention for health problems than men, while men are less likely tthan women to report minor injuries, then the proportion of men recorded in slight injuries in motor accidents will be less than its real value.

Hence, control group is not a valid one for assessing whether women are on average better at road safety than man. Because if proportion of men reported in slight injuries is less than acutal value, thus the probability of experienced a fatal injuries given a men, is higher than women.