A2 Soln

Xiangyu Kong

27/02/2020

Question 1

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

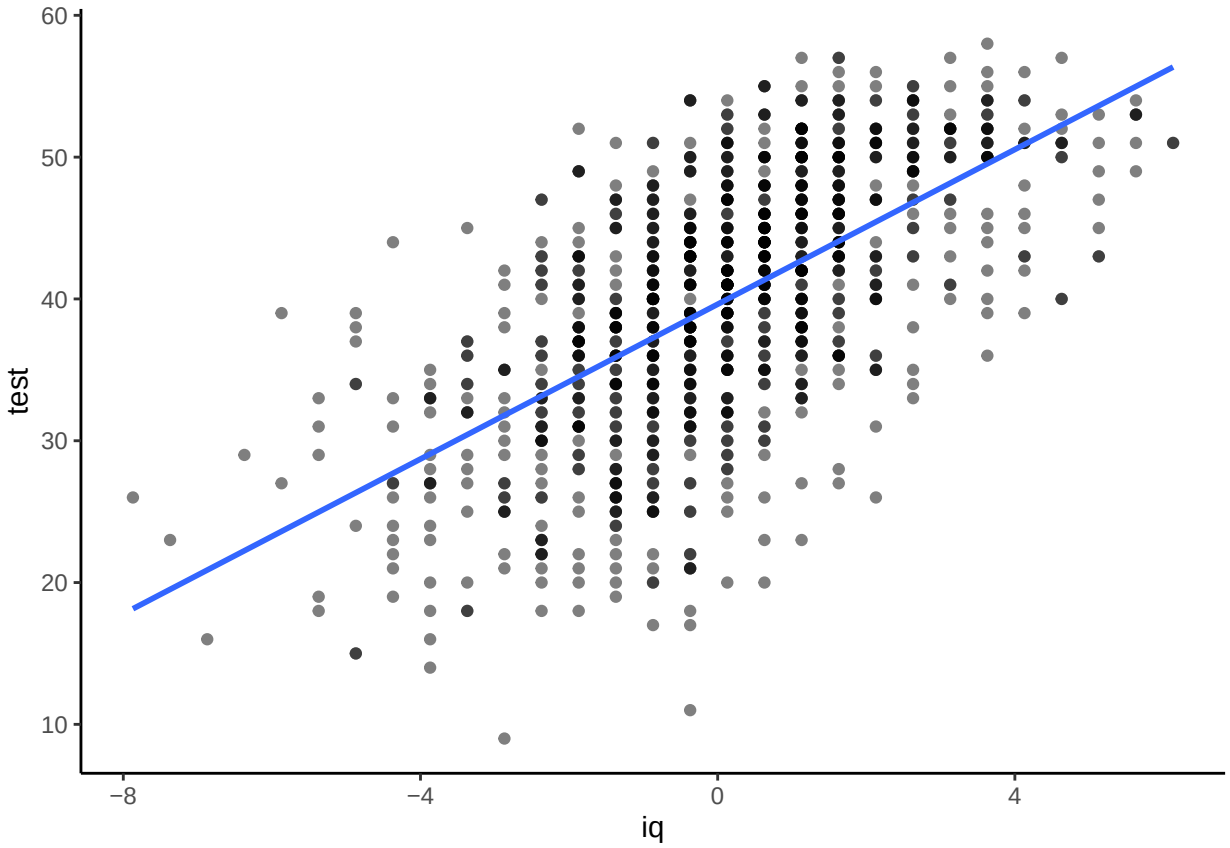
```
## Warning: Missing column names filled in: 'X1' [1]
```

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

```
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
```

```
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

Intercept:

The intercept indicates the average end-of-year language scores for a male, white (non-minority ethnics) student with a verbal IQ score of 0, who lives in the a family with socioeconomic status of 0, and studies in a school with students' mean socioeconomic status of 0 and mean verbal IQ score of 0.

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

```
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
```

```
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

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. This 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.

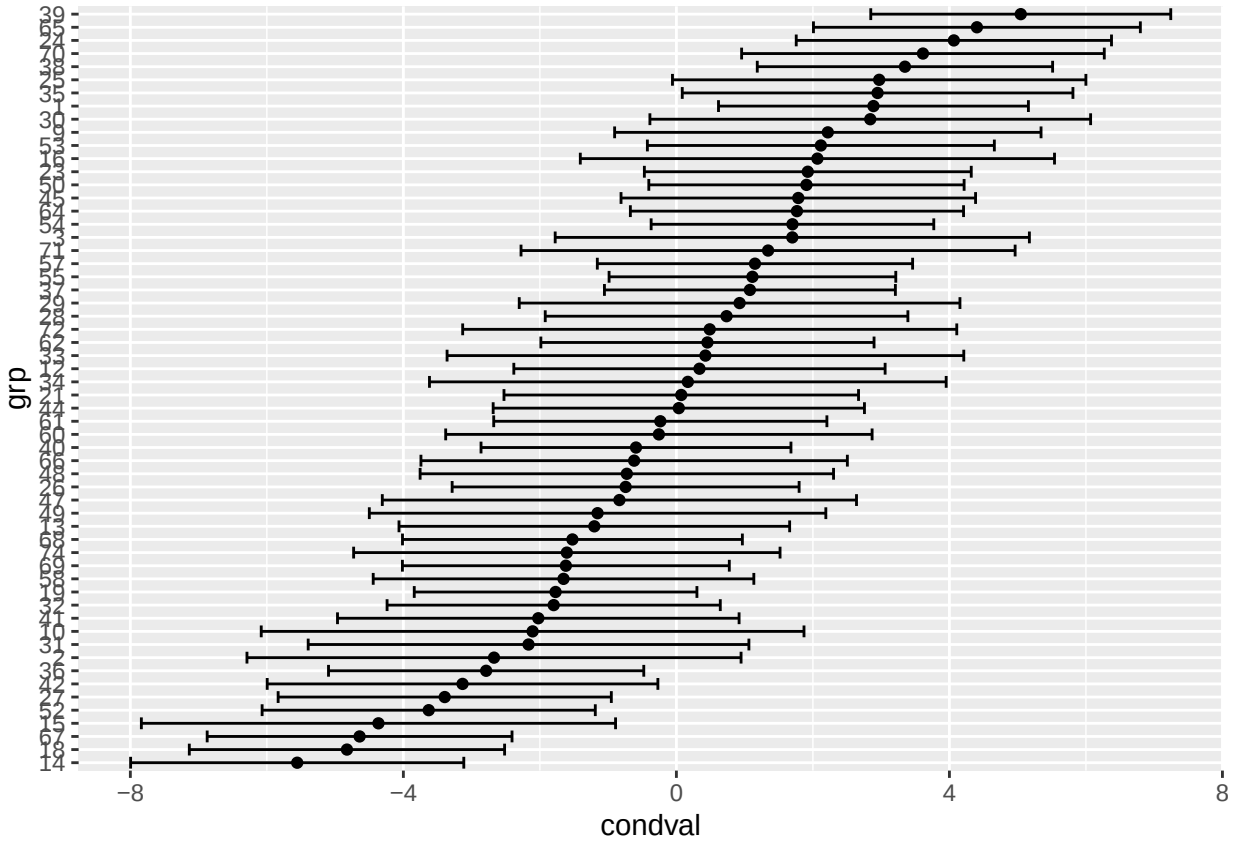
The second line in the linear mixed model's confidence interval is `.sigma`. This is the confidence interval for the residuals' standard diviation. This indicates the variability of students' end-of-year verbal test scores, not caused by different schools.

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

```
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

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")]
```

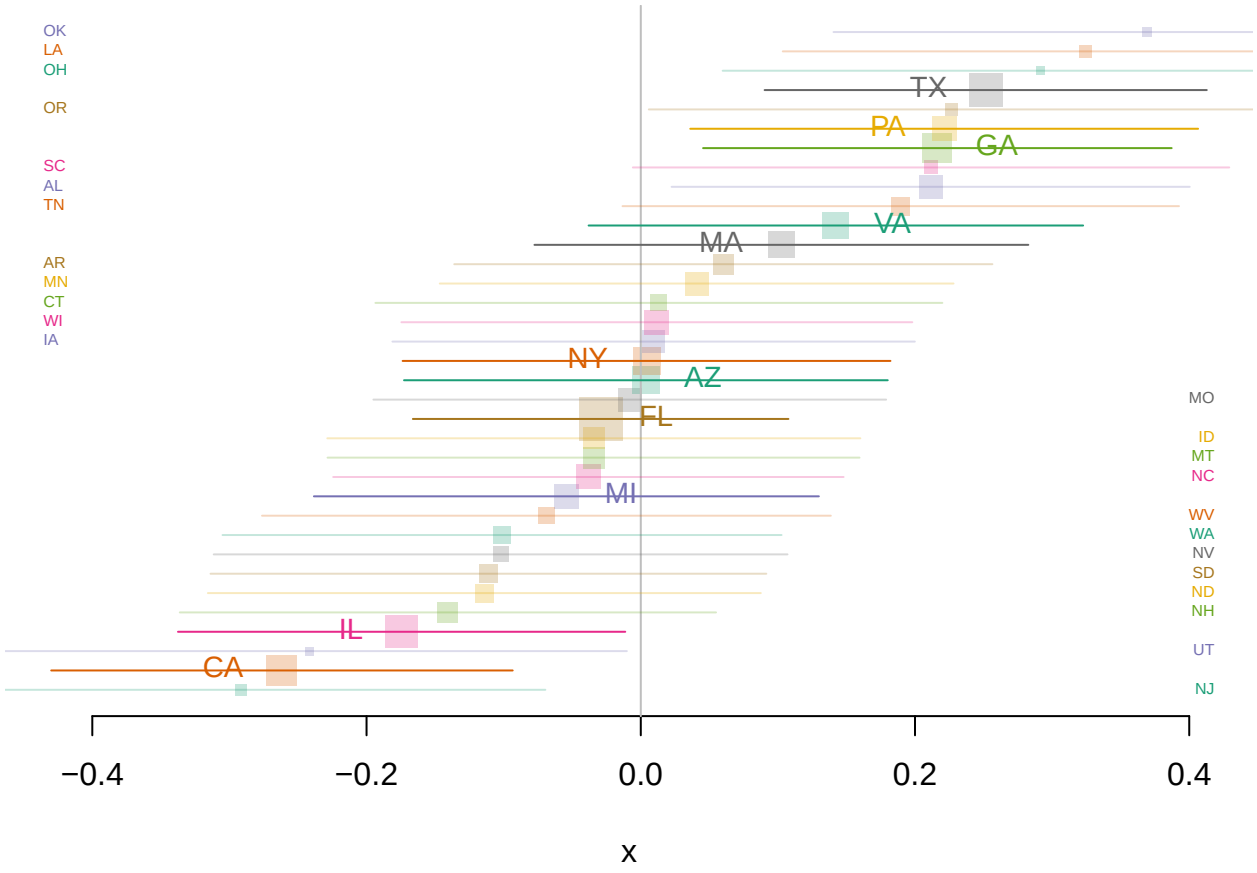
```
##                       colName
## 151 chewing_tobacco_snuff_or
##                                                                                label
## 151 RECODE: Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
```

```
# 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
```

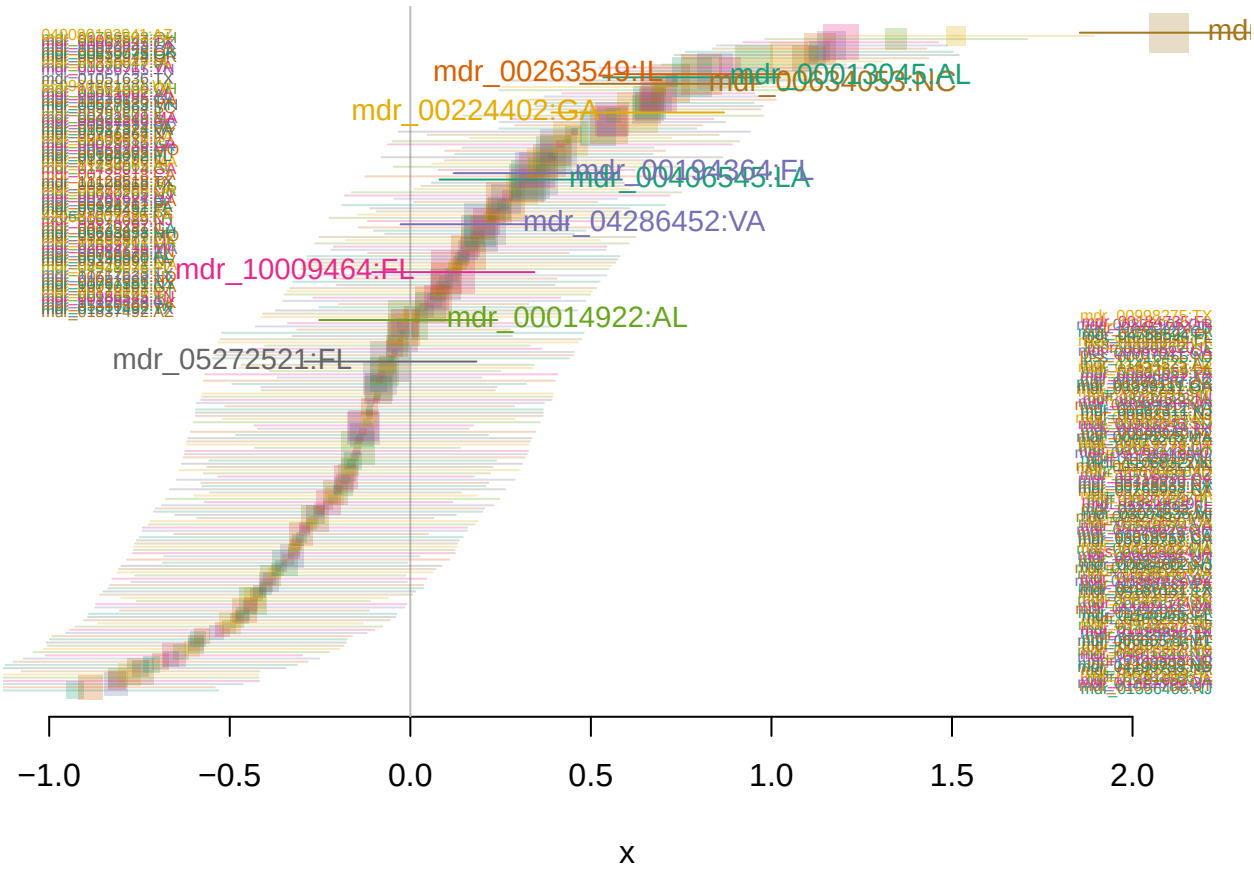
```
library("glmmTMB")

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|)
(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 2.a

Question 2.b

Question 2.c

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'
```

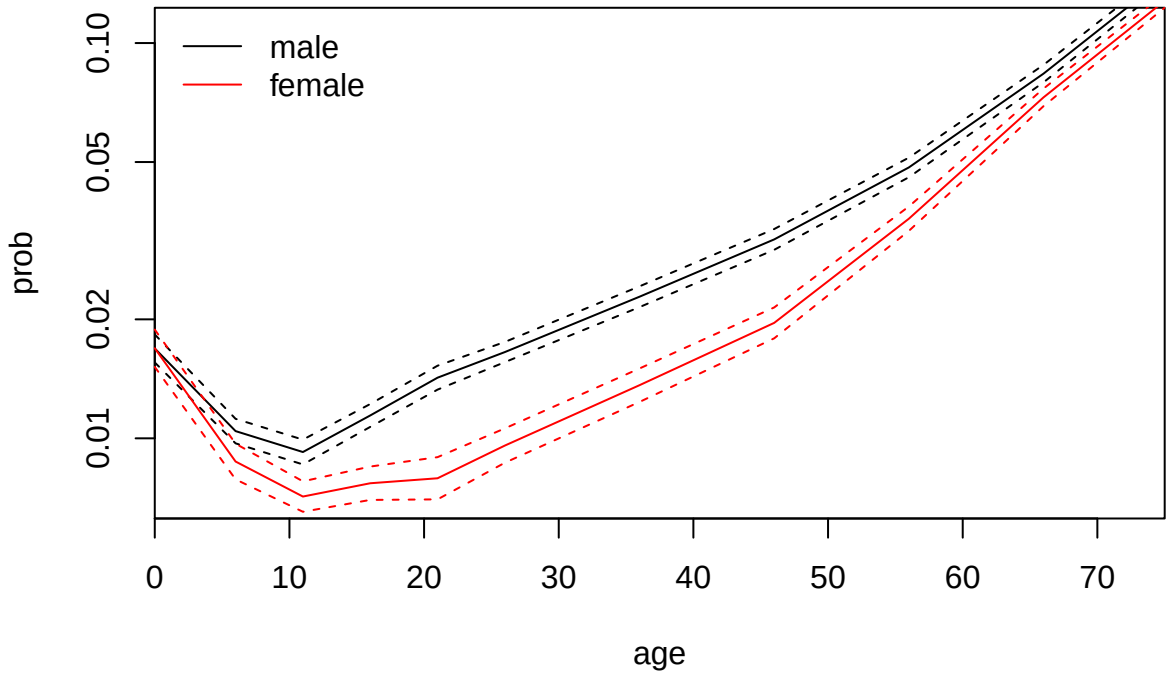
```
theGlm = glm(
  y ~ sex + age + Light_Conditions + Weather_Conditions,
  data = pedestrians,
  family = binomial(link = "logit")
)
knitr::kable(summary(theGlm)$coef, digits = 3)
```

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

Question 3.a

Question 3.b

Question 3.c