# STA303 - Assignment 2

Winter 2020

Due 2020-03-06 11:59 pm

This assignment is worth 5% of your final grade. It is also intended as preparation for Test 2 (worth 20%) and your final exam, so making a good effort here can help you get up to 33% of your final grade. You will get your feedback on Assignment 2 before Test 2.

Submission is via Crowdmark NOT Quercus. You will receive an email from Crowdmark. Contact Head TA Crystal Chen (chy.chen@mail.utoronto.ca) if you do not receive an email.

You should be able to do Question 1 by the end of week 4, Question 2 by the end of week 5 and Question 3 by the end of week 7.

- **Question 1** uses data about the IQ and language test scores for students in the netherlands, (school.csv). You will need to download this from the Assignment 2 Quercus page.
- **Question 2** uses smoking data, (smoking.RData) and instructions for obtaining the data are at the beginning of the question.
- **Question 3** uses Road accident data, (pedestrians.rds) and instructions for obtaining the data are at the beginning of the question.

*Note: You can use whatever packages are useful to you, i.e., tidyverse is not required if you prefer base R or something else. Just make sure you show which packages you are loading in a libraries chunk. Example code for this assignment is shown with tidyverse functions in the `sta303_Assignment2_example-code.Rmd` file on the Assignment 2 Quercus page.*

### Libraries used:

```r
library(tidyverse)
install.packages("Pmisc", repos = "http://R-Forge.R-project.org",
    type = "source")
```

# Question 1: Linear mixed models

The file school.csv (available on Quercus) contains data on 760 Grade 8 students (i.e., most are 11 years old) in 32 primary schools in the Netherlands. The data are adapted from Snijders and Boskers' *Multilevel Analysis*, 2nd Edition (Sage, 2012).

Table 1: Variables in the school.csv data set

| Variable | Description |
|---|---|
| school | an ID number indicating which school the student attends |
| test | the student's score on an end-of-year language test |
| iq | the student's verbal IQ score |
| ses | the socioeconomic status of the student's family |
| sex | the student's sex |
| minority_status | 1 if the student is an ethnic minority, 0 otherwise |

**Question of interest: Which variables are associated with Grade 8 students' scores on an end-of-year language test?**

## Question 1a

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

## 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.

## 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.

## 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.)

### 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.

### 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.

### Question 1g

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

### 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.

## Question 2: Generalised linear mixed models

Data from the 2014 American National Youth Tobacco Survey is available on http://pbrown.ca/teaching/303/data, where there is an R version of the 2014 dataset `smoke.RData`, a pdf documentation file `2014-Codebook.pdf`, and the code used to create the R version of the data `smokingData.R`.

You can obtain the data with:

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

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

The `smoke` object is a `data.frame` containing the data, the `smokeFormats` gives some explanation of the variables. The `colName` and `label` columns of `smokeFormats` contain variable names in `smoke` and descriptions respectively.

```
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
```

Consider the following model and set of results

```
# 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 |

Table 3: Output of Pmisc::coefTable(smokeModelT)

|  | est | 2.5 % | 97.5 % |
|---|---|---|---|
| **ref prob** |  |  |  |
| M:Urban:white | 0.04 | 0.03 | 0.06 |
| **ageC** |  |  |  |
|  | 1.43 | 1.35 | 1.52 |
| **Sex** |  |  |  |
| F | 0.13 | 0.10 | 0.17 |
| **RuralUrban** |  |  |  |
| Rural | 2.72 | 1.88 | 3.95 |
| **Race** |  |  |  |
| black | 0.22 | 0.15 | 0.31 |
| hispanic | 0.60 | 0.47 | 0.76 |
| asian | 0.33 | 0.16 | 0.65 |
| native | 1.03 | 0.58 | 1.82 |
| pacific | 3.07 | 1.43 | 6.60 |
| **ageC:Sex** |  |  |  |
| F | 0.72 | 0.65 | 0.80 |
| **sd** |  |  |  |
| school:state | 0.75 | 0.59 | 0.95 |
| state | 0.31 | 0.13 | 0.74 |

The results from this code are shown in fig. 1.

```
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

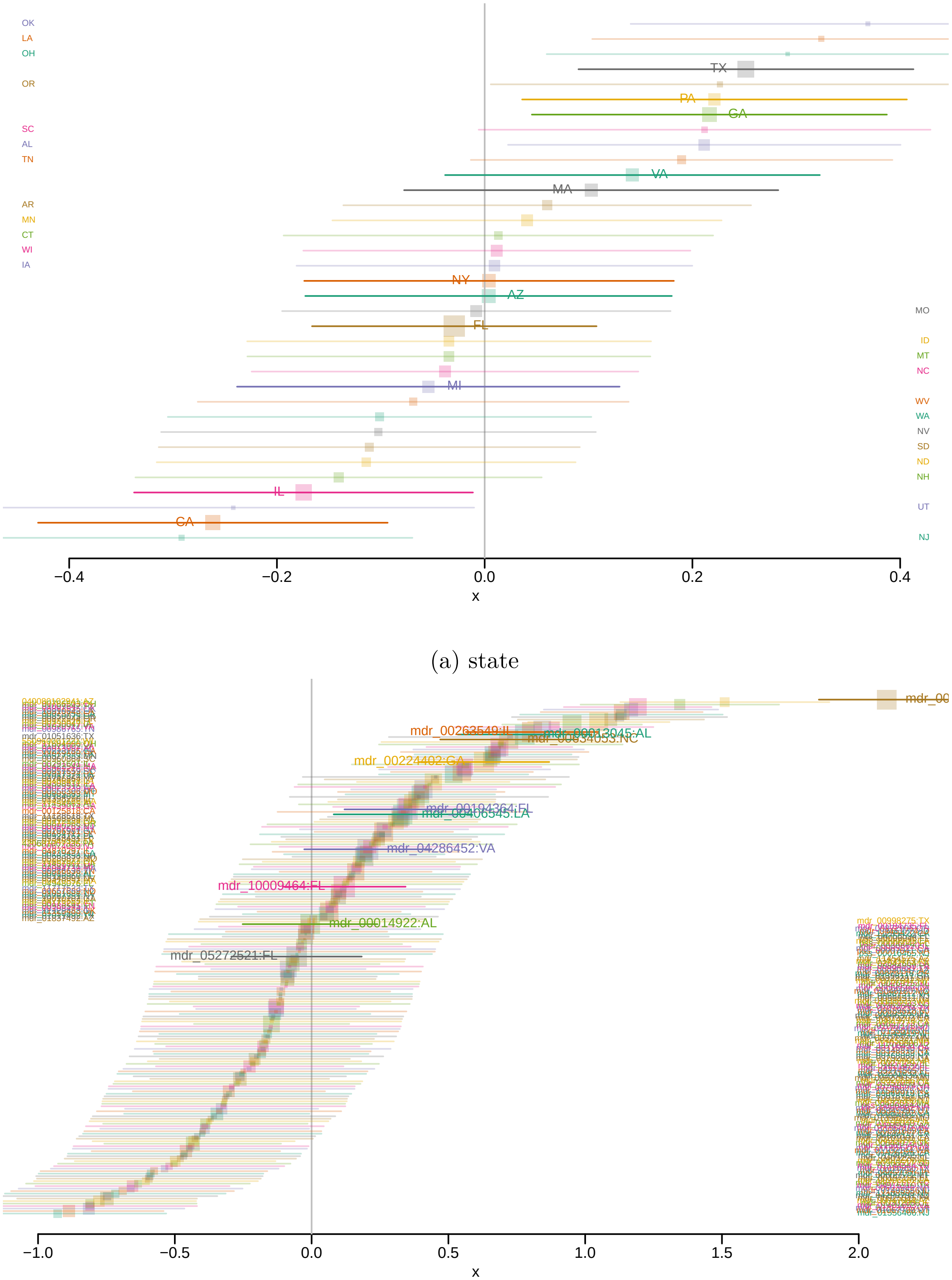
Write down a statistical model corresponding to `smokeModelT`. Briefly explain the difference between this model and a generalized linear model.

## 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.

## Question 2c

Write a paragraph assessing the hypothesis that state-level differences in chewing tobacco usage amongst 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



(a) state

(b) school

Figure 1: Conditional mean and 50pct prediction interval for random effects

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?

# Question 3: Death on the roads

The dataset below is a subset of the data from www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents, with all of the road traffic accidents in the UK from 1979 to 2015. The data below consist of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries (pedestrians with moderate injuries have been removed).

```
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"
```

Notice that men are involved in accidents more than women, and the proportion of accidents which are fatal is higher for men than for women. This might be due in part to women being more reluctant than men to walk outdoors late at night or in poor weather, and could also reflect men being on average more likely to engage in risky behaviour than women.

A glm adjusting for weather and light conditions is below.

```
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 |

Here's another GLM with interactions.

```r
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 |

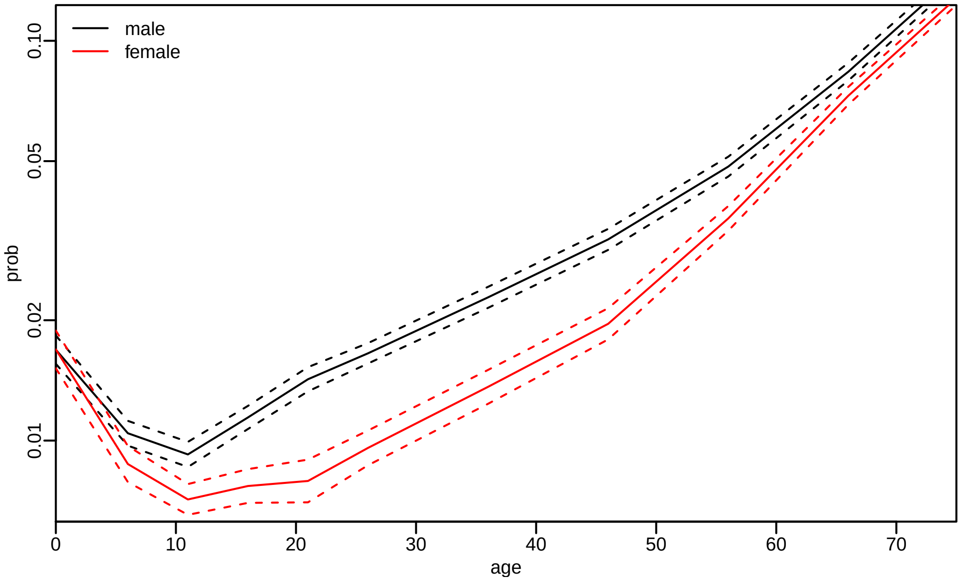


Figure 2: Predicted probability of being a case in baseline conditions (daylight, fine no wind) with 99% CI using `theGlmInt`

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

Table 6: Odds ratios for `theGlm` and `theGlmInt`.

| | model 1 | | | model 2 | | | |
|---|---|---|---|---|---|---|---|
| | est | 2.5 | 97.5 | est | 2.5 | 97.5 | |
| **ref prob** | | | | | | | |
| Male:26 - 35:Daylight:Fine no | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | |
| **sex** | | | | | | | |
| Female | 0.76 | 0.74 | 0.78 | 0.58 | 0.53 | 0.63 | |
| **age** | | | | | | | |
| 0 - 5 | 1.20 | 1.13 | 1.28 | 1.02 | 0.95 | 1.10 | 1.203 |
| 11 - 15 | 0.60 | 0.57 | 0.64 | 0.56 | 0.52 | 0.60 | 0.43 |
| 16 - 20 | 0.71 | 0.68 | 0.75 | 0.69 | 0.65 | 0.74 | 0.46 |
| 21 - 25 | 0.85 | 0.81 | 0.90 | 0.86 | 0.81 | 0.92 | 0.478 |
| 36 - 45 | 1.38 | 1.31 | 1.46 | 1.38 | 1.30 | 1.47 | 0.8244 |
| 46 - 55 | 1.93 | 1.84 | 2.04 | 1.93 | 1.81 | 2.05 | 1.186 |
| 56 - 65 | 3.12 | 2.97 | 3.28 | 2.93 | 2.76 | 3.11 | 2.17 |
| 6 - 10 | 0.70 | 0.66 | 0.74 | 0.63 | 0.59 | 0.68 | 0.526 |
| 66 - 75 | 5.81 | 5.55 | 6.08 | 5.06 | 4.78 | 5.36 | 4.4 |
| Over 75 | 10.26 | 9.82 | 10.71 | 8.84 | 8.38 | 9.33 | |
| **Light Conditions** | | | | | | | |
| Darkness - lighting unknown | 1.30 | 1.13 | 1.48 | 1.29 | 1.13 | 1.48 | |
| Darkness - lights lit | 2.70 | 2.64 | 2.77 | 2.69 | 2.63 | 2.76 | |
| Darkness - lights unlit | 3.24 | 2.92 | 3.59 | 3.23 | 2.92 | 3.58 | |
| Darkness - no lighting | 15.89 | 15.24 | 16.56 | 15.58 | 14.95 | 16.24 | |
| **Weather Conditions** | | | | | | | |
| Fine + high winds | 1.19 | 1.11 | 1.28 | 1.19 | 1.11 | 1.28 | |
| Fog or mist | 1.07 | 0.93 | 1.23 | 1.07 | 0.93 | 1.22 | |
| Raining + high winds | 0.94 | 0.87 | 1.01 | 0.94 | 0.87 | 1.02 | |
| Raining no high winds | 0.81 | 0.78 | 0.83 | 0.81 | 0.78 | 0.84 | |
| Snowing + high winds | 0.58 | 0.41 | 0.81 | 0.58 | 0.41 | 0.81 | |
| Snowing no high winds | 0.47 | 0.39 | 0.57 | 0.47 | 0.40 | 0.57 | |
| **sex:age** | | | | | | | |
| Female:0 - 5 | | | | 1.73 | 1.51 | 1.97 | |
| Female:11 - 15 | | | | 1.33 | 1.18 | 1.50 | |
| Female:16 - 20 | | | | 1.16 | 1.03 | 1.31 | |
| Female:21 - 25 | | | | 0.96 | 0.84 | 1.10 | |
| Female:36 - 45 | | | | 1.03 | 0.91 | 1.16 | |
| Female:46 - 55 | | | | 1.06 | 0.94 | 1.19 | |
| Female:56 - 65 | | | | 1.28 | 1.15 | 1.43 | |
| Female:6 - 10 | | | | 1.44 | 1.27 | 1.64 | |
| Female:66 - 75 | | | | 1.50 | 1.36 | 1.66 | |
| Female:Over 75 | | | | 1.51 | 1.37 | 1.66 | |

## 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.

## 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.

## 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.

## Some code

download data

```
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'
```

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')
```