STA303 - Assignment 2

Winter 2020 Ishaan Nagi 1002452525

Setup

Question 1: Linear mixed models

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

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

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   school = col_double(),
##   ses = col_double(),
##   test = col_double(),
##   iq = col_double(),
##   sex = col_double(),
##   minority_status = col_double(),
##   denomination = col_double()
## )
```

```
school_data= select(school_data,-c(X1))
```

Question 1a

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

- The assumption of **independence** of student test score concerns me. A class of students may share the same teacher, and may have access to better school facilities as compared to other students from different schools. Then, the school they attend may also make a difference in the end-of-year language test scores.

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.

```
ggplot(school_data, aes(x = iq, y = test), xlab="Verbal IQ") +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Verbal IQ", y = "Test Score") +
  labs(title="Verbal IQ vs End-of-Year Test Score")
```

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.

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

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.)

```
lm_1d = lm(data =school_data, test ~ iq + sex
           + ses + minority_status + mean_ses + mean_iq)
summary(lm_1d)
```

```
##
## 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 < 0.0000000000000002 ***
## iq               2.28556    0.11979  19.079 < 0.0000000000000002 ***
## sex              2.34325    0.43385   5.401    0.000000083038062 ***
## ses              0.19332    0.02641   7.319    0.000000000000519 ***
## minority_status -0.17083    0.97592  -0.175                0.861
## mean_ses        -0.21555    0.04641  -4.644    0.000003877027968 ***
## mean_iq          1.42674    0.30264   4.714    0.000002773738749 ***
## ---
## 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: < 0.00000000000000022
```

```
confint(lm_1d)
```

```
##                      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
```

(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.)

- The intercept ($\beta_0 = 38.45808$) represents the average end-of-year language test score for non-ethnic males, given that all other β_i are 0.
- β_0
  - Average end-of-year language test score for non-ethnic majority males is 38.45808.
  - The interval for β_0 ranges from 37.8448162 to 39.0713519. The end of the year score for non-ethnic majority males must lie here 95% of the time.
- IQ
  - The average change in the end-of-year language test score, given a unit increase in IQ, is 2.28556.
  - The interval for iq ranges from 2.0504849 to 2.5206429, indicating that students with a unit higher iq seem to score (in the range of) 2.05 to 2.52 marks higher on the end-of-year language test (95% of the time).
- Sex
  - The difference in the average score, given that a student is female (Sex = 1) is 2.34325 [Everything else being constant].
  - The interval for sex ranges from 1.4918849 to 3.1946222, indicating that females seem to score (in the range of) 1.5 to 3.2 marks higher on the end-of-year language test as compared to males (95% of the time).
- SES

  - The average change in the end-of-year language test score, given a unit increase in SES, is 0.19332 [Everything else being constant].
  - The interval for ses ranges from 0.1414857 to 0.2451566, indicating that students with a higher unit ses seem to score (in the range of) 0.14 to 0.25 marks higher on the end-of-year language test (95% of the time).

- Minority Status
  - The expected change in the end-of-year language test score, given a unit increase in minority_status, is -0.17083 [Everything else being constant].
  - Note: this confidence interval includes 0 and so there may not be a difference. P-value indicates this variable is statistically insignificant.

- Mean SES
  - The expected change in the end-of-year language test score of a student, given a unit increase in their school's SES, is -0.21555 [Everything else being constant].
  - The interval for mean_ses ranges from -0.3066319 to -0.1244709, indicating that students whose school SES increases by 1 unit seem to score (in the range of) 0.3 to 0.1 lower, on the end-of-year language test (95% of the time).

- Mean IQ
  - The average change in the end-of-year language test score of a student, given a unit increase in the average iq of the school, is 1.42674 [Everything else being constant].
  - The interval for mean_iq ranges from 0.8328516 to 2.0206247, indicating that students whose school average IQ increases by 1 unit seem to score (in the range of) 0.8 to 2.0 higher, on the end-of-year language test (95% of the time).

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.

```
lmer_1e = lmer(data =school_data, test ~ iq + sex
               + ses + minority_status + mean_ses + mean_iq + (1 | school))
summary(lmer_1e)
```

```
## 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
```

```
confint(lmer_1e)
```

```
##                       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
```

- Estimated variance for the random effect
  - The variance associated with the Schools is 8.177 [standard deviation of 2.859].
  - After adding iq, sex, ses, minority_status, mean_ses and mean_iq, school effects explain $\left(\frac{8.177}{8.177+38.240}\right) = 17.6\%$ of the residual variance in the model.
  - This s.d. ranges from 2.1818595 to 3.51821014 (with 95% certainty), which doesn't include zero. Then this effect is statistically significant, and schools do have an effect on the end-of-year language test scores.
- The standard deviance associated with all other factors other than schools is 6.184, the associated confidence interval doesn't include 0 and is significant.
- The intercept ($\beta_0 = 38.37951$) represents the average end-of-year language test score for non-ethnic majority males.
- β_0
  - Average end-of-year language test score for non-ethnic males is 38.37951.
  - The interval for β_0 ranges from 37.4412106 to 39.31755070. The end of the year language test score for non-ethnic males lies here 95% of the time.
- IQ
  - The average change in the end-of-year language test score, given a unit increase in IQ, is 2.27784 [Everything else being constant].
  - The interval for iq ranges from 2.0649432 to 2.49094360, indicating that students with a unit higher iq seem to score (in the range of) 2.06 to 2.49 marks higher on the end-of-year language test (95% of the time).

- Sex
  - The average increase given that a student is female (Sex = 1) is 2.29199 [Everything else being constant].
  - The interval for sex ranges from 1.5044771 to 3.08014874, indicating that females seem to score (in the range of) 1.50 to 3.08 marks higher on the end-of-year language test as compared to males (95% of the time).
- SES
  - The average change in the end-of-year language test score, given a unit increase in SES, is 0.19283 [Everything else being constant].
  - The interval for ses ranges from 0.1459275 to 0.23975452, indicating that students with a higher unit ses seem to score (in the range of) 0.145 to 0.239 marks higher on the end of the year test (95% of the time).
- Minority Status
  - The expected change in the end-of-year language test score, given a unit increase in minority_status, is -0.65259 [Everything else being constant].
  - The interval for minority_status ranges from -2.5423935 to 1.24925972
  - Note: this confidence interval includes 0 and so there may not be a difference. P-value indicates this variable is statistically insignificant.
- Mean SES
  - The expected change in the end-of-year language test score of a student, given a unit increase in their school's SES, is -0.20131 [Everything else being constant].
  - The interval for mean_ses ranges from -0.3564217 to -0.04606047, indicating that students whose school SES increases by 1 unit seem to score (in the range of) 0.35 to 0.046 lower, on the end of the year test (95% of the time).
- Mean IQ
  - The average change in the end-of-year language test score of a student, given a unit increase in the average iq of the school, is 1.62512 [Everything else being constant].
  - The interval for mean_iq ranges from 0.6166461 to 2.63522563, indicating that students whose school average IQ increases by 1 unit seem to score (in the range of) 0.6 to 2.6 higher, on the end-of-year language test (95% of the time).

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.

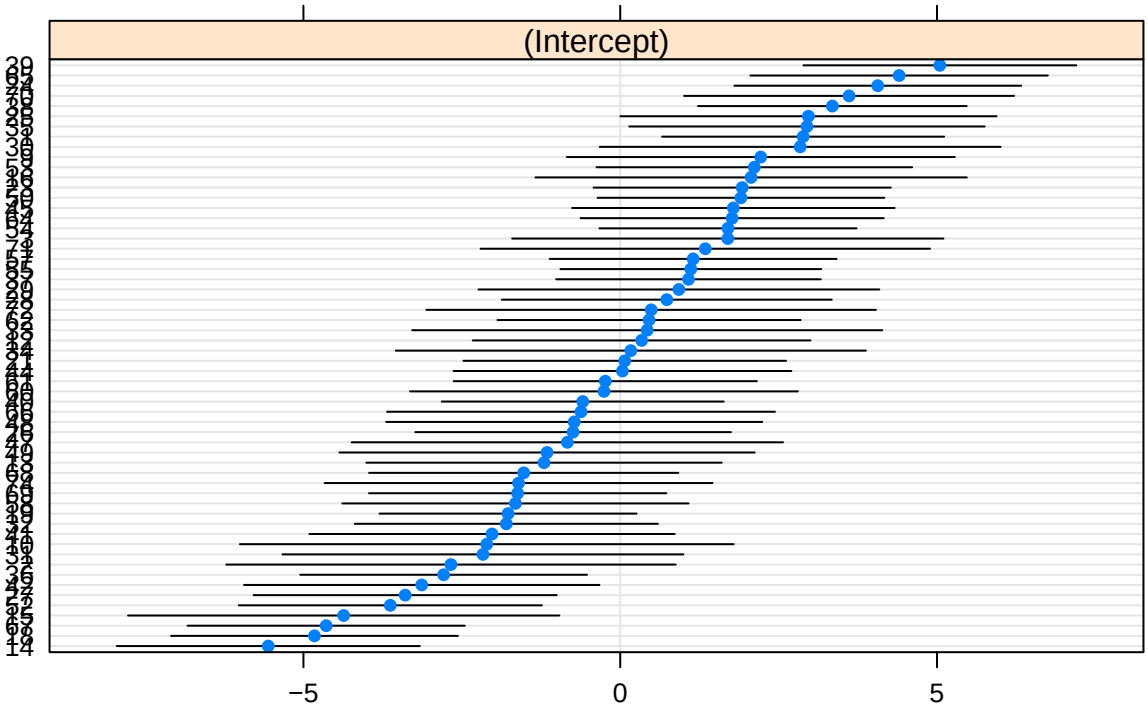
- There is a difference of the coefficients in model from 1d and in 1e. This is simply because model 2 computes different intercepts for individual schools.
- Note, in model2 each school is considered independently. Then number of observations fall from 992 to 53. This may give slighter larger confidence intervals.

Question 1G

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

```
plt = (lme4::ranef(lmer_1e, condVar=TRUE))
lattice::dotplot(plt)
```

```
## $school
```



- Yes, there is a non-zero variance between the schools. The 95% confidence interval does not contain 0. This indicates schools do make a difference in the end-of-year language test scores. It was **correct** to include 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.

- From our analysis it is evident that variables such as Verbal IQ (iq), Sex (sex), Socio-Economic Status of the student's family (ses), the average socio-economic status of a school (mean_ses), and the average Verbal IQ of the school (mean_iq) have an impact on the end-of-year language test score and are statistically significant.
- The school a student goes to also seems significant in having an impact on the end-of-year language test score.
- Then the score on the end-of-year language test is not independent for each student. The factors mentioned above, do have an impact on the end-of-year language test score.

Question 2: Generalised linear mixed models

```
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
```

```
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)), ]
smokeSub$ageC = smokeSub$Age - 16
```

Question 2a

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

$$Y_{ijk}|U \sim Binomial(N_i, \mu_{ijk})$$

$$log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = X_i\beta + U_{jk}$$

$$U \sim MVN(0, \Sigma)$$

- $\mu_i jk$ refers to the probability that the ith subject who went to jth school in the kth state used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days, given X_i and $U_j k$.
- $U_j k$ represents the effect (deviation from population average) of jth school in the kth state on a subject in relation to probability that a subject Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days.
- GLMMs estimate random intercepts, slopes and models variations within and among groups, unlike GLMs.

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 Generalized Linear Model with logit link is more appropriate for this dataset than a mixed model since we are trying to ascertain probabilities in [0 1]. The logit link, using the sigmoid function enforces this constraint, whereas the linear mixed model would not.
- The link, logit is also beneficial for its computational advantages in the case of small probabilities. Instead of multiplying small floating point numbers, we can sum of log-odds to calculate the log-odds joint probability.

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 tosell 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?

- According to our analysis, variance among schools is higher than the variance between states. Then 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 **false**.
- 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), it would be important to **find individual schools** with high chewing rates since schools explain a higher percentage of the variance in the model.

Question 3: Death on the roads

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.

$$Y_i \sim Binomial(p_i)$$

$$logit(p_i) = X_i\beta$$

* p_i is the probability that a pedestrian involved in motor vehicle accident had a severity of injury as 'Fatal'.

- Case-control study with male and female subjects, age [0 , 75+], lighting conditions [lit, unlit, no lighting, unknown] and weather conditions [some combination of Snow, Rain, yes/no high-wind, Fogs or Mist].
- Cases were pedestrians involved in motor vehicle accidents with Casualty_Severity as 'Fatal'.
- Pedestrians involved in motor vehicle accidents with Casualty_Severity as 'slight' were identified as controls.
- Covariates: Sex, age, lighting conditions, weather conditions.

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.

- 'theGlmInt' (model2) with the Sex-Age interaction, has a majority of statistically significant estimates that significantly affect the odds of being a case. Then 'theGlmInt' is more appropriate for addressing this research question.
- Yes, UK road accident data are consistent with the hypothesis that women tend to be safer on average:
  - Odds of being a Females case is 0.58 times the odds of being a Male case.
- On being safer as pedestrians than men as teenagers and in early adulthood:
  - The confidence intervals seem to overlap in the early half of childhood [0-7/8], hence nothing can be said with high certainty.
  - After the age of 7/8, women seem to be safer (as indicated by the data and Figure 2).

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 men.

$$e^{\beta_{\text{female}}} = \frac{\left(\frac{\text{Female Cases}}{\text{Female Control}}\right)}{\left(\frac{\text{Male Cases}}{\text{Male Control}}\right)}$$

- Female control is over-represented because women are generally more willing to seek medical attention for health problems than men.
- Also, it is hypothesized that men are less likely than women to report minor injuries caused by road accidents.
  - Then, β_{Female} is under-estimated, then this control group is not evidence to assess whether women are on average better at road safety than men.
  - Then our case-control assumptions are violated.