A2 Soln

Xiangyu Kong, 1002109620

05/03/2020

Question 1

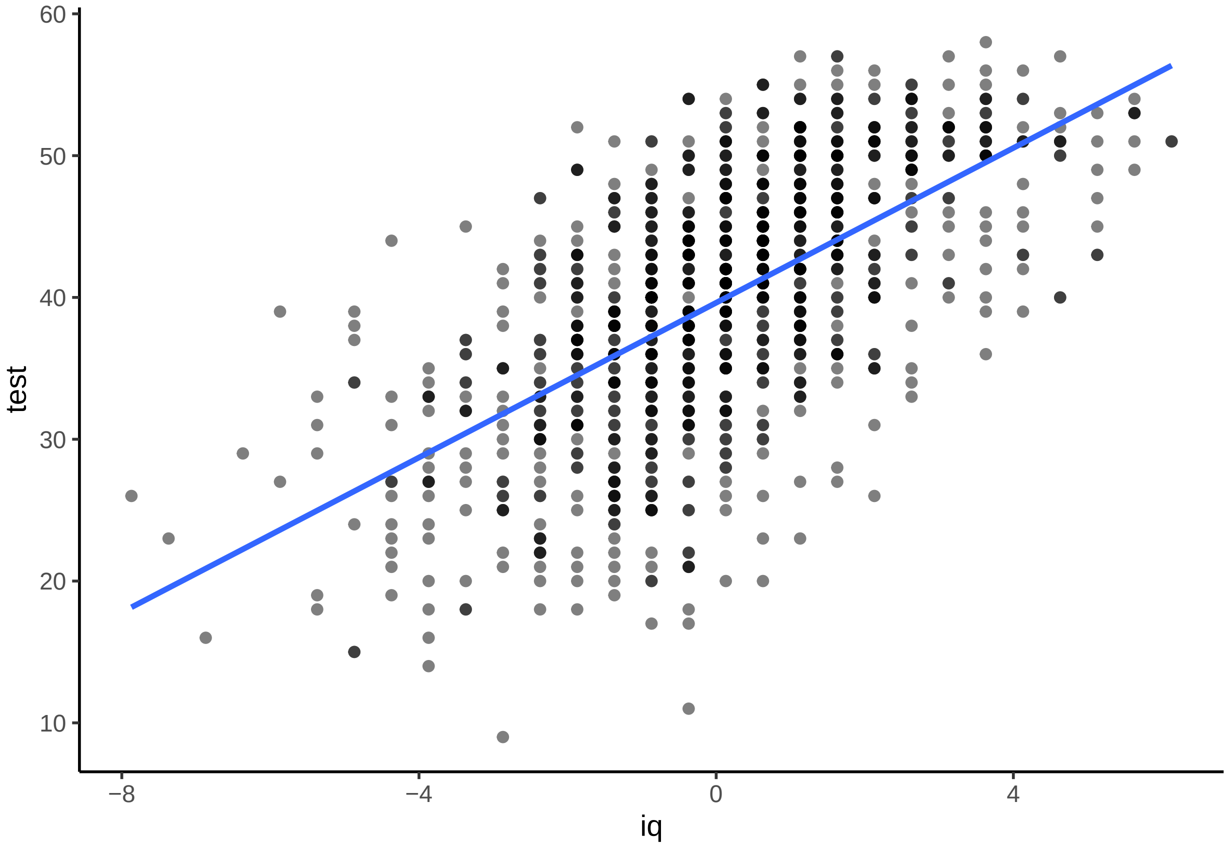
```
school_data = read_csv("school.csv")
```

Question 1.a

The independence assumption may be violated. Students from the same school may produce end-of-year language scores similar to each other. For example, A school with better teaching resources or environment may be more likely to have students with better end-of-year language scores. Thus the observations may not independent of each other.

Question 1.b

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



From the scatter plot, we observe an upward ascending trend as `iq` increases. The best fit blue line also suggests a positive relationship between `iq` and `test`. Thus we can claim that according to the plot that students' verbal iq score 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

```r
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
```

```r
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

Estimates:

- The intercept shows that the average end-of-year language scores for the baseline subgroup is 38.46.
    - This baseline subgroup consists of male, white (non-minority ethnics) students with verbal IQ score of 0, who live in families with socioeconomic status of 0, and study in schools with students' mean socioeconomic status of 0 and mean verbal IQ score of 0.
- An increase in student's verbal iq score by 1 tends to make a student's end-of-year language score increase by 2.29.
- A female student tend to have an end-of-year language score 2.34 higher than a male student.
- An increase in student's ses level by 1 tends to make a student's end-of-year language score increase by 0.19.
- A minority student tend to have an end-of-year language score 0.17 lower than a non-minority (white) student.
- An increase in the student's school's mean ses level by 1 tends to decrease the student's end of year language score by 0.22.
- An increase in the student's school's mean iq score by 1 tends to increase the student's end of year language score by 1.43.

Confidence Intervals:

- The 95% confidence interval for the model's intercept is $(37.84, 39.07)$.
- 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 is not associated to the students' language 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
```

```r
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

Random Effects:

By looking at the Random effects section in the summary, we can see that:

- The School random effect has variance of 8.177
- The Residuals has variance of 38.240
- The School random effect explain about $\frac{8.177}{8.177+38.240} \times 100 = 17.62\%$ of the total variance in the model.

Estimates:

The interpretation for the fixed effect esitmates are similar to that of the linear model's estimates.

- The intercept shows that the average end-of-year language scores for the baseline subgroup is 38.38.
  - This baseline subgroup consists of male, white (non-minority ethnics) students with verbal IQ score of 0, who live in families with socioeconomic status of 0, and study in schools with students' mean socioeconomic status of 0 and mean verbal IQ score of 0.
- An increase in student's iq level by 1 tends to make a student's end-of-year language score increase by 2.28.
- A female student tend to have an end-of-year language score 2.29 higher than a male student.
- An increase in student's ses level by 1 tends to make a student's end-of-year language score increase by 0.19.
- A minority student tend to have an end-of-year language score 0.65 lower than a non-minority (white) student.
- An increase in the student's school's mean ses level by 1 tends to decrease the student's end of year language score by 0.20.
- An increase in the student's school's mean iq level by 1 tends to increase the student's end of year language score by 1.64.

Confidence Intervals:

- `.sig01` is the confidence interval for the standard deviation for the first random effect (i.e. the `school` random effect). It means that with 95% confidence, the standard deviation explained by `school` is $(2.1819, 3.5182)$
- `.sigma` is the confidence interval for the residuals' standard deviation. This represents the standard deviation not explained by `school`.
- The 95% confidence interval for the model's intercept is $(37.44, 39.32)$.
- 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 is not associated to the students' language scores.

Question 1.f

The estimated fixed effects for `iq`, `sex` and `ses` in the mixed linear model are very similar to the estimates given by the simple linear regression model. In the mixed linear model, their 95% confidence intervals are slightly tighter.

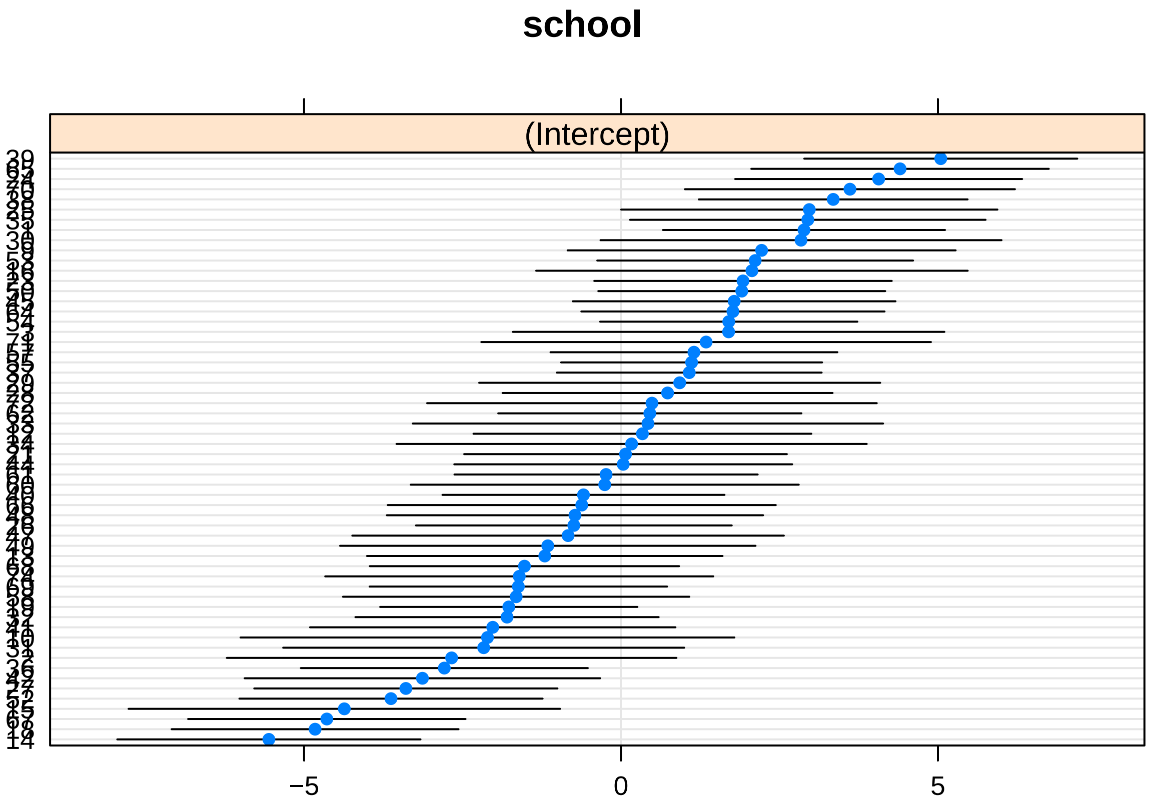
The estimated fixed effect for `mean_iq` and `mean_ses` both increased comparing to the linear regression model. By considering `school` as a random effect, we are grouping the observations into different groups according to school. However, the `mean_iq` and `mean_ses` for each school are the same. This means for each school, there is only one observation of `mean_iq` and `mean_ses`. The sample size within each group is less than the sample size used in the linear regression model. This causes their estimates to increase, and also causes their 95% confidence interval to be wider than the simple linear regression model.

The estimated fixed effect for `minority_status` has significantly decreased (more negative), but for both the generalized linear model and linear regression model, the 95% confidence intervals include 0, so the minority status is still likely to be unrelated to the students' language score.

Question 1.g

```
rand_effects <- lme4::ranef(school_lmm, condVar = TRUE)
lattice::dotplot(rand_effects)
```

```
## $school
```



The plot shows that for different schools, there conditional mean values differ from the grand mean. The range of the dots are wide apart and the plot forms a visible trend. This indicates that adding the random effect for `school` is appropriate in this case because it explains part of the the variation in mean.

Question 1.h

Schools that the students are studying in affects the student's language score. This can be visually shown in the conditional mean and confidence interval plot above, and can also be analytically shown after fitting a linear mixed model with `school` as the random effect. The `school` effects explain about $\frac{8.177}{8.177+38.240} \times 100 =$ 17.62% of the total variance in the model. To perform regression, we need to make it a grouping random effect.

After considering the `school` as a random effect, by looking at the confidence interval for `iq`, `sex`, `ses`, `mean_iq` and `mean_ses` we can see that the 95% confidence interval do not include 0. We can conclude with 95% confidence that considering the effect explained by students studying in different schools, as the students' verbal iq, their families' socioeconomic status and their schools' mean iq increase, they more likely to achieve a better end-of-year language scores. Female students are more likely to obtain a better end-of-year language score than male students. Students in schools with higher mean socialeconomic status will likely to obtain a lower score. The ethnicity of the student is likely to be uncorrelated to the students' score because the p-value for the estimate is insignificant, and the confidence intervals include 0.

Question 2

Question 2.a

The model can be represented by

$$
\begin{aligned}
Y_{ij} \mid A, B &\sim Binomial(N_{ij}, \mu_{ij}) \\
h(\mu_{ij}) = logit(\mu_{ij}) = \frac{\mu_{ij}}{1-\mu_{ij}} &= X_{ij}\beta + A_i + B_{ij} + \epsilon_{ij} \\
A_i &\sim N(0, \sigma_A^2) \\
B_{ij} &\sim N(0, \sigma_B^2)
\end{aligned}
$$

where

- Y_{ij} is the number of people who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days, from the jth school of the ith state.
- N_{ij} is the number of people in the jth school of the ith state.
- μ_{ij} is the proportion of the people who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days, from the jth school of the ith state.
- A_i is the state i's deviation from the population average
- B_{ij} is the ith state's j's school's deviation from the population average.
- X_{ij} is the covariate matrix for the jth school of the ith state.
- ϵ_{ij} is the random error term for the jth school of the ith state.
- $h(\mu_{ij})$ is the logit function.

`smokeModelT` is a Generalized Linear Mixed Model. The difference between `smokeModelT` and the Generalized Linear Model is that `smokeModelT` contains nested random effects from `state` and `school`. It assumes that the number of students who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days is dependent to each other for students from the same state and school.

The covariates are age (centered around 16 years old), sex (male, female), rural urban area (rural, urban), race (white, black, hispanic, asian, native, pacific). The covariates also include an intersection term for age and sex.

The baseline group includes people who are 16 years old, male, white living in rural area.

Question 2.b

The generalized linear mixed model with a logit link is more appropriate for this dataset than a linear mixed model because the responses are binary, and we are interested in the number of success (number of people who have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days) out of a fixed number of trials (the total number of people within the specific subgroup). A binomial model with logit link fits well under this scenario.

Question 2.c

In the table output using `Pmisc::coefTable(smokeModelT)`, the last two lines of the table represents the standard deviation explained by schools nested within states and by states. From the 95% confidence intervals we can see that the the confidence interval for the standard deviation explained by `school` intersecting `state` is $(0.59, 0.95)$. The 95% confidence interval for standard deviation explained by `state` is $(0.13, 0.74)$. The two confidence intervals do overlap, thus it is hard to tell for certain which difference is larger using confidence interval.

However, with point estimate as the next best metric option, we can see that the point estimate for standard deviation explained by school intersecting state is 0.75, and standard deviation explained by state alone is 0.31. Using point estimates, we conclude that `school` intersecting `state` explains the standard deviation more than using `state` alone.

This claim is also supported by the plots. The plot generated using `state` as the group variable does have a trend and the x axis has a wide range of values. This means that using `state` as a random variable is appropriate. However, the plot generated using `school:state` indicates that there is a greater trend than the x axis is wider than that of the plot output using `state`. Thus using `school` intersecting `state` explains more difference than only using `state`.

The two evidence above indicate that differences between schools within a state in chewing tobacco usage amongst high school students are much larger than state-level differences, contradicting to the hypothesis. This suggests that if one was interested in identifying locations with many tobacco chewers, it would be more important to find individual schools with high chewing rates, rather than just targeting those states where chewing is most common.

Question 3

Question 3.a

A case-control model that correspond to `theGlm` and `theGlmInt` models can be sampled from a pool of patients who are injured in motor vehicle accidents. The case in this study is the group of people who have been fatally injured in motor vehicle accidents. The control in this study is the group of the people who are slightly injured in motor vehicles accidents.

The covariates are the age (26-35, 0-5, 6-10, 11-15, 16-20, 21-25, 36-45, 46-55, 56-65, 66-75 and Over 75), sex (male, female), lighting condition (daylight, lights lit, lights unlit, no lighting, lighting unknown) and weather conditions (Fine no high winds, Raining no high winds, Snowing no high winds, Fine + high winds, Raining + high winds, Snowing + high winds, Fog or mist). The group corresponding to the baseline is the group of people who are male, 26-35 years old who got into motor accidents under daylight lighting condition and under fine no high winds weather condition.

Question 3.b

To address this research question, we would need to use to a model that contains intersection between `sex` and `age` in order to compare the probability of male and female getting fatal injuries when they are teenagers and in early adulthood. Thus `theGlmInt` is more appropriate.

First we look at the Odds Ratio Table. By looking at the `sex` section under `model2`, we see that the 95% confidence interval for the odds of `female` is $(0.53, 0.63)$, which is less than 1. This suggests that with 95% confidence, comparing to males, females have lower odds of receiving fatal injuries.

Then we look at the intersection term in the `sex:age` section in the table. We define young adults to be the age group from 16 to 25 years old. For age between 16 to 20, we see that `Female:16-20` has 95% confidence interval of $(1.03, 1.31)$. This is slightly greater than 1, suggesting that females at the age of $16-20$ have a slightly larger odds of receiving fatal injuries than males at the same age. For age between 21 and 25, we see that `Female:21-25` has 95% confidence interval of $(0.84, 1.10)$. This interval includes 1, thus we cannot conclude whether females at the age of $21-25$ have different probability of receiving fatal injuries than males at the same age.

Combining the two intersection results, we can conclude that the intersections between `sex` and `weight` do not weigh heavily in determing whether females have higher probability of receiving fatal injury.

This is also confirmed by the probability against age graph. Within the age range around $16-25$ years old, males have a higher probability of getting fatal injuries in motor accidents than female do, and the 99% confidence intervals do not overlap.

Combining these two evidence, we can say that the hypothesis that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood is true.

Question 3.c

The control group is not a valid one for assessing whether women are on average better at road safety than man. This is because if the claim that men are less likely than women to report minor injuries caused by road accidents is true, the control group for male will be biased, and will be underrepresented. This violates the assumption that object inclusion in the study doesn't depend on covariates. In this specific case, the sex of the patient is a covariate, and males not seeking medical attention prevents participation in the study. For the control group to be valid, gender should not prevent participation of the study.