Final

Jinhong Du - 12243476

2020/01/03

Contents

1. The first dataset we analyze contains the responses of our second survey about final options. Some students have left comments and we can study if whether someone leaves a comment or not is associated with his/her choice. You can find the dataset in the file folder, the ID information has been removed and the actual comment is replaced by whether a comment is left or not.

(a) (15 points) Convert the score y_{ij} for each option into $y'_{ij} = |y_{ij} - 3|$ for student i and option j. Write down a model for y'_{ij} which takes into account: 1) y'_{ij} across j are choices made by the same student i; 2) values for y'_{ij} are ordered categories; 3) whether someone leaves a comment can have different effects on different final options. Would building a random effect model on the original scores y_{ij} be appropriate? Why? (Hint: check the correlations of the scores for different options across students)

We can use the cumulative logit model with random intercept,

$$\begin{aligned}\operatorname{logit}[\mathbb{P}(y'_{i1} \leq k|u_i)] &= \alpha_k + u_i\\ \operatorname{logit}[\mathbb{P}(y'_{ij} \leq k|u_i)] &= \alpha_k + \beta_j + \beta'_j x_i + u_i,\end{aligned}$$

for $k = 0, 1$ and $j = 2, 3, 4$, where x_i is an indicator of whether `HasComment` is `TRUE` or `FALSE` for student i, and u_i is a random effect. Here we don't include a random slope for x_i since it will result in singular fits with estimated correlations being -1. Equivalently, by treating `Option1`,...,`Option4` as covariates, and rearrange y'_{ij} as $\tilde{y}_i$, we have

$$\begin{aligned}\operatorname{logit}[\mathbb{P}(\tilde{y}_i \leq k|u_i)] = \alpha_k &+ \beta_2 \mathbb{1}_{\{\texttt{Option}=2\}} + \beta_3 \mathbb{1}_{\{\texttt{Option}=3\}} + \beta_4 \mathbb{1}_{\{\texttt{Option}=4\}}\\ &+ \beta'_2 x_i \mathbb{1}_{\{\texttt{Option}=2\}} + \beta'_3 x_i \mathbb{1}_{\{\texttt{Option}=3\}} + \beta'_4 x_i \mathbb{1}_{\{\texttt{Option}=4\}}\\ &+ u_i,\end{aligned}$$

where we treat `Option`=1 as the baseline.

For cumulative logit model with random intercept based on original scores y_{ij}, consider the latent variable $y^*_{ij} = -\beta_j - \beta'_j x_i - u_i + \epsilon_{ij}$ such that $y_{ij} = k$ if $y^*_{ij} \in (\alpha_{k-1}, \alpha_k]$, where ϵ_{ij}'s are independent logistic random variables. So y_{ij} and y^*_{ij} are positively correlated.

$$\begin{aligned}\operatorname{Cov}(y^*_{ij_1}, y^*_{ij_2}) &= \operatorname{Cov}(u_i + \epsilon_{ij_1}, u_i + \epsilon_{ij_2}) = \sigma_u^2 + \mathbb{1}_{\{j_1=j_2\}}\sigma_\epsilon^2\\ \operatorname{Corr}(y^*_{ij_1}, y^*_{ij_2}) &= \frac{\sigma_u^2 + \mathbb{1}_{\{j_1=j_2\}}\sigma_\epsilon^2}{\sigma_u^2 + \sigma_\epsilon^2} > 0.\end{aligned}$$

Therefore, when most pairwise correlation of the scores for different options across students is positive, the random effect model will be appropriate.

Let's look at the correlation between different options, i.e., the correlation between y_{ij} for different j:

```r
df <- read.csv('responses_final_options.csv', header = TRUE, sep = ",")
cor(df[,2:5])
```

```
##             Option1     Option2     Option3     Option4
## Option1  1.00000000 -0.33647738  0.08111134 -0.05292106
## Option2 -0.33647738  1.00000000 -0.56123929  0.06015207
## Option3  0.08111134 -0.56123929  1.00000000  0.02660314
## Option4 -0.05292106  0.06015207  0.02660314  1.00000000
```

As we can see, some pairwise correlation is negative (the smallest -0.62 indicates strongly negatively correlation), which means that the random effect model is unsuitable.

(b) (15 points) Can you use R to fit the above model? If not, you can fit other reasonable models and explain why the other model is reasonable. Compare the results with a model without random effects and comment.
Possibly useful R packages for this problem:

- `reshape2`: http://www.cookbook-r.com/Manipulating_data/Converting_data_between_wide_and_long_format/.
- `anytime`: https://www.displayr.com/r-date-conversion/.

We can use the function `clmm` from the `R` package `ordinal` to fit a cumulative logit model with one or more random effects. Also note that the model fitted by `clmm` is slightly different from our presumed model in (a):

$$\begin{aligned}\operatorname{logit}[\mathbb{P}(\tilde{y}_i \leq k|u_i)] = \alpha_k &- \beta_2 \mathbb{1}_{\{\texttt{Option}=2\}} - \beta_3 \mathbb{1}_{\{\texttt{Option}=3\}} - \beta_4 \mathbb{1}_{\{\texttt{Option}=4\}}\\ &- \beta_2' x_i \mathbb{1}_{\{\texttt{Option}=2\}} - \beta_3' x_i \mathbb{1}_{\{\texttt{Option}=3\}} - \beta_4' x_i \mathbb{1}_{\{\texttt{Option}=4\}}\\ &- u_i.\end{aligned}$$

So the estimated coefficients of fixed effects in our model should be the negative of the estimated coefficients of the output of `clmm`, and so does the estimated random effect.

As we can see, for model in (a) and the one without random effects:

- The estimated coefficients of fixed effects (including α_k) in both models are similar. While most estimated coefficients in the model without random effects tend to shrink to 0 compared to the first model, except for `HasCommentTRUE` and `Option2:HasCommentTRUE`. This may be caused by computational error.
- The estimated standard errors in the first model is a little larger than the ones in the second model, which may be caused by random effects.

In both models, only the hypothesis tests for covariates `Option2` and `Option2:HasCommentTRUE` are significant, which means that

- When keeping x_i fixed, there is a difference of the scores y_{i1}' of option 1 and y_{i2}' of option 2, for the same student i. More specifically, the estimated $\mathbb{P}(y_{i2}' \leq k|u_i)$ is smaller than the estimated $\mathbb{P}(y_{i1}' \leq k|u_i)$ when keeping other covariates fixed, which means that y_{i2}' is more likely to be 2 than y_{i1}', i.e. the number of extreme scores in option 2 is larger than the one in option 1.
- The covariate `HasComment` has an effect on scores of option 2. Students who has comments tend to give more extreme scores in option 2.

```r
library(tidyr)
data <- df[,2:6]
data[,1:4] <- abs(data[,1:4] - 3)
data['Student'] <- 1:length(data[,1])
df2 <- gather(data, Option, y, Option1:Option4, factor_key=TRUE)
df2$Option <- as.factor(regmatches(df2$Option, regexpr("+\\d", df2$Option)))
df2$HasComment <- factor(df2$HasComment)
df2$y <- ordered(factor(df2$y), levels = c(0:2))
library(ordinal)
fm1 <- clmm(y ~ Option*HasComment + (1|Student),
             link='logit', data=df2)
summary(fm1)
```

```
## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: y ~ Option * HasComment + (1 | Student)
## data:    df2
##
```

```
##  link  threshold nobs logLik  AIC    niter     max.grad cond.H
##  logit flexible  248  -154.07 328.13 481(1020) 3.52e-05 1.8e+02
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  Student (Intercept) 0.2056   0.4534
## Number of groups:  Student 62
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## Option2                  1.89256    0.81322   2.327  0.01995 *
## Option3                 -0.18243    0.49366  -0.370  0.71172
## Option4                 -0.01569    0.51144  -0.031  0.97553
## HasCommentTRUE          -0.10142    0.69503  -0.146  0.88398
## Option2:HasCommentTRUE -3.22765    1.10772  -2.914  0.00357 **
## Option3:HasCommentTRUE  0.52741    0.99699   0.529  0.59681
## Option4:HasCommentTRUE  0.25843    1.00831   0.256  0.79772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##     Estimate Std. Error z value
## 0|1  -2.4309     0.4346  -5.594
## 1|2  -1.3075     0.3829  -3.414
```

```r
library(VGAM)
fit2 <- vglm(y ~ Option + HasComment + Option*HasComment,  data=df2,
             family = cumulative(link = "logitlink", parallel = T))
summary(fit2)
```

```
##
## Call:
## vglm(formula = y ~ Option + HasComment + Option * HasComment,
##     family = cumulative(link = "logitlink", parallel = T), data = df2)
##
## Pearson residuals:
##                         Min      1Q  Median       3Q   Max
## logitlink(P[Y<=1]) -1.450 -0.2097 -0.1932 -0.08028 7.955
## logitlink(P[Y<=2]) -1.025 -0.4989 -0.4989 -0.19573 5.704
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept):1           -2.344e+00  3.880e-01  -6.042 1.52e-09 ***
## (Intercept):2           -1.251e+00  3.486e-01  -3.588 0.000333 ***
## Option2                 -1.856e+00  8.001e-01  -2.320 0.020366 *
## Option3                  1.848e-01  4.801e-01   0.385 0.700291
## Option4                  2.199e-15  4.922e-01   0.000 1.000000
## HasCommentTRUE           1.037e-01  6.921e-01   0.150 0.880841
## Option2:HasCommentTRUE  3.159e+00  1.111e+00   2.844 0.004461 **
## Option3:HasCommentTRUE -5.153e-01  1.012e+00  -0.509 0.610738
## Option4:HasCommentTRUE -2.522e-01  1.008e+00  -0.250 0.802322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 308.4731 on 487 degrees of freedom
##
## Log-likelihood: -154.2366 on 487 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##                Option2                Option3                Option4
##              0.1563328              1.2029770              1.0000000
##         HasCommentTRUE Option2:HasCommentTRUE Option3:HasCommentTRUE
##              1.1093205             23.5446363              0.5973303
## Option4:HasCommentTRUE
##              0.7770508
```

2. Everyone is talking about coronavirus, and you may be following the news everyday to see how it spreads across world. Let's analyze the public coronavirus data ourselves and try to get more insights than what have learnt from the media.

Background:

People are sharing Figure 1 on twitter where the figure comes from this report: https://medium.com/@andreasbackhausab/coronavirus-why-its-so-deadly-in-italy-c4200a15a7bf . This figure compares between Italy and South Korea. It catches people's eyes because this may show that there can be a lot of infected young people in Italy that are missed as they show no or mild symptoms, and they may spread the virus to others unconsciously. As pointed out in the source report, *Italy has predominantly been testing people with symptoms of a coronavirus infection, while South Korea has been testing basically everyone since the outbreak had become apparent. Consequently, South Korea has detected more asymptomatic, but positive cases of coronavirus than Italy, in particular among young people.* However, looking at Figure 1 alone is not enough to get the above conclusion as the age structure in the two countries are also different. Even if both two countries can diagnoses every effected case, you would see a higher fraction of elder people diagnosed in Italy than in South Korea since Italy has one of the oldest populations in the world.

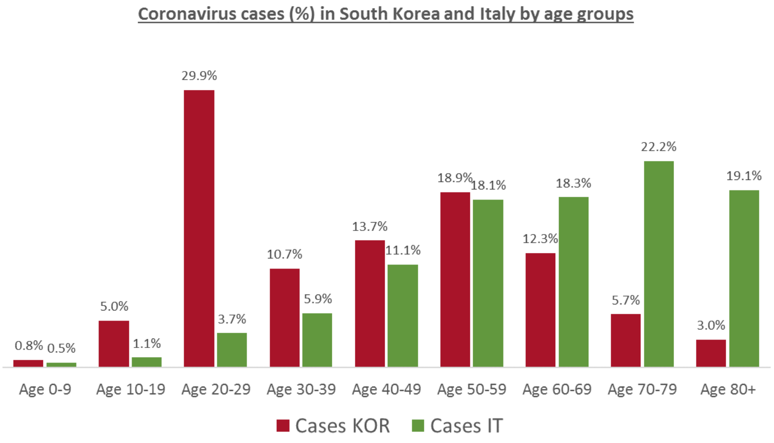

Figure 1

Can we show if in Italy elder people are more easily diagnosed as they have more severe symptoms? If Italy has missed a lot of younger patients who have no or milder symptoms, can we give an estimate of how many people are actually infected in Italy at this moment?

Datasets:

Patient level information in South Korea can be downloaded here: https://www.kaggle.com/kimjihoo/coronavirusdataset. In the file folder on Canvas, you can find the data file I downloaded which is updated to 3/11/2020. This is one of the dataset we are going to use in this problem, and I download it so that everyone has the same dataset to analyze.

Patient level information in Italy is not available, so one dataset we are going to use is the region level data by date, which was downloaded from this kaggle website: https://www.kaggle.com/sudalairajkumar/covid19-in-italy. You can also find it in the file folder on Canvas, which is updated to 3/14/2020. The age information for patients in Italy is hard to find, so we use the proportions in Figure 1 for Italy and assume that these proportions remain the same on day 3/14/2020.

The age structure of the Italy and South Korea population can be found here: https://www.populationpyramid.net/world/2019/. Here are a few lines of R code that you can use to download them:

```
library(data.table)
italy.pop <- fread("https://www.populationpyramid.net/api/pp/380/2019/?csv=true")
south.korea.pop <- fread("https://www.populationpyramid.net/api/pp/410/2019/?csv=true")
```

For all the above datasets, if you do not understand any of them, you can refer to the source websites where you can easily find explanations.

Problems:

Let's simplify our analysis by dividing people into four age groups: 0-19, 20-49, 50-69 and 70+ years old. The plots in Figure 2 are what I generated using the datasets mentioned above.

The left is similar to Figure 1, which compares the current proportions of diagnosed among the four age groups (Italy: 3/14/2020, Korea: 3/11/2020). The right compares the age structure in the whole population between the two countries.

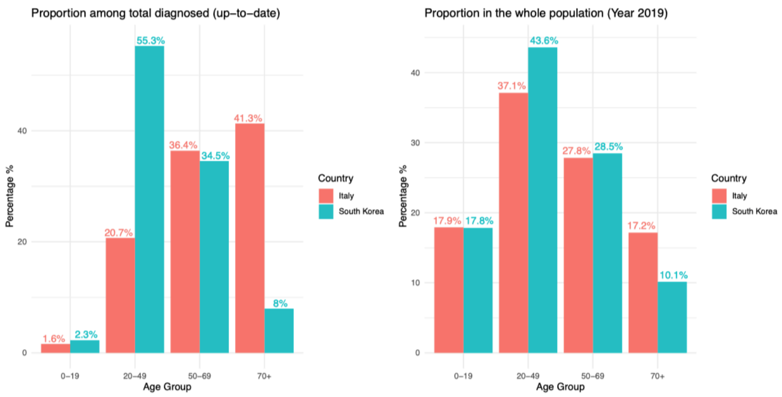

Figure 2

(a) (20 points) Build a logistic regression to study the country effects on the diagnosis probabilities across age groups. You can use the relationship below

$$\begin{aligned}&\text{Current number of people diagnosed in country } i \text{ and age group } j\\=&\text{Total population in country } i \text{ and group } j\times\\&\text{Proportion being infected and diagnosed in country } i \text{ and age group } j\end{aligned}$$

Notice that the infection and diagnosis rates in different countries can be different, and can also vary across age groups.

Write down the model and fit it in `R`. Then test whether the country effects on the diagnosis probability vary across age groups.

Let $y_{ij} \sim \text{Binomial}(n_{ij}, p_{ij})$ be the number of people diagnosed in country i and age group j for $i = 1, 2$ (denoting Italy and South Korea, respectivley) and $j = 1, 2, 3, 4$ (denoting 0-19, 20-49, 50-69 and 70+ years old, respectivley), where n_{ij} is the total population in country i and age group j and p_{ij} is the theoretical proportion being infected and diagnosed in country i and age group j. So the model is given by

$$\begin{aligned}\text{logit}(p_{ij}) = &\beta_0 + \beta_{c2}\mathbb{1}_{i=2} + \beta_{a2}\mathbb{1}_{j=2} + \beta_{a3}\mathbb{1}_{j=3} + \beta_{a4}\mathbb{1}_{i=4}\\&+ \beta_{ca2}\mathbb{1}_{i=2}\mathbb{1}_{j=2} + \beta_{ca3}\mathbb{1}_{i=2}\mathbb{1}_{j=3} + \beta_{ca4}\mathbb{1}_{i=2}\mathbb{1}_{i=4}.\end{aligned}$$

We add the interaction terms since the infection and diagnosis rates in different countries can be different, and can also vary across age groups.

Next we fit this model in `R`:

```r
# Computes population in different groups
library(data.table)
italy.pop <- fread("https://www.populationpyramid.net/api/pp/380/2019/?csv=true")
south.korea.pop <- fread("https://www.populationpyramid.net/api/pp/410/2019/?csv=true")
is_in <- function(x){
    ages <- strsplit(x, '-', fixed=TRUE)[[1]]
    if(length(ages)==1){
        # '100+' is in the group 4
        return(4)
    }
    l <- matrix(c(0,19,20,49,50,69,70,100), nrow=2)

    for(i in 1:4){
        if((as.numeric(ages[1])>=l[1,i]) & (as.numeric(ages[2])<=l[2,i]))
            return(i)
    }
}
italy.pop[['group']] = apply(italy.pop[,'Age'], 1, is_in)
italy.pop <- aggregate(italy.pop[,c('M', 'F')], by=list(group=italy.pop[['group']]), FUN=sum)
italy.pop.age_count <- apply(italy.pop[,c('M', 'F')], 1, FUN=sum)
south.korea.pop[['group']] = apply(south.korea.pop[,'Age'], 1, is_in)
south.korea.pop <- aggregate(south.korea.pop[,c('M', 'F')],
                             by=list(group=south.korea.pop[['group']]), FUN=sum)
south.korea.pop.age_count <- apply(south.korea.pop[,c('M', 'F')], 1, FUN=sum)

# Compute numebrs of diagnosed patients
italy.patient <- read.csv('covid19_italy_region.csv', header = TRUE, sep = ",")
italy.patient <- italy.patient[italy.patient['Date']=='2020-03-14 17:00:00',]
italy.diag_count <- round(c(1.6, 20.7, 36.4, 41.3)/100*
                              sum(italy.patient$TotalPositiveCases))

south.korea.patient <- read.csv('patient_south_korea.csv', header = TRUE, sep = ",")
is_in <- function(x){
    l <- matrix(c(0,19,20,49,50,69,70,100), nrow=2)
    for(i in 1:4){
        if((as.numeric(x)>=l[1,i]) & (as.numeric(x)<=l[2,i]))
            return(i)
    }
}
south.korea.patient <- south.korea.patient
south.korea.total_patient <- dim(south.korea.patient)[1]
south.korea.patient <- south.korea.patient[!is.na(south.korea.patient$birth_year),]
south.korea.diag_age <- apply(2020-south.korea.patient['birth_year'], 1, is_in)
south.korea.diag_count <- aggregate(south.korea.diag_age,
                                   by=list(group=south.korea.diag_age), FUN=length)[,'x']
south.korea.diag_count <- round(south.korea.diag_count/sum(south.korea.diag_count)
                            *south.korea.total_patient)

counts <- matrix(append(c(italy.diag_count, south.korea.diag_count),
                        c(italy.pop.age_count-italy.diag_count,
                          south.korea.pop.age_count-south.korea.diag_count)),ncol=2)
df <- data.frame(
           country=factor(rep(c(1,2),each=4)),
           age=factor(rep(c(1:4), 2)))
fit <- glm(counts~country*age, data=df, family=binomial())
summary(fit)
```

```
##
## Call:
```

```
## glm(formula = counts ~ country * age, family = binomial(), data = df)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                Estimate Std. Error  z value Pr(>|z|)
## (Intercept)   -10.37308    0.05431 -190.986  < 2e-16 ***
## country2       -0.41195    0.09078   -4.538 5.68e-06 ***
## age2            1.83056    0.05638   32.470  < 2e-16 ***
## age3            2.68312    0.05550   48.348  < 2e-16 ***
## age4            3.29197    0.05536   59.467  < 2e-16 ***
## country2:age2   0.38342    0.09331    4.109 3.97e-05 ***
## country2:age3  -0.45559    0.09342   -4.877 1.08e-06 ***
## country2:age4  -1.49293    0.09947  -15.009  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  1.9029e+04  on 7  degrees of freedom
## Residual deviance: -1.1916e-10  on 0  degrees of freedom
## AIC: 90.944
##
## Number of Fisher Scoring iterations: 4
```

As we can see:

- The Wald test for $\beta_{c2} = \hat{\beta}_{c2}$ is very significant, which menas that the country effects exist.
- The Wald tests for $\beta_{aj} = \hat{\beta}_{aj}$ $(j = 2, 3, 4)$ are very significant, which menas that when keeping the country fixed, the probability of infected and diagnosed in group j $(j = 2, 3, 4)$ is large than in group 1. Moreover, the effect gets larger as the age becomes larger.
- For the interaction terms, $\beta_{caj} = \hat{\beta}_{caj}$ is significant for $j = 2, 3, 4$. Compared to age group 1, elder people in age group 2, 3 and 4 are more likely to be infected in Italy than in South Korea. What's more, diagnosis probability decreases more for groups with older people in the same age group, comparing those in South Korea to those in Italy. So the country effects on the diagnosis probability vary across age groups.

(b) (20 points) One can intuitively account for the population age structure difference when intepreting the results in the left plot of Figure 2. For example, for age group 70+, the diagnosed coronavirus case proportions ratio between Italy and Korea is about 5, however, that ratio drops to 1.7 when comparing the total population. Thus, the diagnosed cases ratio is three times of the whole population ratio for people with 70+, showing that Italy diagnoses more elder people. Give interpretations of the logistic regression coefficients. Specifically, quantitatively illustrate how the estimates of the country effect coefficients in your logistic model correspond to the ratios mentioned in the above intuitive illustration for Figure 2. (Hint: for the quantitative illustration, p use the fact that $\frac{p}{1-p} \approx p$ when p is close to 0)

Since p_{ij} the proportion being diagnosed is smalle, we have

$$\begin{aligned}\log(p_{ij}) \approx \operatorname{logit}(p_{ij}) &= \beta_0 + \beta_{c2}\mathbb{1}_{i=2} + \beta_{a2}\mathbb{1}_{j=2} + \beta_{a3}\mathbb{1}_{j=3} + \beta_{a4}\mathbb{1}_{i=4} \\ &\quad + \beta_{ca2}\mathbb{1}_{i=2}\mathbb{1}_{j=2} + \beta_{ca3}\mathbb{1}_{i=2}\mathbb{1}_{j=3} + \beta_{ca4}\mathbb{1}_{i=2}\mathbb{1}_{i=4} \\ p_{ij} &\approx e^{\beta_0 + \beta_{c2}\mathbb{1}_{i=2} + \beta_{a2}\mathbb{1}_{j=2} + \beta_{a3}\mathbb{1}_{j=3} + \beta_{a4}\mathbb{1}_{i=4}} \\ &\quad \times e^{\beta_{ca2}\mathbb{1}_{i=2}\mathbb{1}_{j=2} + \beta_{ca3}\mathbb{1}_{i=2}\mathbb{1}_{j=3} + \beta_{ca4}\mathbb{1}_{i=2}\mathbb{1}_{i=4}}.\end{aligned}$$

As we can see from the results in (a):

- When keeping other covariates fixed, the probability of being infected and diagnosed in South Korea is about $e^{\hat\beta_{c2}+\hat\beta_{caj}}$ more than the one in Italy in age group j.
- Compared to age group 1, the probability of being infected and diagnosed in group j is about $e^{\hat\beta_{aj}}$ more than the one in age group 1 in Italy. Similarly, the probability of being infected and diagnosed in group j is about $e^{\hat\beta_{aj}+\hat\beta_{caj}}$ more than the one in age group 1 in South Korea.

The ratio in the left plot for age group 4 can be computed by

$$\frac{\frac{\hat p_{14}}{\sum_{j=1}^4 \hat p_{1j}}}{\frac{\hat p_{24}}{\sum_{j=1}^4 \hat p_{2j}}} \approx \frac{\frac{e^{\beta_{a4}}}{\sum_{j=1}^4 e^{\beta_{aj}}}}{\frac{e^{\beta_{a4}+\beta_{ca4}}}{\sum_{j=1}^4 e^{\beta_{aj}+\beta_{caj}}}} = \frac{e^{-\beta_{ca4}}\sum_{j=1}^4 e^{\beta_{aj}}}{\sum_{j=1}^4 e^{\beta_{aj}+\beta_{caj}}} = 4.917024 \approx 5$$

when we only look at the proportion of each age group only in diagnosed cases. However, when we take the age stucture in the total population into consideration, we will get

$$\frac{\hat\pi_{24}}{\hat\pi_{14}} \times \frac{\frac{\hat p_{14}}{\sum_{j=1}^4 \hat p_{1j}}}{\frac{\hat p_{24}}{\sum_{j=1}^4 \hat p_{2j}}} = 2.902712 \approx 3,$$

where $\hat\pi_{ij}$ is the sample proportion of number of people in age group j in country i.

```
cat(exp(-fit$coefficients[8]) / sum(1+exp(fit$coefficients[c(3,4,5)])) *
        sum(1+exp(fit$coefficients[c(3:8)])))
```

```
## 4.917024
```

```
pi1 <- italy.pop.age_count/sum(italy.pop.age_count)
pi2 <- south.korea.pop.age_count/sum(south.korea.pop.age_count)
cat(pi2[4]/pi1[4]*exp(-fit$coefficients[8]) / sum(1+exp(fit$coefficients[c(3,4,5)])) *
        sum(1+exp(fit$coefficients[c(3:8)])))
```

```
## 2.902712
```

(c) (20 points) Use the decomposition below:

$$\begin{aligned}&\text{Probability of being infected and diagnosed in country } i \text{ and age group } j\\ =&\text{Chance of getting infected in country } i \text{ for age group } j\times\\ &\text{The diagnosis ability (rate) in country } i \text{ for age group } j\end{aligned}$$

and further assume that:

- $\frac{\text{The infection rate in South Korea and group } j}{\text{The infection rate in Italy and group } j}$ is the same across j.
- South Korea has tested so many people so that almost everyone infected have been diagnosed. (The diagnosis ability in South Korea for group $j = 1$)
- Italy is able to diagnose every infected people for age 70 and above. (The diagnosis ability in Italy for age group $70+ = 1$)

Give an estimate of the total number of people infected by coronavirus in Italy at 3/14/2020. Also, use delta method to estimate the standard deviation of the estimated number of infected. (Hint: again you can simplify your calculation using the fact that $\frac{p}{1-p} \approx p$ when p is close to 0)

Let p_{ij} be the probability of being infected and diagnosed in country i and age group j, q_{ij} be the chance of getting infected in country i for age group j and r_{ij} be the diagnosis ability (rate) in country i for age group j. Then $p_{ij} = q_{ij}r_{ij}$.

From the assumptions, we have

- $\frac{q_{21}}{q_{11}} = \frac{q_{22}}{q_{12}} = \frac{q_{23}}{q_{13}} = \frac{q_{24}}{q_{14}}$.
- $r_{2j} = 1$ for all j.
- $r_{14} = 1$.

So $q_{2j} = p_{2j}$ for all j and $q_{14} = p_{14}$. Then $\frac{q_{2j}}{q_{1j}} = \frac{q_{24}}{q_{14}} = \frac{p_{24}}{p_{14}}$ and $q_{1j} = \frac{p_{2j}p_{14}}{p_{24}}$ for all j. Let n_{ij} be the number of people in age group j in country i. Then the total number of people infected by coronavirus in Italy at 3/14/2020 is $\sum_{j=1}^{4} n_{1j}q_{1j}$. For real data, we use the estimated probability to compute it by $\sum_{j=1}^{4} n_{1j}\hat{q}_{1j}$.

```r
p <- fitted(fit)
p1 <- p[1:4]
p2 <- p[5:8]
q1 <- p2*p1[4]/p2[4]
est_total <- sum(q1 * italy.pop.age_count)
round(est_total)
```

```
## [1] 60539
```

To compute the standard error, consider

$$\hat{q}_{1j} = \frac{\hat{p}_{2j}\hat{p}_{14}}{\hat{p}_{24}} \approx \begin{cases} e^{\hat{\beta}_0 - \hat{\beta}_{ca4}} & , \ j = 1 \\ e^{\hat{\beta}_0 + \hat{\beta}_{aj} + \hat{\beta}_{caj} - \hat{\beta}_{ca4}} & , \ j = 2, 3, 4 \end{cases}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) \approx (\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})^{-1}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_{c2}, \hat{\beta}_{a2}, \hat{\beta}_{a3}, \hat{\beta}_{a4}, \hat{\beta}_{ca2}, \hat{\beta}_{ca3}, \hat{\beta}_{ca4})^\top \in \mathbb{R}^8$, $\boldsymbol{X} \in \mathbb{R}^{8\times 8}$ is the design matrix and $\hat{\boldsymbol{W}} = \mathrm{diag}(n_{ij}\hat{p}_{ij}(1 - \hat{p}_{ij}))$.

Let $f(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{4} n_{1j}\hat{q}_{1j}$. Then

$$\begin{array}{lll} \dfrac{\partial f}{\partial \hat{\beta}_0} = \displaystyle\sum_{j=1}^{4} n_{1j}\hat{q}_{1j} & \dfrac{\partial f}{\partial \hat{\beta}_{c2}} = 0 & \dfrac{\partial f}{\partial \hat{\beta}_{ca4}} = -\displaystyle\sum_{j=1}^{3} n_{1j}\hat{q}_{1j} \\ \dfrac{\partial f}{\partial \hat{\beta}_{aj}} = n_{1j}\hat{q}_{1j} \qquad j = 2,3,4 & & \dfrac{\partial f}{\partial \hat{\beta}_{caj}} = n_{1j}\hat{q}_{1j}, \qquad j = 2,3 \end{array}$$

The variance for the total number of people infected in Italy at 3/14/2020 is given by

$$\mathrm{Var}\left(\sum_{j=1}^{4} n_{1j}\hat{q}_{1j}\right) = \nabla f^{\top}\mathrm{Var}(\hat{\boldsymbol{\beta}})\nabla f$$

```
X <- model.matrix(fit, df)
nabla_f <- matrix(c(est_total, 0, q1[-1] * italy.pop.age_count[-1],
                    q1[c(2,3)] * italy.pop.age_count[c(2,3)],
                    -sum(q1[-4] * italy.pop.age_count[-4])), ncol=1)
sqrt(t(nabla_f) %*% solve(t(X) %*% diag(c(italy.pop.age_count, south.korea.pop.age_count)
        * p * (1-p)) %*% X) %*% nabla_f)
```

```
##          [,1]
## [1,] 2218.571
```

(d) (10 points) Using the logistic regression, we implicitly assume that people inside one age group in one country are i.i.d.. However, this is not true as (1) the virus spread from person to person, so the probability of someone get infected is correlated with people around (2) people inside the same age group have heterogeneity, for instance the chance of getting infected for a 70-year-old can be different from that for a 90-year-old, though they are in the same group. Discuss how the sample correlation and heterogeneity can affect our estimates in (a) and (c). Provide potential solutions if you can.

The sample correlation and heterogeneity will result in more variability, which we called overdispersion.

Let y_{ijk} for $k = 1, \ldots, n_{ij}$ be ungrouped Bernoulli random variables that constitute the group data y_{ij}. If $\mathrm{Corr}(y_{ijs}, y_{ijt}) = \rho > 0$ for $s \neq t$, we have

$$\begin{aligned}\mathrm{Var}\left(y_{ij}\right) &= \mathrm{Var}\left(\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk}\right) = \frac{1}{n_{ij}^2}\left[\sum_{k=1}^{n_{ij}} \mathrm{Var}\left(y_{ijk}\right) + 2\sum_{s<t}\mathrm{Cov}\left(y_{is}, y_{it}\right)\right] \\ &= \frac{1}{n_{ij}^2}\left[n_{ij}p_{ij}\left(1-p_{ij}\right) + n_{ij}\left(n_{ij}-1\right)\rho p_{ij}\left(1-p_{ij}\right)\right] = \left[1+\rho\left(n_{ij}-1\right)\right]\frac{p_{ij}\left(1-p_{ij}\right)}{n_{ij}},\end{aligned}$$

which is larger than the variance $\frac{p_{ij}(1-p_{ij})}{n_{ij}}$ we assumeed. So if positive sample correlation exists, our estimated variance will tend to get larger than the theoretical variance of our presumed model.

Heterogeneity results in an overall response distribution at that weight having greater variation than the Binomial. So it also cause overdispersion.

Some possible solutions are:

- To use mixture models, e.g. the beta-binomial model $y|p \sim \mathrm{Binomial}(n, p)$ and $p \sim \mathrm{Beta}(\alpha_1, \alpha_2)$, a mixture of Binomial distribution.
- To use quasi-likelihood methods, by assuming a suitable mean-variance relationship.
- To use GLMMs, model the variability across person/group as random effects.