The model I fitted is

$$\text{logit}(P(Y_{ij}=1)) = B_0 + B_i^C \mathbb{1}_{\{country = Korea\}} + B_j^A \mathbb{1}_{\{age\ group\}} + B_{ij}^{AC} \mathbb{1}_{\{age\ group\}} \mathbb{1}_{\{Korea\}}$$

$i = 1$, Korea

$$j = \begin{cases} 2, & 20-49 \\ 3, & 50-69 \\ 4, & 70+ \end{cases}$$

This is a saturated model, I chose to fit this model because I was unable to fit the ungrouped data model, which will solve the degree of freedom issue, due to limited memory on my computer.

Therefore, the model I fitted is an overfit.

Model estimates are: From R

$\hat{B}_0 = -10.40303$

$\hat{B}_1^C = -0.36626$

$\hat{B}_2^A = 1.83130$

$\hat{B}_3^A = 2.68367$

$\hat{B}_4^A = 3.29251$

$\hat{B}_{12}^{AC} = 0.36361$

$\hat{B}_{13}^{AC} = -0.46654$

$\hat{B}_{14}^{AC} = -1.51384$

```r
#loading the data
i_pop <- fread("https://www.populationpyramid.net/api/pp/380/2019/?csv=true")
k_pop <- fread("https://www.populationpyramid.net/api/pp/410/2019/?csv=true")

sk_patient <- read_csv("C:/Users/Jun Jie Choo/Desktop/patient_south_korea.csv")
```

```
## Parsed with column specification:
## cols(
##   patient_id = col_double(),
##   sex = col_character(),
##   birth_year = col_double(),
##   country = col_character(),
##   region = col_character(),
##   disease = col_double(),
##   group = col_character(),
##   infection_reason = col_character(),
##   infection_order = col_double(),
##   infected_by = col_double(),
##   contact_number = col_double(),
##   confirmed_date = col_date(format = ""),
##   released_date = col_date(format = ""),
##   deceased_date = col_date(format = ""),
##   state = col_character()
## )
```

```r
sk_patient2 = sk_patient
i_patient <- read_csv("C:/Users/Jun Jie Choo/Desktop/covid19_italy_region.csv")
```

```
## Parsed with column specification:
## cols(
##   SNo = col_double(),
##   Date = col_datetime(format = ""),
##   Country = col_character(),
##   RegionCode = col_double(),
##   RegionName = col_character(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   HospitalizedPatients = col_double(),
##   IntensiveCarePatients = col_double(),
##   TotalHospitalizedPatients = col_double(),
##   HomeConfinement = col_double(),
##   CurrentPositiveCases = col_double(),
##   NewPositiveCases = col_double(),
##   Recovered = col_double(),
##   Deaths = col_double(),
##   TotalPositiveCases = col_double(),
##   TestsPerformed = col_double()
## )
```

```r
#Cleaning data
#I noticed there are a lot of missing values in birth year

sum(is.na(sk_patient$birth_year))/nrow(sk_patient) #proportion of missing data
```

```
## [1] 0.9153641
```

```
nrow(sk_patient)*0.085 #number of non missing data
```

```
## [1] 668.865
```

```
#we have roughly 650 datapoints without missing birth year, which is enough to fit our model
#we assume missing values are not missing not at random (MNAR) and remove the na values
#this is a reasonable assumption because birth year missingness doesn't depend on itself

#removing na values
sk_patient = sk_patient[(!is.na(sk_patient$birth_year)),]


#removing non koreans from dataset
sk_patient = sk_patient[(sk_patient$country=="Korea"),]


#keeping only latest date in italian dataset
i_patient$Date = as.Date(i_patient$Date)
i_patient = i_patient[(i_patient$Date=="2020-03-14"),]
```

Setting up the covariates and response for model

```
#south korea pop
sk_patient$age = 2020 - sk_patient$birth_year #calculating age for korea


n1 =   (sum(k_pop$M[1:4])+sum(k_pop$F[1:4]))
n2 =   (sum(k_pop$M[5:10])+sum(k_pop$F[5:10]))
n3 =   (sum(k_pop$M[11:14])+sum(k_pop$F[11:14]))
n4 =   (sum(k_pop$M[15:21])+sum(k_pop$F[15:21]))

y1 = ceiling(sum(sk_patient$age <= 19)/nrow(sk_patient) * nrow(sk_patient2))
y2 = ceiling(sum(sk_patient$age > 19 & sk_patient$age <= 49)/nrow(sk_patient) * nrow(sk_patient2))
y3 = ceiling(sum(sk_patient$age > 49 & sk_patient$age <= 69)/nrow(sk_patient) * nrow(sk_patient2))
y4 = ceiling(sum(sk_patient$age > 69)/nrow(sk_patient) * nrow(sk_patient2))


#italy pop
m1 =  (sum(i_pop$M[1:4])+sum(i_pop$F[1:4]))
m2 =  (sum(i_pop$M[5:10])+sum(i_pop$F[5:10]))
m3 =  (sum(i_pop$M[11:14])+sum(i_pop$F[11:14]))
m4 =  (sum(i_pop$M[15:21])+sum(i_pop$F[15:21]))

z1 =  round(0.016*(sum(i_patient$CurrentPositiveCases)+sum(i_patient$NewPositiveCases)))
z2 =  round(0.207*(sum(i_patient$CurrentPositiveCases)+sum(i_patient$NewPositiveCases)))
z3 =  round(0.364*(sum(i_patient$CurrentPositiveCases)+sum(i_patient$NewPositiveCases)))
z4 =  round(0.413*(sum(i_patient$CurrentPositiveCases)+sum(i_patient$NewPositiveCases)))


diagnosed = t(t(c(y1,y2,y3,y4,z1,z2,z3,z4)))
undiagnosed = t(t(c(n1-y1,n2-y2,n3-y3,n4-y4,m1-z1,m2-z2,m3-z3,m4-z4)))
```

```r
country = t(t(rep(c("korea","italy"),times=c(4,4))))
age_group = t(t((c("0-19","20-49","50-69","70+","0-19","20-49","50-69","70+"))))
```

```r
#Fitting the model

#fitting the model with interactions
model <- glm(matrix(append(diagnosed,undiagnosed),ncol=2) ~ factor(country)+factor(age_group)+factor(cou

summary(model)
```

```
##
## Call:
## glm(formula = matrix(append(diagnosed, undiagnosed), ncol = 2) ~
##     factor(country) + factor(age_group) + factor(country):factor(age_group),
##     family = binomial)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                                             Estimate Std. Error  z value
## (Intercept)                                -10.40303    0.05513 -188.691
## factor(country)korea                        -0.36626    0.09082   -4.033
## factor(age_group)20-49                       1.83130    0.05723   32.001
## factor(age_group)50-69                       2.68367    0.05633   47.639
## factor(age_group)70+                         3.29251    0.05619   58.594
## factor(country)korea:factor(age_group)20-49  0.36361    0.09338    3.894
## factor(country)korea:factor(age_group)50-69 -0.46654    0.09347   -4.991
## factor(country)korea:factor(age_group)70+   -1.51384    0.09956  -15.206
##                                             Pr(>|z|)
## (Intercept)                                  < 2e-16 ***
## factor(country)korea                        5.51e-05 ***
## factor(age_group)20-49                       < 2e-16 ***
## factor(age_group)50-69                       < 2e-16 ***
## factor(age_group)70+                         < 2e-16 ***
## factor(country)korea:factor(age_group)20-49 9.87e-05 ***
## factor(country)korea:factor(age_group)50-69 6.00e-07 ***
## factor(country)korea:factor(age_group)70+    < 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:  1.8291e+04  on 7  degrees of freedom
## Residual deviance: -1.4313e-10  on 0  degrees of freedom
## AIC: 90.839
##
## Number of Fisher Scoring iterations: 4
```

Yes. Country effects on diagnosis probability vary across age group as seen by the country:age_group effects.

Let π_{ij} be diagnosis probability of country i, age group j

$i = 1$, Korea

$j = \begin{cases} 2, & 20-49 \\ 3, & 50-69 \\ 4, & 70+ \end{cases}$

Then $\dfrac{\hat{\pi}_{01}}{\hat{\pi}_{11}} \simeq \dfrac{\dfrac{\hat{\pi}_{01}}{1-\hat{\pi}_{01}}}{\dfrac{\hat{\pi}_{11}}{1-\hat{\pi}_{11}}}$

$$= \frac{\exp(\hat{B}_0)}{\exp(\hat{B}_0 + \hat{B}_1^C)}$$

$$= \exp(-\hat{B}_1^C) = \exp(-0.36626) = 1.44233$$

(from R 2a)

0-19 Diagnosed case ratio $= \dfrac{\hat{\pi}_{01}}{\hat{\pi}_{11}} \div \dfrac{17.9}{17.8} = 1.434 > 0.69 = \dfrac{1.6}{2.3} \div \dfrac{17.9}{17.8}$

(Figure 1 + Figure 2)

Similarly, $\dfrac{\hat{\pi}_{02}}{\hat{\pi}_{12}} \simeq \dfrac{\exp(\hat{B}_0 + \hat{B}_2^A)}{\exp(\hat{B}_0 + \hat{B}_1^C + \hat{B}_2^A + \hat{B}_{12}^{AC})}$

$$= \exp(-\hat{B}_1^C - \hat{B}_{12}^{AC})$$

$$= 1.002654$$

20-49 Diagnosed case ratio $0.208 \div \dfrac{37.1}{43.6} = 1.1752 > 0.44$

$$\frac{\hat{\pi}_{03}}{\hat{\pi}_{13}} = \frac{\exp(\hat{B}_0 + \hat{B}_3^A)}{\exp(\hat{B}_0 + \hat{B}_1^C + \hat{B}_3^A + \hat{B}_{13}^{AC})}$$

$$= \exp(-\hat{B}_1^C - \hat{B}_{13}^{AC})$$

$$= 2.3$$

$$2.3 \doteq \frac{27.8}{28.5} = 2.358 \geq \frac{36.4}{34.5} \doteq \frac{27.8}{28.5} = 1.08$$

$$\frac{\hat{\pi}_{04}}{\hat{\pi}_{14}} = \frac{\exp(\hat{B}_0 + \hat{B}_4^A)}{\exp(\hat{B}_0 + \hat{B}_1^C + \hat{B}_4^A + \hat{B}_{14}^{AC})}$$

$$= \exp(-\hat{B}_1^C - \hat{B}_{14}^{AC})$$

$$= 6.55416$$

$$6.55416 \doteq \frac{17.2}{10.1} = 3.85 \geq 3$$

Diagnose case ratio for 70+

$\exp(-\hat{B}_i^C - \hat{B}_{ij}^{AC})$ Logistic regression coefficients are the ratio of proportion diagnosed of age group j.

$$\frac{\pi_{1j}}{1-\pi_{1j}} \approx \pi_{1j} = P(\text{Korean Infect and Diag group } j)$$

$= P(\text{Korean Infect group } j)$ since $P(\text{KR diag}) = 1, \forall j$

Similarly, $P(\text{Italian Infect and Diag group } 4)$
$= P(\text{Italian Infect group } 4) = \pi_{04}$

$$\Rightarrow \frac{\pi_{04}}{\pi_{14}} = c \quad \leftarrow \text{constant}$$

Subby R values for π_{14}, π_{04}:

$$\frac{\hat{\pi}_{04}}{\hat{\pi}_{14}} \approx 6.55416 = c$$

Then use c to find Italy infection rates for groups 1, 2, 3

$P(\text{Italy infect rate group } 3)$
$= P(\text{Korean infect rate group } 3) \times c$
$= \hat{\pi}_{13} \times 6.55416$ (since $\text{Prob}(\text{KR diag}) = 1$
$\approx \exp(\hat{B}_0 + \hat{B}_1^c + \hat{B}_3^c + \hat{B}_{13}^{0c}) \times 6.55416$

$= 0.00126592$

P(Italy infect rate group 2)

$= \exp(\hat{B}_0 + \hat{B}_1^C + \hat{B}_2^A + \hat{B}_{12}^{AC}) \times c$

$= 0.0012 37978$

P(Italy infect rate group 1)

$= \exp(\hat{B}_0 + \hat{B}_1^C) \times c$

$= 0.0001378717$

P(Italy infect rate group 4)

$= \exp(\hat{B}_0 + \hat{B}_1^C + \hat{B}_4^A + \hat{B}_{14}^{AC}) \times c$

$= 0.0008164676$

Total infected in Italy:

$= \sum_{j=1}^{4} P(\text{Infect rate group } j) \times \text{Total Pop Italy group } j$

$= 59114.71$

2c) $\sqrt{n}\left(h(\hat{B}) - h(B^*)\right) \rightarrow N\left(0,\ \dot{h}(B^*)^T V_{B_0} \dot{h}(B^*)\right)$

$h(B) = \pi_{01} \cdot I_1 + \pi_{02} I_2 + \pi_{03} I_3 + \pi_{04} I_4$

where I_j is total pop of Italy in age group j.

$\simeq e^{B_0} I_1 + e^{B_0 + B_2^A} I_2 + e^{B_0 + B_3^A} I_3 + e^{B_0 + B_4^A} I_4$

1x8 vector

$$\dot{h}(B) = \begin{pmatrix} e^{B_0} I_1 + e^{B_0 + B_2^A} I_2 + e^{B_0 + B_3^A} I_3 + e^{B_0 + B_4^A} I_4 \\ 0 \\ e^{B_0 + B_2^A} I_2 \\ e^{B_0 + B_3^A} I_3 \\ e^{B_0 + B_4^A} I_4 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

```r
    find estimated infection and std error
#getting pi_ij
p11 = exp(coef(model)[1] + coef(model)[2] ) * 6.55416
p12 = exp(coef(model)[1] + coef(model)[2] + coef(model)[3] + coef(model)[6] ) * 6.55416
p13 = exp(coef(model)[1] + coef(model)[2] + coef(model)[4] + coef(model)[7]) * 6.55416
p14 = exp(coef(model)[1] + coef(model)[2] + coef(model)[5] + coef(model)[8]) * 6.55416

#estimate of the total number of people infected by coronavirus in Italy
estimated_infection = p11 * m1 + p12*m2 + p13*m3 + p14*m4


#calculating std errors
first_derirative = c(estimated_infection,0,p12*m2,p13*m3,p14*m4,0,0,0)
se_hat = sqrt(t(first_derirative) %*% vcov(model) %*% first_derirative) / sqrt(8)

print(estimated_infection)
```

```
## (Intercept)
##    59114.71
```

```r
print(se_hat)
```

```
##          [,1]
## [1,] 179.5665
```

Standard errors of our estimates will be underestimated if we don't take into account correlation within age groups or heterogeneity. We can fit a Binomial GLMM with logit link to take into account the correlation within age groups. Or use quasi likelihood methods to deal with overdispersion due to heterogeneity.