Survival Analysis - Winter 2024
Assignment 1

The assignment is due January 26 before the class, via Quercus. Please return your answers in a single pdf file.

1. Let $x_1, ..., x_n$ be an IID sample from $N(\mu, \sigma^2)$ distribution, with known σ^2.

(a) Find the log-likelihood function $l(\mu)$ and the maximum likelihood estimator $\hat{\mu}$.

- Given $x_i \sim N(\mu, \sigma^2)$, the likelihood function of the data can be written as:

$$L_n(\mu; x_1, ...x_n, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\{\frac{-(x_i-\mu)^2}{2\sigma^2}\}$$
$$= (\frac{1}{\sqrt{2\pi\sigma}})^n \exp\{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}\}$$

- the log-likelihood function $\ell_n(\mu)$ then can expressed as:

$$\ell_n(\mu) = n\log(1/\sqrt{2\pi\sigma}) + \sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}$$

- the MLE of $\hat{\mu}$ then can be solved by setting the first derivative of $\ell_n(\mu)$ w.r.t μ to zero:

$$\frac{\partial \ell_n(\mu)}{\partial \mu} = 1/\sigma^2 \sum_{i=1}^{n} (x_i - \mu)$$
$$= 1/\sigma^2 (\sum_{i=1}^{n} x_i - n\mu)$$

Let $\frac{\partial \ell_n(\mu)}{\partial \mu} = 0$ gives that:

$$1/\sigma^2 (\sum_{i=1}^{n} x_i - n\hat{\mu}) = 0$$

So,MLE of $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- one last step, we check the second derivative to verify it is a maximum.

$$\frac{\partial \ell_n^2(\mu)}{\partial \mu^2} = -n/\sigma^2 < 0$$

so the log-likelihood function is a concave function, and MLE of μ is a maximum.

(b) Find the variance $V[\hat{\mu}]$.

- from (a), we know $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- the variance of $\hat{\mu}$ can be expressed as:

$$\begin{aligned} V[\hat{\mu}] &= Var(\frac{1}{n}\sum_{i=1}^{n} x_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) \\ &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

(c) Find the score variance $V[l'(\hat{\mu})]$.

- from (a) we know that $\ell_n'(\hat{\mu}) = 1/\sigma^2(\sum_{i=1}^{n} x_i - n\mu)$
- So

$$\begin{aligned} V[\ell_n'(\hat{\mu})] &= Var(1/\sigma^2(\sum_{i=1}^{n} x_i - n\mu)) \\ &= \frac{1}{\sigma^4} Var(\sum_{i=1}^{n} x_i)[\text{Var(n}\mu\text{)=0}] \\ &= \frac{n\sigma^2}{\sigma^4} = \frac{n}{\sigma^2} \end{aligned}$$

(d) Find the expected information $E[-l''(\hat{\mu})]$.

From (a), we have:

$$\ell''(\hat{\mu}) = \frac{\partial \ell_n^2(\mu)}{\partial \mu^2} = -n/\sigma^2$$

So

$$E(-\ell''(\hat{\mu})) = E(-(-n/\sigma^2)) = n/\sigma^2$$

(e) What is the connection between the quantities in (b)-(d)?

$$V[\hat{\mu}] = \frac{1}{E(-\ell''(\hat{\mu}))}$$

$$V[\ell_n'(\hat{\mu})] = \frac{1}{V[\hat{\mu}]}$$

$$V[\ell_n'(\hat{\mu})] = E(-\ell''(\hat{\mu}))$$

2.

(a) Find the log-likelihood function for the parameters in the model fitted in 1st lecture slides, slide 33, and calculate its value at the maximum likelihood point.

- we know that $D_{zx} \sim pois(\lambda_{zx} Y_{zx})$, the PDF of Poisson distribution is:

$$f(D_{zx}; \lambda_{zx}) = \frac{(\lambda_{zx} Y_{zx})^{D_{zx}} \exp\{-\lambda_{zx} Y_{zx}\}}{D_{zx}!}$$

- so the likelihood of the data can be written as:

$$\begin{aligned} L_n(\lambda) &= \prod_{z=1}^{2} \prod_{x=1}^{3} f(D_{zx}; \lambda_{zx}) \\ &= \prod_{z=1}^{2} \prod_{x=1}^{3} \frac{(\lambda_{zx} Y_{zx})^{D_{zx}} \exp\{-\lambda_{zx} Y_{zx}\}}{D_{zx}!} \end{aligned}$$

- So take the log, the log-likelihood function of the data is:

$$\ell_n(\lambda) = \sum_{z=1}^{2} \sum_{x=1}^{3} \{D_{zx} \log(\lambda_{zx} Y_{zx}) - \lambda_{zx} Y_{zx} - log(D_{zx}!)\}$$

  where $\log(\lambda_{zx} Y_{zx}) = \alpha + \beta Z + \gamma_1 \mathbf{1}\{X = 1\} + \gamma_2 \mathbf{1}\{X = 2\} + \log(Y_{zx})$

- From lecture 1, slide 33, we know the MLE of $\hat{\alpha} = -5.4177$, $\hat{\beta} = 0.8697$, $\hat{\gamma}_1 = 0.1290$, $\hat{\gamma}_2 = 0.6920$,we sub in the MLE estimates into the log-likelihood function, and $\ell_n(\lambda) = -11.898$. Below are the codes for the calculation.

```
alpha <- -5.4177
beta <- 0.8697
gamma1 <- 0.1290
gamma2 <- 0.6920

D <- c(4, 5, 8, 2, 12, 14)
y <- c(607.9,1272.1,888.9,311.9,878.1,667.5)

## log(lambdaY)
log_mu <- c(
  alpha+log(y[1]),
  alpha + gamma1+log(y[2]),
  alpha + gamma2+log(y[3]),
  alpha + beta+log(y[4]),
  alpha + beta + gamma1+log(y[5]),
  alpha + beta + gamma2+log(y[6])
)
## negative lambdaY
```

```
neg_mu <- -exp(log_mu)

## negative log(D!)
neg_logD <- -log(factorial(D))

library(dplyr)
dat <- data.frame(
  D = D,
  log_mu = log_mu,
  neg_mu = neg_mu,
  neg_logD = neg_logD
) %>%
  mutate(loglike = D * log_mu + neg_mu + neg_logD)

sum(dat$loglike)
```

(b) The model in (a) assumed that the exposure effect is constant across the age groups (the proportionality assumption). Write the log-likelihood function for the parameters in a model that relaxes this assumption. Fit this model using glm and calculate the value of the likelihood function at the maximum likelihood point.

- The log-likelihood function has the same form as in (a):

$$\ell_n(\lambda) = \sum_{z=1}^{2}\sum_{x=1}^{3}\{D_{zx}\log(\lambda_{zx}Y_{zx}) - \lambda_{zx}Y_{zx} - \log(D_{zx}!)\}$$

  However, the $\lambda_{zx}Y_{zx}$ term in the log-likelihood function changed. By including the interaction term, the model can be specified as:

$$\begin{aligned}\log(\lambda_{zx}Y_{zx}) &= \alpha + \beta Z + \gamma_1\mathbf{1}\{X = 1\} + \gamma_2\mathbf{1}\{x = 2\} \\ &\quad + \rho_1\mathbf{1}\{X = 1\}Z + \rho_2\mathbf{1}\{X = 2\}Z + \log(Y_{zx}) \qquad (2)\end{aligned}$$

- The model fit and main output is displayed in the following:

```
 interaction_m <- glm(formula = d ~ z*as.factor(x) +
                      offset(log(y)),
                      family = poisson(link = "log"))
summary(interaction_m)
 Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.02372    0.50000 -10.047   <2e-16 ***
z               -0.02582    0.86603  -0.030    0.976
as.factor(x)1   -0.51527    0.67082  -0.768    0.442
as.factor(x)2    0.31317    0.61237   0.511    0.609
```

```
z:as.factor(x)1  1.27195    1.01653   1.251    0.211
z:as.factor(x)2  0.87188    0.97285   0.896    0.370
```

1. From the model fit, we can get the MLE estimate of the parameters, and we sub in these estimates into (2) to get $\log(\lambda_{zx}Y_{zx})$, and sub $\log(\lambda_{zx}Y_{zx})$ into the log-likelihood function to calculate the likelihood value, which gives 11.06186.

(c) Use the likelihood ratio test to test the assumption mentioned in (b).
The LR statistics can be written as:

$$2(\ell_n(saturated) - \ell_n(main)) \sim \chi^2(2)$$

in this question, the LR test statistics equals $2(\ell_n(saturated) - \ell_n(main)) = 2(-11.06186 - (-11.89823)) = 1.672747$. The LR test statistics is less than the critical value of 22.36 (Using a significance level of .05), so the corresponding P-value is greater than .05, and we conclude that the interaction and main effect model do not have statistical difference, and the proportionality assumption is met.

(d) How is the residual deviance reported in the glm output related to the quantities calculated in (a) and (b)?
The residual deviance is calculated as $2(\ell_n(saturated) - \ell_n(main))$, where $\ell_n(saturated)$ is the log-likelihood value of the saturated model calculated in (b) and $\ell_n(main)$ is the log-likelihood value of the nested (main effect) model calculated in (a).

3. Hoem (1987, International Statistical Review 55) considered the association between marital status and mortality of young German men using the below dataset.

(a) Specify a model for the mortality rate, including an intercept term, age group effects, and marital status effect. Assume that the marital status effect is propor-tional over age.
Since the outcome is mortality rate, I would fit a Poisson model for it.And also notice that we assume the marital status effect is proportional over age, which suggests that we do not consider the interactive effect between marital status and age. The model can be specified as:

$$\lambda = \exp(\alpha + \beta X + \gamma_1 \mathbf{1}\{Z = 23\} + \gamma_2 \mathbf{1}\{Z = 24\} + \gamma_3 \mathbf{1}\{Z = 25\} \\ + \gamma_4 \mathbf{1}\{Z = 26\} + \gamma_5 \mathbf{1}\{Z = 27\} + \gamma_6 \mathbf{1}\{Z = 28\} + \gamma_7 \mathbf{1}\{Z = 29\})$$

where Z represents age, and X is the marital status, α is the intercept term, β and γ_s are the coefficients for marital status and age, and

$$X = \begin{cases} 1 & \text{if marital status = single} \\ 0 & \text{if marital status = married} \end{cases}$$

(b) Show how the regression coefficient parameter for marital status can be interpreted in terms of a rate ratio.

After adjusting for age, say age = 22, the mortality rate for single is $\lambda_{X=1} = \exp(\alpha + \beta)$, and $\lambda_{X=0} = \exp(\alpha)$, so the rate ratio is:

$$\begin{aligned} RR_X &= \frac{\lambda_{X=1}}{\lambda_{X=0}} \\ &= \frac{\exp(\alpha+\beta)}{\exp(\alpha)} \\ &= \exp(\alpha+\beta-\alpha) = \exp(\beta) \end{aligned}$$

So, the adjusted RR for marital status can be derived from the regression model output by taking the exponential of the marital status coefficient, which gives $\exp(\beta)$.

(c) Enter the data, fit the model and interpret the results.

the model can be fitted using the `glm` function. the output is as follows:

```
Q3_m_main <- glm(formula = death ~ (marital_status)+as.factor(age)
+ offset(log(py)),
family = poisson(link = "log"),
 data = Q3data_long)
Call:  glm(formula = death ~ (marital_status) + as.factor(age) +
offset(log(py)), family = poisson(link = "log"), data = Q3data_long)

Coefficients:
          (Intercept)  marital_statussingle      as.factor(age)23
             -5.95948               0.61114               0.00493
     as.factor(age)24      as.factor(age)25      as.factor(age)26
              0.04224               0.09898               0.14755
     as.factor(age)27      as.factor(age)28      as.factor(age)29
              0.18265               0.18736               0.20167

Degrees of Freedom: 15 Total (i.e. Null);  7 Residual
Null Deviance:      232.6
Residual Deviance: 1.047  AIC: 130.2
```

- Regarding model fitting: the residual deviance is 1.047, which is less than the critical value of 14.07($\chi^2(7)$ at α level of .05). So we conclude that the fitted model is not statistically significant different from the saturated model, the model fit well.
- Regarding the significance of exposure of interest (which is marital status): the coefficient of marital status is 0.61115(SE:0.0417), so the mortality rate ratio comparing single to married is $\exp(0.61115) =$

1.84 with 95% CI of (1.63,2.08), suggesting the mortality rate for single is 1.84 times that for married after adjusting for age. and the association is of statistical significance.

(d) Calculate the expected number of events in each age/marital status category. Compare the expected numbers to observed event counts to assess the overall model fit using the chi-squared goodness of fit test.

- the expected number of cases (denoted by μ) can be calculated from the fitted model.

$$E(\mu_{XZ}) = \hat{\lambda}_{XZ}Y_{XZ} = Y_{XZ}\{\exp(\hat{\alpha} + \hat{\beta}X + \hat{\gamma}_1\mathbf{1}\{Z = 23\} + \hat{\gamma}_2\mathbf{1}\{Z = 24\} + \hat{\gamma}_3\mathbf{1}\{Z = 25\} \\ + \hat{\gamma}_4\mathbf{1}\{Z = 26\} + \hat{\gamma}_5\mathbf{1}\{Z = 27\} + \hat{\gamma}_6\mathbf{1}\{Z = 28\} + \hat{\gamma}_7\mathbf{1}\{Z = 29\})\}$$

where

- for single, age = 22: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta}$
- for single, age = 23: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}_1$
- for single, age = 24: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}_2$
- for single, age = 25: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}_3$
- for single, age = 26: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}_4$
- for single, age = 27: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}_5$
- for single, age = 28: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}_6$
- for single, age = 29: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}_7$
- for married, age = 22: $\log(\hat{\lambda}) = \hat{\alpha}$
- for married, age = 23: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\gamma}_1$
- for married, age = 24: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\gamma}_2$
- for married, age = 25: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\gamma}_3$
- for married, age = 26: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\gamma}_4$
- for married, age = 27: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\gamma}_5$
- for married, age = 28: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\gamma}_6$
- for married, age = 29: $\log(\hat{\lambda}) = \hat{\alpha} + \hat{\gamma}_7$

The calculations are done using R, expected number of events are shown in the following table where the column name is expect_n: item the chi-squared test statistics is written as:

$$\chi^2 = \sum_{XZ} \frac{(obs\# - expected\#)^2}{expected\#}$$

with df=7. the χ^2 test statistics is 1.06335.The critical value for chi-sqaured statistics with df of 7 is 14.07(α level 0f 0.05), so the test statistics of 1.06 is less than the critical value, and P-value is greater than 0.05,indicating that the difference between the observed number of cases and the expected number of cases is not statistically significant. So the overall model fit is well.

(e) Fit also a model that allows for interaction between marital status and age. What can you say about the model fit now?

The model can be specified as:

$$\begin{aligned}\lambda = \exp(\alpha + \beta X + \gamma_1 \mathbf{1}\{Z = 23\} + \gamma_2 \mathbf{1}\{Z = 24\} + \gamma_3 \mathbf{1}\{Z = 25\} \\ +\gamma_4 \mathbf{1}\{Z = 26\} + \gamma_5 \mathbf{1}\{Z = 27\} + \gamma_6 \mathbf{1}\{Z = 28\} + \gamma_7 \mathbf{1}\{Z = 29\} \\ +\rho_1 \mathbf{1}\{Z = 23\} X + \rho_2 \mathbf{1}\{Z = 24\} X + \rho_3 \mathbf{1}\{Z = 25\} X \\ +\rho_4 \mathbf{1}\{Z = 26\} X + \rho_5 \mathbf{1}\{Z = 27\} X + \rho_6 \mathbf{1}\{Z = 28\} X + \rho_7 \mathbf{1}\{Z = 29\} X)\end{aligned}$$

```
Q3_m <- glm(formula = death ~ (marital_status)*as.factor(age) +
offset(log(py)),family = poisson(link = "log"),data = Q3data_long)
Q3_m_summary <- summary(Q3_m)
```

the `lrtest` is used to compare the main effect and interactive effect model, the codes and output are given below:

```
lrtest(Q3_m_main,Q3_m)
Likelihood ratio test

Model 1: death ~ (marital_status) + as.factor(age) + offset(log(py))
Model 2: death ~ (marital_status) * as.factor(age) + offset(log(py))
  #Df  LogLik Df Chisq Pr(>Chisq)
1   9 -56.117
2  16 -55.593  7 1.047      0.994
```

The results showed that the chi-squared test statistic with value of 1.047 (df=7), the P-value is 0.994, which is greater than a significance level of 0.05, so the interactive effect is not statistically significant. Given the fact that the main and interactive effect is not statistically different from each other, and incorporating interaction terms will resulting in estimating more parameters, and we may probably go with the main effect model. A visual check of the RR(single vs married) plot also showed that it does not change a lot over different ages, so this is reassuring.

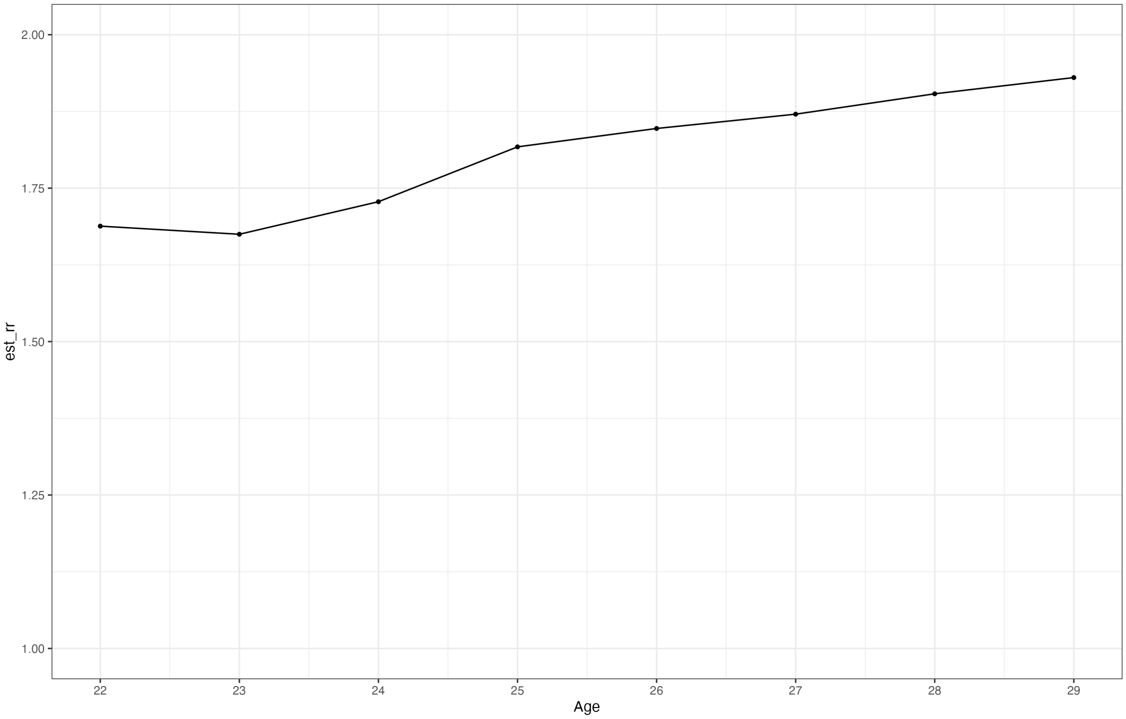

Figure 1: RR (single vs. married) plot over different ages

4. (a) The code plots the empirical daily case rates among vaccinated/unvaccinated (we take the latter to comprise both unvaccinated and partially vaccinated). Produce a plot of the empirical daily rate ratios between vaccinated/unvaccinated.

The Rate Ratio(RR) can be calculated as follows:

$$RR = \frac{\frac{\text{num. of cases in the vaccinated population}}{\text{total num. of vaccinated people}}}{\frac{\text{num. of cases in the unvaccinated population}}{\text{total num. of unvaccinated people}}}$$

The main codes are displayed below:

```
cases$rr <- cases$vacrate/cases$unvacrate
   plot_dat <- cases
   library(ggplot2)

   rr_plot <- plot_dat %>%
     ggplot(aes(x=date,y=rr))+
     geom_line()+
     scale_x_date(date_breaks = "1 month", date_labels =  "%b %Y") +
     theme_bw()+
     xlab(’Date’)+
     geom_hline(yintercept = 1, linetype = "dashed", color = "grey")+
     ylab(’Daily COVID-19 Case Rate Ratio (Vaccinated/Unvaccinated)’)
```

the output of the RR plot is:

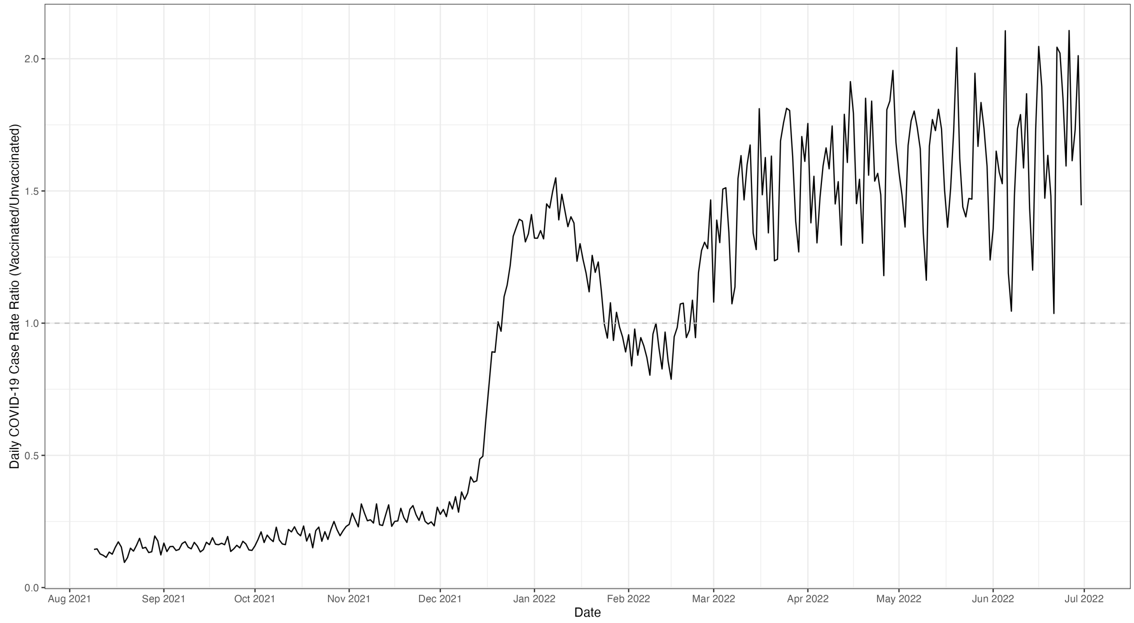

Figure 2: empirical daily rate ratios between vaccinated/unvaccinated

(b) Fit an appropriate saturated Poisson regression model to verify that you can use the model to reproduce the same daily rate ratios. Add also the 95% confidence bands into the plot. What can you say about the "vaccination effect" on preventing infection/positive test over time? Would the assumption of constant vaccination effect be reasonable in this case? How can you test this?

The saturated model can be specified as follows, that is,the estimated number of cases μ_{XZ} (or estimated event rate λ_{XZ}) can be modeled as:

$$\log(\mu_{XZ}) = log(\lambda_{XZ} Y_{XZ}) = 0 + \sum_{i}^{326} \alpha_i \mathbf{1}_{\{X=i\}} + \sum_{i}^{326} \beta_i \mathbf{1}_{\{X=i\}} Z + \log(Y_X Z)$$

where i represents the i^{th} date, X represents date variable, and Z represents vaccinate status, with Z =1 for vaccinated, and Z=0 for unvaccinated.

For the i^{th} date, the rate(λ_{i1}) for vaccinated group is:

$$\lambda_{i1} = \exp(\alpha_i + \beta_i)$$

For the i^{th} date, the rate(λ_{i0}) for unvaccinated group is:

$$\lambda_{i0} = \exp(\alpha_i)$$

it then gives the RR as:

$$RR = \frac{\lambda_{i1}}{\lambda_{i0}} = \frac{\exp(\alpha_i + \beta_i)}{\exp(\alpha_i)} = \exp(\beta_i)$$

The 95%CI is given by:

$$(\exp(\beta_i - 1.96SE_i), \exp(\beta_i + 1.96SE_i))$$

where SE is the standard error of β_i The codes for fitting the saturated model are presented below:

```
sat_m1 <- glm(formula = d ~ 0 + as.factor(date) +
fullyvac:as.factor(date)+ offset(log(y)),
family = poisson(link = "log"),
              data = cases)
```

the RR plot (codes provided in the Appendix) is presented below, the blue line is the point estimate of daily RR, and grey lines are the corresponding 95% CI :

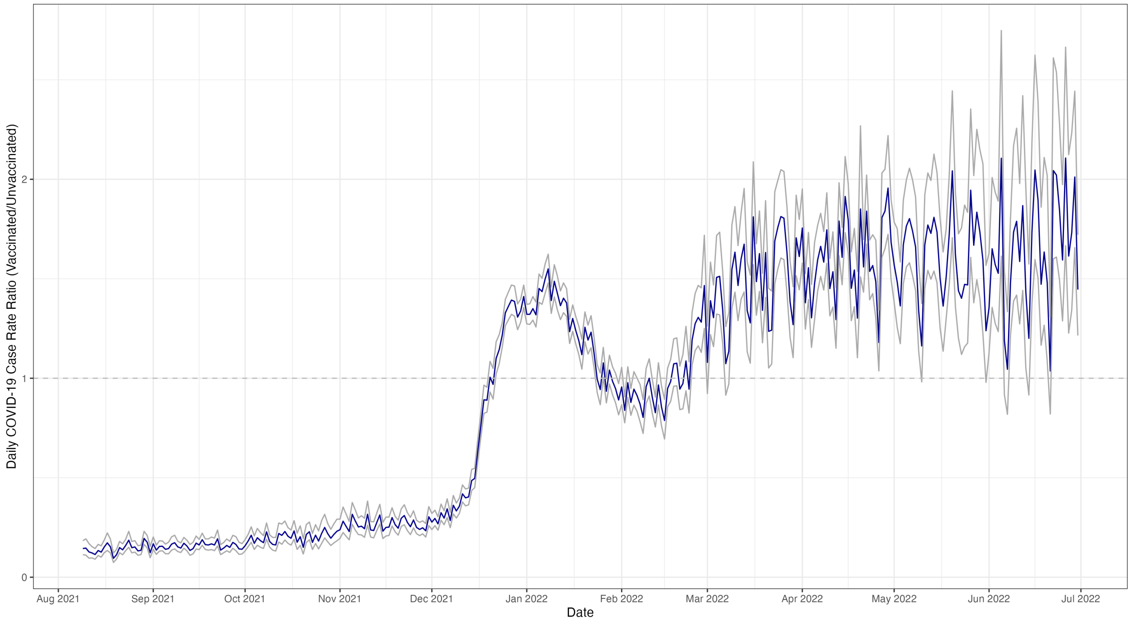

Figure 3: The Daily RR plot

- From the plot, we see that the vaccination effect's direction changed over time. From 2021/08 to 2021/12/21, the RR is below 1, which indicates that vaccination has protective effect on protecting people from getting infected. and after that, the direction of the effect reversed, we see most of the dates (except for 2022/01/24 - 2022/02/23) have RR greater than 1, indicating that the vaccinated group has higher risk of getting infected.
- as we can see from the plot, the RR changed over time, so the assumption of constant vaccination effect is not reasonable. this assumption can be tested using likelihood ratio test where the likelihoods of the

saturated model and the main effect model(nested model) are compared. Specifically, $2(\ell(saturated) - \ell(main)) \sim \chi^2$, using `lrtest` command in R, we get $\chi^2 = 55725$ with df =325, and the corresponding P-value is less than 2.2e-16. So we can conclude that the interactive effect between date and vaccination status is statistically significant, and the constant vaccination effect assumption is again, not reasonable.

(c) The daily rate ratios are quite noisy due to small counts. Using JAGS, fit an appropriate Bayesian Poisson regression model to smooth the daily rates, and plot the resulting posterior mean rate ratios along with the 95% credible intervals. Compare the results to the unsmoothed estimates.

The JAGs setup is as follows:

- 1. the data:

```
datalist <- list(
  "fullyvac" = cases$fullyvac,
  "date_n" = cases1$date_n
    %>% as.factor(),
  "d" = cases$d,
  "date" = cases$date,
  "y" = cases$y,
  "N" = nrow(cases))
```

- 2. the parameters: we have in total 326 params for the date variable and 326 params for the interaction term, and the initial values of these parameters are set to 0

```
ndate = 326
initslist <- list( 'beta'=rep(0, ndate),
    'gamma'=rep(0,ndate), 'phi'=1)
```

- 3. the model : as we specified in (b), the model set up is as below:

```
model {
  for (i in 1:652) {
    d[i] ~ dpois(mu[i])
    mu[i] <- y[i] * exp(gamma[date_n[i]] + beta[date_n[i]]*fullyvac[i])
  }
  betamean[1] <- 0.0
  betaprec[1] <- phi * 0.001
  betamean[2] <- 0.0
  betaprec[2] <- phi * 0.001
  for (i in 3:326) {
    betamean[i] <- 2 * beta[i-1] - beta[i-2]
    betaprec[i] <- phi
  }
  for (i in 1:326) {
    beta[i] ~ dnorm(betamean[i],betaprec[i])
    logRR[i] <- beta[i]
  }
  phi ~ dgamma(0.001,0.001)
  for (i in 1:326) {
    gamma[i] ~ dnorm(0.0,0.001)
  }
}
```

the smoothed RR plot (codes in the Appendix) is presented below, where the RR of the i^{th} date, RR_i is calculated as $\exp(\beta_i)$, the 95% credible region is calculated from the 2.5%-tile and 97.5%-tile

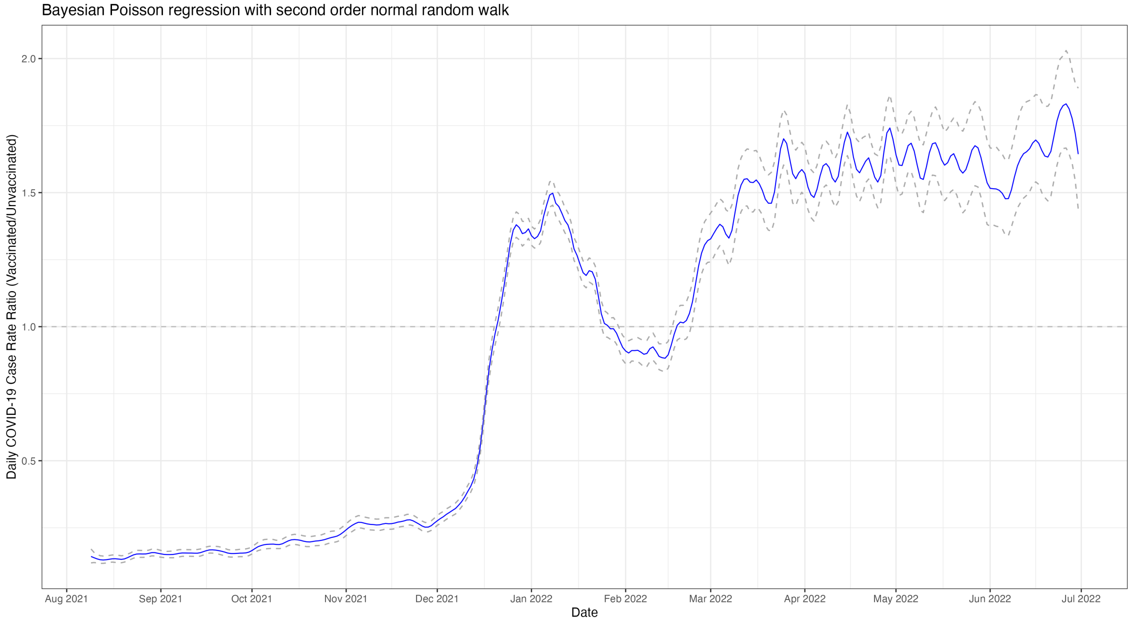

Figure 4: Bayesian Poisson regression

In the Bayesian Poisson regeression, the second order normal random walk is used, so the RR plot is more smoothed compared to the unsmoothed estimates. and In general, the unsmoothed and smoohted RRs showed similar pattern: From 2021/08 to 2021/12/21, the RR is below

1, which indicates that vaccination has protective effect on protecting people from getting infected. and after that, the direction of the effect reversed, we see most of the dates (except for 2022/01/24 - 2022/02/23) have RR greater than 1, indicating that the vaccinated group has higher risk of getting infected

5. (a) Fit an appropriate Poisson regression model to estimate the case rate ratios by vaccination status, adjusting for age group. Plot the resulting rate ratios and their 95% confidence intervals and compare them to the crude rate ratios. Here we assume the age effect to be proportional over time. Based on the model we fitted in Q4, this model will adjust for the age effect by including agegroup variable as a main effect in the model. the model can be specified as:

$$\log(\mu_{XZ}) = log(\lambda_{XZ}Y_{XZ}) = 0+\sum_{i}^{326}\alpha_i\mathbf{1}_{\{X=i\}}+\sum_{i}^{326}\beta_i\mathbf{1}_{\{X=i\}}Z+\sum_{j}^{5}\gamma_j\mathbf{1}_{\{agegroup=j\}}+\log(Y_{XZ})$$

the relevant R codes are presented below:

```
source("R/cases_hosp_age.r")
Q5_dat <- cases
Q5_m <- glm(formula = d ~ 0 + as.factor(date) +
fullyvac:as.factor(date)+ as.factor(agegroup)+offset(log(y)),
              family = poisson(link = "log"),
              data = Q5_dat)
```

the output of the RR plots are shown here.

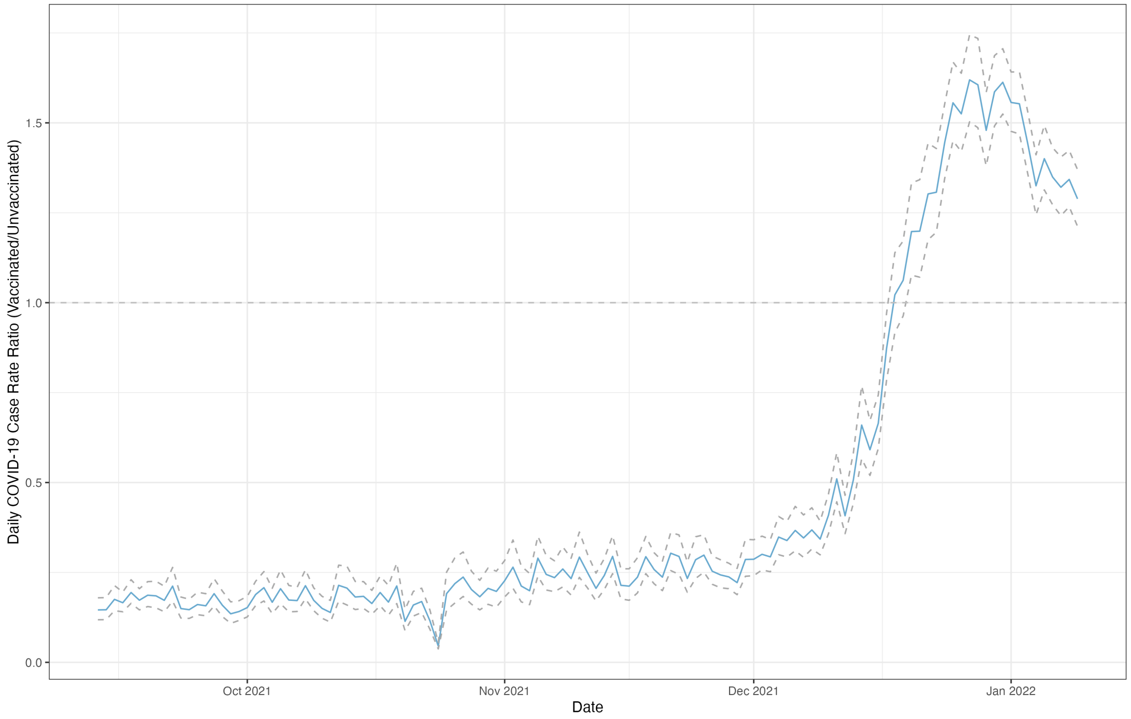

Figure 5: Adjusted RR plot (adjusted for agegroup)

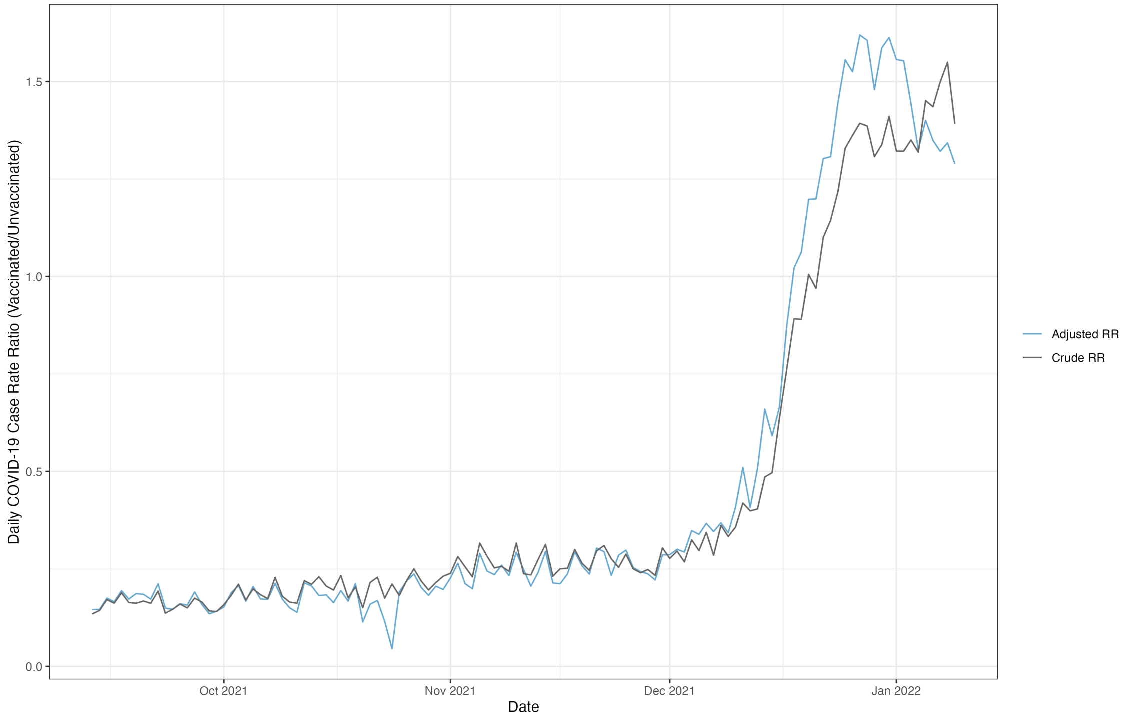

Figure 6: Crude vs Adjusted RR

We can see from the plot that:

- before 2021 mid Dec, the difference between the crude RR and adjusted RR is small.But after that we see that the difference between adjusted crude RR became greater, with adjusted RR greater than the crude RR. This indicates that before 2021 mid Dec, the age distributions in the vaccinated and unvaccinated groups are similar, but after 2021 mid Dec, the age distributions in two comparison groups are different. The RR values are then distorted.
- Also, using mid Dec 2021 as cutoff date (approximately), both adjusted and crude RR showed protective effect (RR less than 1) for the infection, however, after that, the crude and adjusted RR became greater than 1, indicating a reversed effect of the vaccination effect, that is, the vaccinated group has a higher risk of getting infected.

(b) Again, the daily rate ratios are quite noisy. Using JAGS, fit an appropriate Bayesian Poisson regression model to smooth the daily rates, and plot the resulting posterior mean rate ratios along with the 95% credible intervals. Interpret the results.

The JAGs setup is as follows:

- 1. the data:

```
datalist <- list('fullyvac'=Q5_dat1$fullyvac,
                              'date_n'=Q5_dat1$date_n,
                              'd'=Q5_dat1$d,
                              'y'=Q5_dat1$y,
                              'agegrp_n' =Q5_dat1$agegrp_n,
                              'N'= nrow(Q5_dat1))
```

- 2. the parameters: we have in total 119 params for the date variable and 119 params for the interaction term, and 5 param for the agegroup variable, and the initial values of these parameters are set to 0

```
ndate <- 119
  initslist <- list( 'alpha'=rep(0,n_agegrp),
                     'beta'=rep(0, ndate),
                     'gamma'=rep(0,ndate), 'phi'=1)
```

- 3. the model : the model set up is as below:

```
model {
  for (i in 1:N) {
    d[i] ~ dpois(mu[i])
    mu[i] <- y[i] * exp(alpha[agegrp_n[i]] +
    gamma[date_n[i]] + beta[date_n[i]]*fullyvac[i])
  }
  betamean[1] <- 0.0
  betaprec[1] <- phi * 0.001
  betamean[2] <- 0.0
  betaprec[2] <- phi * 0.001
  for (i in 3:119) {
    betamean[i] <- 2 * beta[i-1] - beta[i-2]
    betaprec[i] <- phi
  }
  for (i in 1:119) {
    beta[i] ~ dnorm(betamean[i],betaprec[i])
    logRR[i] <- beta[i]
  }
  phi ~ dgamma(0.001,0.001)
  for (i in 1:119) {
    gamma[i] ~ dnorm(0.0,0.001)
  }
for (i in 1:5) {
    alpha[i] ~ dnorm(0.0,0.001)
  }
}
```

the smoothed RR plot (codes in the Appendix) is presented below, where the RR of the i^{th} date, RR_i is calculated as $\exp(\beta_i)$, the 95% credible region is calculated from the 2.5%-tile and 97.5%-tile

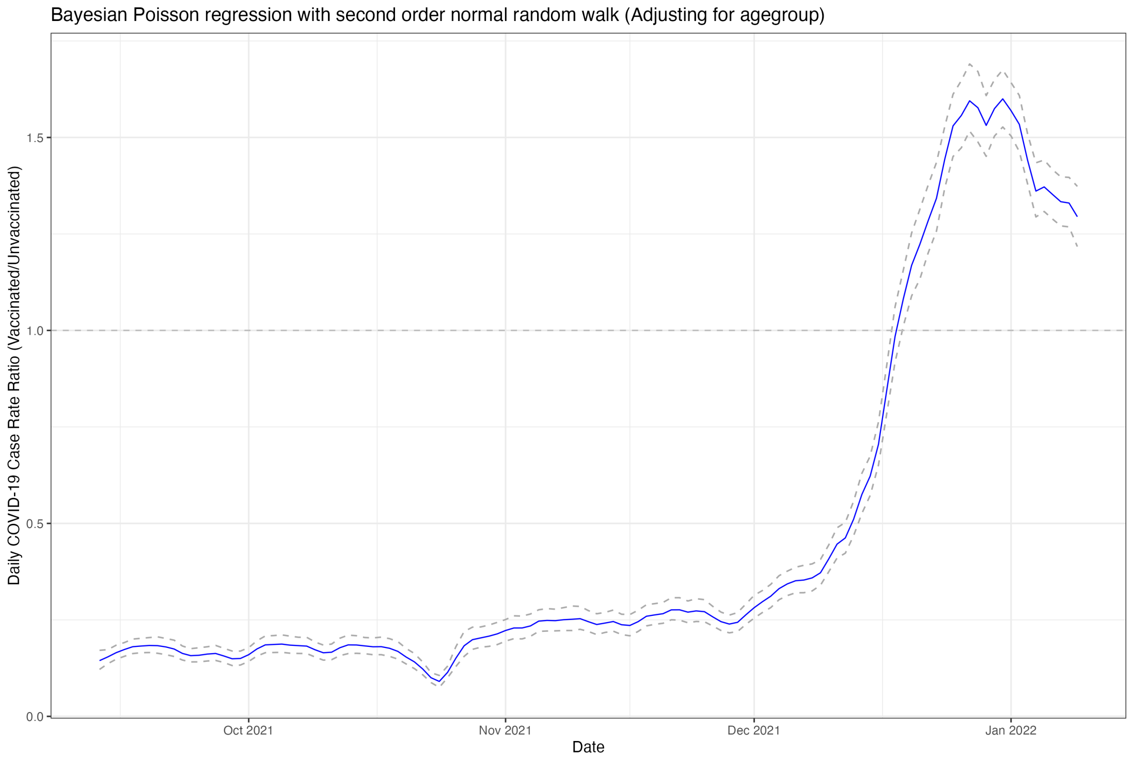

Figure 7: Bayesian Poisson regression

Again, compared to the unsmoothed estimates, the Bayesian Poisson regression is much more smoother. After adjusting for age, before mid Dec 2021, the RRs are below 1, and after that effect reversed, and the RRs are above 1.

Codes for this assignment

Q2.

(a)
```
alpha <- -5.4177
beta <- 0.8697
gamma1 <- 0.1290
gamma2 <- 0.6920

D <- c(4, 5, 8, 2, 12, 14)
y <- c(607.9,1272.1,888.9,311.9,878.1,667.5)

## log(lambdaY)
log_mu <- c(
  alpha+log(y[1]),
  alpha + gamma1+log(y[2]),
  alpha + gamma2+log(y[3]),
  alpha + beta+log(y[4]),
  alpha + beta + gamma1+log(y[5]),
  alpha + beta + gamma2+log(y[6])
)
## negative lambdaY
neg_mu <- -exp(log_mu)

## negative log(D!)
neg_logD <- -log(factorial(D))

library(dplyr)
dat <- data.frame(
  D = D,
  log_mu = log_mu,
  neg_mu = neg_mu,
  neg_logD = neg_logD
) %>%
  mutate(loglike = D * log_mu + neg_mu + neg_logD)

sum(dat$loglike)


d <- c(4,5,8,2,12,14)
y <- c(607.9,1272.1,888.9,311.9,878.1,667.5)
z <- c(0,0,0,1,1,1)
x <- c(0,1,2,0,1,2)
main_m <- glm(formula = d ˜ z+as.factor(x) +
offset(log(y)),family = poisson(link = "log"))
```

(b)

```
d <- c(4,5,8,2,12,14)
y <- c(607.9,1272.1,888.9,311.9,878.1,667.5)
z <- c(0,0,0,1,1,1)
x <- c(0,1,2,0,1,2)
main_m <- glm(formula = d ˜ z+as.factor(x) +
offset(log(y)), family = poisson(link = "log"))


interaction_m <- glm(formula = d ˜ z*as.factor(x) +
offset(log(y)), family = poisson(link = "log"))


summary(interaction_m)
```

```
alpha1 <- -5.02372
beta1 <- -0.02582
gamma1_1 <- -0.51527
gamma2_1 <- 0.31317
rho1 <- 1.27195
rho2 <- 0.87188

log_mu1 <- c(alpha1+log(y[1]),
             alpha1+gamma1_1+log(y[2]),
             alpha1+gamma2_1+log(y[3]),
             alpha1+beta1+log(y[4]),
             alpha1+beta1+gamma1_1+rho1+log(y[5]),
             alpha1+beta1+gamma2_1+rho2+log(y[6])
             )

dat1 <- data.frame(D = d,
           log_mu1 = log_mu1,
           mu1 = exp(log_mu1),
           logD = log(factorial(d)),
           dlog_lambda = d*log_mu1) %>%
  mutate(loglike1 = D*log_mu1-mu1-logD)




dat_all <- cbind(dat,dat1)


2*(sum(dat_all$loglike1) - sum(dat_all$loglike) )

main_m_residual_Deviance = 2*(sum(dat_all$loglike1) - sum(dat_all$loglike) )
## double check
##anova(main_m, interaction_m, test = "LRT")
```

Q3.

(c)

```
Q3data <- read_excel("data/CHL5209-HW1-Q3data.xlsx")%>%
  select(-total_death, -total_py)

Q3data_long <- Q3data %>% select(age,single,married) %>%
  tidyr::pivot_longer(cols = c(single,married),
                      names_to = 'marital_status',
                      values_to = 'death') %>%
  bind_cols(
    Q3data %>% select(single_py,married_py) %>%
      tidyr::pivot_longer(cols = c(single_py,married_py),
                          names_to = ’py_tmp’,
                          values_to = ’py’) %>%
      select(-py_tmp)
  )

Q3_m_main <- glm(formula = death ~
(marital_status)+as.factor(age) + offset(log(py)),
            family = poisson(link = "log"),
            data = Q3data_long)

Q3_m_main_summary <- summary(Q3_m_main)$coefficients %>% data.frame()


alpha_hat <- Q3_m_main_summary$Estimate[1]
beta_hat <- Q3_m_main_summary$Estimate[2]

log_lambda_dat <- Q3_m_main_summary %>%
  mutate( single_log_lambda = ifelse(rownames(.)==’marital_statussingle’,
                 alpha_hat+Estimate,
                 alpha_hat+beta_hat+Estimate),
          married_log_lambda = ifelse(rownames(.)==’marital_statussingle’,
                                     alpha_hat,
                                     alpha_hat+Estimate)
  ) %>%
  filter(rownames(.)!="(Intercept)")
```

(d)

```
expected_dat <- data.frame(
  age = rep(22:29, 2),
  marital_status = c(
    rep("single", length(22:29)),
    rep("married", length(22:29))
  ),
  log_lambda_hat = c(log_lambda_dat$single_log_lambda,
  log_lambda_dat$married_log_lambda)
) %>% full_join(Q3data_long,by=c(’age’,’marital_status’))
%>%
  mutate(expect_n = exp(log_lambda_hat)*py,
         chi_square = (death - expect_n)^2/expect_n
         )
```

```
expected_dat$chi_square %>% sum()
```

(e)

```
    ## saturated model
Q3_m <- glm(formula = death ~ (marital_status)*as.factor(age)
+ offset(log(py)),
             family = poisson(link = "log"),
             data = Q3data_long)



Q3_m_summary <- summary(Q3_m)$coefficients %>% data.frame() %>%
  mutate(est_rr = ifelse(rownames(.)==’marital_statussingle’,
        exp(Q3_m$coefficients[’marital_statussingle’]),
        exp(Q3_m$coefficients[’marital_statussingle’]+ Estimate)),
         var_name = rownames(.))
library(stringr)

Q3_m_coef <- Q3_m_summary %>%
  filter(grepl(":",var_name)|var_name =="marital_statussingle")%>%
  mutate(age = str_replace_all(rownames(.),
  "marital_statussingle:as\\.factor\\(age\\)", ’’)) %>%
  mutate(age=ifelse(var_name==’marital_statussingle’,
  min(Q3data_long$age),age))

Q3_rr_plot <- Q3_m_coef %>%
  ggplot(aes(x=as.numeric(age),y=est_rr))+
  geom_line()+
  geom_point(size=1)+
  scale_x_continuous(breaks =seq(min(Q3_m_coef$age),max(Q3_m_coef$age) ,
  by = 1)) +
  theme_bw()+
  xlab(’Age’)+
  ylab(’Rate Rito (Single vs Married)’)

ggsave(Q3_rr_plot,height = 7.36,width = 11.5,
       file=’output/Q3_rr_plot.png’,dpi=300)
```

Q4.

(a) and (b)

```
rr_table <- cases[,c(’date’,’unvac’,’vac’,
’unvacpop’,’vacpop’,’rr’)]
cases <- cases[,c(’date’,’unvac’,’vac’,
’unvacpop’,’vacpop’)]
nrow(cases)
cases <- reshape(cases, varying=list(c(’unvac’,’vac’),
c(’unvacpop’, ’vacpop’)),
                 direction=’long’, times=c(0,1))
cases <- cases[order(cases$date),]
nrow(cases)
names(cases) <- c(’date’, ’fullyvac’, ’d’, ’y’, ’id’)


sat_m1 <- glm(formula = d ˜ 0 + as.factor(date)
+fullyvac:as.factor(date)+ offset(log(y)),
             family = poisson(link = "log"),
             data = cases)

sat_m1_main <- glm(formula = d ˜ as.factor(date)
+as.factor(fullyvac)+ offset(log(y)),
              family = poisson(link = "log"),
              data = cases)
install.packages("lmtest")
library(lmtest)
lrtest(sat_m1,sat_m1_main)
lrtest(sat_m1_main,sat_m1)

sat_m1_summary <- summary(sat_m1)$coefficients %>% data.frame() %>%
  filter(grepl(":",rownames(.))) %>%
  mutate(est_rr = exp(Estimate),
         date = regmatches(rownames(.),
 regexpr("\\d{4}-\\d{2}-\\d{2}",
  as.character(rownames(.)))) %>%
           as.Date(),
         est_rr_lower = exp(Estimate - 1.96*Std..Error),
         est_rr_upper = exp(Estimate + 1.96*Std..Error)

         )

Q4_1_plot_dat <- left_join(rr_table,sat_m1_summary,by=’date’)


rr_plot_saturated <- Q4_1_plot_dat %>%
  ggplot(aes(x=date,y=rr))+
  geom_line(color=’darkblue’)+
  geom_line(aes(y=est_rr_lower),color=’darkgrey’)+
  geom_line(aes(y=est_rr_upper),color=’darkgrey’)+
  scale_x_date(date_breaks = "1 month", date_labels =  "%b %Y") +
  geom_hline(yintercept = 1, linetype = "dashed", color = "grey")+
  theme_bw()+
  xlab('Date')+
  ylab('Daily COVID-19 Case Rate Ratio (Vaccinated/Unvaccinated)')

rr_plot_saturated
ggsave(rr_plot_saturated,height =  7.47 ,width =13.5,
       file='output/Q4_rr_ci_plot.png',dpi=300)

rr_plot+
  geom_line(data=Q4_1_plot_dat,aes(y=est_rr,color='red'))
```

(c)

```
cases1 <- cases %>% mutate(date_n =  as.numeric(date - min(date)+1),
                           interc = as.numeric(date)*fullyvac )

# datalist_tmp <- data.frame('fullyvac'=cases$fullyvac,
#                            'date_n'=cases1$date_n,
#                            'd'=cases$d,
#                            'y'=cases$y,
#                  'N'= nrow(cases))

datalist <- list(
  "fullyvac" = cases$fullyvac,
  "date_n" = cases1$date_n
    %>% as.factor(),
  "d" = cases$d,
  "date" = cases$date,
  "y" = cases$y,
  "N" = nrow(cases)
)
dput(datalist, file.path("data/vacdata.txt"))

ndate <- cases$date %>% unique() %>% length()

initslist <- list( 'beta'=rep(0, ndate),
                   'gamma'=rep(0,ndate), 'phi'=1)
dput(initslist,"data/vacinits.txt")

model <- jags.model('data/vacmodel.txt',
                    data=datalist, inits=initslist,
                    n.chains=2, quiet=FALSE)
```

```
ndate <- cases$date %>% unique() %>% length()
# Check convergence:

samples <- coda.samples(model, c(’logRR’),
n.iter=10000, n.burnin=5000, thin = 10)
#plot(samples, ask=TRUE)

# Some summary statistics:

summary(samples)
logrrs <- as.matrix(samples)
#boxplot(logrrs,use.cols=TRUE)

logrr <- colMeans(logrrs)
ciu <- apply(logrrs, 2, quantile, probs=0.975)
cil <- apply(logrrs, 2, quantile, probs=0.025)

baye_pios <- data.frame(
  bayes_rr = logrr %>% exp(),
  bayes_lower =  cil%>% exp(),
  bayes_upper = ciu %>% exp(),
  date = Q4_1_plot_dat$date %>% sort()
)


 bayes_rr_plot <-
   baye_pios %>%
   ggplot(aes(x=date,y=bayes_rr))+
   geom_line(color=’blue’,size=0.4)+
   scale_x_date(date_breaks = "1 month", date_labels =  "%b %Y") +
   theme_bw()+
   xlab(’Date’)+
   ggtitle(’Bayesian Poisson regression
   with second order normal random walk’)+
   ylab(’Daily COVID-19 Case Rate Ratio (Vaccinated/Unvaccinated)’)+
 geom_line(data=baye_pios,aes(y=bayes_lower),
 color=’darkgrey’,linetype=’dashed’)+
   geom_line(data=baye_pios,aes(y=bayes_upper),
   color=’darkgrey’,linetype=’dashed’)+
   geom_hline(yintercept = 1, linetype = "dashed", color = "grey")

 bayes_rr_plot
ggsave(bayes_rr_plot,height = 7.47,width = 13.5,
        file=’output/Q4_bayes_rr_plot.png’,dpi=300)
```

```
### JAGs
model {
  for (i in 1:N) {
    d[i] ~ dpois(mu[i])
    mu[i] <- y[i] * exp(gamma[date_n[i]] + beta[date_n[i]]*fullyvac[i])
  }
  betamean[1] <- 0.0
  betaprec[1] <- phi * 0.001
  betamean[2] <- 0.0
  betaprec[2] <- phi * 0.001
  for (i in 3:326) {
    betamean[i] <- 2 * beta[i-1] - beta[i-2]
    betaprec[i] <- phi
  }
  for (i in 1:326) {
    beta[i] ~ dnorm(betamean[i],betaprec[i])
    logRR[i] <- beta[i]
  }
  phi ~ dgamma(0.001,0.001)
  for (i in 1:326) {
    gamma[i] ~ dnorm(0.0,0.001)
  }
}
```

Q5.

(a)

```
       source("R/cases_hosp_age.r")
Q5_dat <- cases
Q5_m <- glm(formula = d ~ 0 + as.factor(date) +fullyvac:as.factor(date)+
as.factor(agegroup)+offset(log(y)),
              family = poisson(link = "log"),
              data = Q5_dat)

Q5_m_summary <- summary(Q5_m)$coefficients %>% data.frame() %>%
  filter(grepl(":",rownames(.))) %>%
  mutate(est_rr = exp(Estimate),
         date = regmatches(rownames(.),
                           regexpr("\\d{4}-\\d{2}-\\d{2}",
                                   as.character(rownames(.)))) %>%
           as.Date(),
         est_rr_lower = exp(Estimate - 1.96*Std..Error),
         est_rr_upper = exp(Estimate + 1.96*Std..Error)
  )

Q5plot_dat <- left_join(Q5_m_summary,rr_table,by=’date’) %>%
  select(date,rr,est_rr) %>%
  rlang::set_names(’date’,’Crude RR’,’Adjusted RR’) %>%
  tidyr::pivot_longer(cols = c(‘Crude RR‘,‘Adjusted RR‘),
                      names_to = ’model’,
                      values_to = ’RR’)


Q5plot <- Q5plot_dat %>%
  ggplot(aes(x=date,y=RR,color=model))+
  geom_line()+
  theme_bw()+
  scale_x_date(date_breaks = "1 month",
  date_labels =  "%b %Y") +
  theme_bw()+
  theme(legend.title = element_blank())+
  scale_color_manual(values = c("Crude RR" = "#666666",
  "Adjusted RR" = "#67a9cf"))+
  xlab(’Date’)+
  ylab(’Daily COVID-19 Case Rate Ratio (Vaccinated/Unvaccinated)’)


  ggsave(Q5plot,height = 7.36,width = 11.5,
         file=’output/Q5_rr_plot.png’,dpi=300)


  Q5plot_ci <- Q5_m_summary %>%
    ggplot(aes(x=date,y=est_rr))+
    geom_line(color=’#67a9cf’)+
    geom_line(aes(y=est_rr_lower),color=’darkgrey’,
    linetype=’dashed’)+
    geom_line(aes(y=est_rr_upper),color=’darkgrey’,
    linetype=’dashed’)+
    scale_x_date(date_breaks = "1 month",
    date_labels =  "%b %Y") +
    geom_hline(yintercept = 1, linetype = "dashed",
    color = "grey")+
    theme_bw()+
    xlab(’Date’)+
    ylab(’Daily COVID-19 Case Rate Ratio (Vaccinated/Unvaccinated)’)


  ggsave(Q5plot_ci,height = 7.36,width = 11.5,
          file=’output/Q5_Q5plot_ci.png’,dpi=300)
```

(b)

```
Q5_dat1 <- Q5_dat %>% mutate(date_n =  as.numeric(date - min(date)+1),
                                agegrp_n = case_when(agegroup==’12-17yrs’~1,
                                                     agegroup==’18-39yrs’~2,
                                                     agegroup==’40-59yrs’~3,
                                                     agegroup==’60-79yrs’~4,
                                                     agegroup==’80+’~5
                                                     ))

   datalist <- list(’fullyvac’=Q5_dat1$fullyvac,
                              ’date_n’=Q5_dat1$date_n,
                              ’d’=Q5_dat1$d,
                              ’y’=Q5_dat1$y,
                              ’agegrp_n’ =Q5_dat1$agegrp_n,
                              ’N’= nrow(Q5_dat1))

   dput(datalist, file.path("data/vacdata_Q5.txt"))

   ndate <- Q5_dat1$date %>% unique() %>% length()
   n_agegrp <- Q5_dat1$agegroup %>% unique() %>% length()

   initslist <- list( ’alpha’=rep(0,n_agegrp),
   ’beta’=rep(0, ndate), ’gamma’=rep(0,ndate), ’phi’=1)
   dput(initslist,"data/vacinits_Q5.txt")

   model_Q5 <- jags.model(’data/vacmodel_Q5.txt’,
                       data=datalist, inits=initslist, n.chains=2, quiet=FALSE)


   # Check convergence:

   samples <- coda.samples(model_Q5, c(’logRR’),
   n.iter=10000, n.burnin=5000, thin = 10)
   #plot(samples, ask=TRUE)

   # Some summary statistics:

   summary(samples)
   logrrs <- as.matrix(samples)
   #boxplot(logrrs,use.cols=TRUE)
```

```
  logrr <- colMeans(logrrs)
  ciu <- apply(logrrs, 2, quantile, probs=0.975)
  cil <- apply(logrrs, 2, quantile, probs=0.025)

  baye_pios_Q5 <- data.frame(
    bayes_rr = logrr %>% exp(),
    bayes_lower =  cil%>% exp(),
    bayes_upper = ciu %>% exp(),
    date = Q5_dat1$date %>% unique() %>%  sort()
  )



  bayes_rr_plot_Q5 <-
    baye_pios_Q5 %>%
    ggplot(aes(x=date,y=bayes_rr))+
    geom_line(color=’blue’,size=0.4)+
    scale_x_date(date_breaks = "1 month", date_labels =  "%b %Y") +
    theme_bw()+
    xlab(’Date’)+
    ggtitle(’Bayesian Poisson regression
    with second order normal random walk (Adjusting for agegroup)’)+
    ylab(’Daily COVID-19 Case Rate Ratio (Vaccinated/Unvaccinated)’)+
    geom_line(data=baye_pios_Q5,aes(y=bayes_lower),
    color=’darkgrey’,linetype=’dashed’)+
    geom_line(data=baye_pios_Q5,aes(y=bayes_upper),
    color=’darkgrey’,linetype=’dashed’)+
    geom_hline(yintercept = 1, linetype = "dashed",
    color = "grey")

  bayes_rr_plot_Q5

  ggsave(bayes_rr_plot_Q5,width = 11.1,height = 7.47,
         file=’output/Q5_bayes_rr_plot.png’,dpi=300)
##JAGs
model {
  for (i in 1:N) {
    d[i] ~ dpois(mu[i])
    mu[i] <- y[i] * exp(alpha[agegrp_n[i]] +
    gamma[date_n[i]] + beta[date_n[i]]*fullyvac[i])
  }
  betamean[1] <- 0.0
  betaprec[1] <- phi * 0.001
  betamean[2] <- 0.0
  betaprec[2] <- phi * 0.001
  for (i in 3:119) {
    betamean[i] <- 2 * beta[i-1] - beta[i-2]
    betaprec[i] <- phi
  }
  for (i in 1:119) {
    beta[i] ~ dnorm(betamean[i],betaprec[i])
    logRR[i] <- beta[i]
  }
  phi ~ dgamma(0.001,0.001)
  for (i in 1:119) {
    gamma[i] ~ dnorm(0.0,0.001)
  }
for (i in 1:5) {
    alpha[i] ~ dnorm(0.0,0.001)
  }
}
```