A3 Soln

Xiangyu Kong, 1002109620

24 March, 2020

Question 1

```
birth_file = 'birthData.rds'
if (!file.exists(birth_file)) {
  download.file('http://pbrown.ca/teaching/303/data/birthData.rds',
                birth_file)
}
x = readRDS(birth_file)
```

```
x$bygroup = factor(gsub(
  "[[:space:]]",
  "",
  paste0(x$MetroNonmetro, x$MothersHispanicOrigin)
))
x$timeInt = as.numeric(x$time)
x$y = as.matrix(x[, c('Male', 'Female')])
x$sin12 = sin(x$timeInt / 365.25)
x$cos12 = cos(x$timeInt / 365.25)
x$sin6 = sin(2 * x$timeInt / 365.25)
x$cos6 = cos(2 * x$timeInt / 365.25)
baselineDate = as.Date('2007/1/1')
baselineDateInt = as.integer(baselineDate)
```

```
res = mgcv::gam(
  y ~ bygroup +
    cos12 + sin12 + cos6 + sin6 +
    s(timeInt, by = bygroup, k = 120, pc = baselineDateInt),
  data = x,
  family = binomial(link = 'logit')
)
```

```
res2 = gamm4::gamm4(
  y ~ bygroup +
    cos12 + sin12 + cos6 + sin6 +
    s(timeInt, by = bygroup, k = 120, pc = baselineDateInt),
  random = ~ (1 | bygroup:timeInt),
  data = x,
  family = binomial(link = 'logit')
)
```

```
coefGamm = summary(res2$mer)$coef

knitr::kable(cbind(mgcv::summary.gam(res)$p.table[, 1:2],
                   coefGamm[grep("^Xs[(]", rownames(coefGamm), invert = TRUE), 1:2]),
             digits =5)
```

	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	0.03237	0.00583	0.04223	0.00128
bygroupMetroNotHispanicorLatino	0.01942	0.00640	0.00678	0.00149
bygroupNonmetroHispanicorLatino	-0.02340	0.02013	-0.00643	0.00455
bygroupNonmetroNotHispanicorLatino	0.01550	0.00604	0.00593	0.00209
cos12	0.00060	0.00125	-0.00026	0.00048
sin12	-0.00021	0.00123	0.00046	0.00047
cos6	0.00165	0.00116	0.00092	0.00045
sin6	0.00071	0.00118	0.00010	0.00046

```
1/sqrt(res$sp)
```

```
##       s(timeInt):bygroupMetroHispanicorLatino
##                                    5.201104e-01
##    s(timeInt):bygroupMetroNotHispanicorLatino
##                                    2.224808e-01
##    s(timeInt):bygroupNonmetroHispanicorLatino
##                                    1.284191e+00
## s(timeInt):bygroupNonmetroNotHispanicorLatino
##                                    2.847145e-05
```

```
lme4::VarCorr(res2$mer)
```

```
##  Groups          Name                                          Std.Dev.
##  bygroup:timeInt (Intercept)                                   0.0022596
##  Xr.2            s(timeInt):bygroupNonmetroNotHispanicorLatino 0.0000000
##  Xr.1            s(timeInt):bygroupNonmetroHispanicorLatino    0.0000000
##  Xr.0            s(timeInt):bygroupMetroNotHispanicorLatino    0.0000000
##  Xr              s(timeInt):bygroupMetroHispanicorLatino       0.0000000
```

```
timeJan = as.numeric(as.Date('2010/1/1')) / 365.25

toPredict = expand.grid(
  timeInt = as.numeric(seq(
    as.Date('2007/1/1'), as.Date('2018/12/1'), by = '1 day'
  )),
  bygroup = c('MetroHispanicorLatino', 'NonmetroNotHispanicorLatino'),
  cos12 = cos(timeJan),
  sin12 = sin(timeJan),
  cos6 = cos(timeJan / 2),
  sin6 = sin(timeJan / 2)
)

predictGam = mgcv::predict.gam(res, toPredict, se.fit = TRUE)
predictGamm = predict(res2$gam, toPredict, se.fit = TRUE)
```

```
ranef2 = lme4::ranef(res2$mer, condVar = TRUE, whichel = 'bygroup:timeInt')
ranef2a = exp(cbind(est = ranef2[[1]][[1]],
                    se = sqrt(attributes(ranef2[[1]])$postVar))
              %*% Pmisc::ciMat())
```

1.

Write down statistical models corresponding to res and res2

Answer:

TODO

2.

Which of the two sets of results is more useful for investigating this research hypothesis?

Answer:

TODO

3.

Write a short report (a paragraph or two) addressing the following hypothesis: The long-term trend in sex ratios for urban Hispanics and rural Whites is consistent with the hypothesis that discrimination against Hispanics, while present in the full range of the dataset, has been increasing in severity over time.

Answer:

TODO

4.

Write a short report addressing the following hypothesis: The election of Trump in November 2016 had a noticeable effect on the sex ratio of Hispanic-Americans roughly 5 months after the election.

Answer:

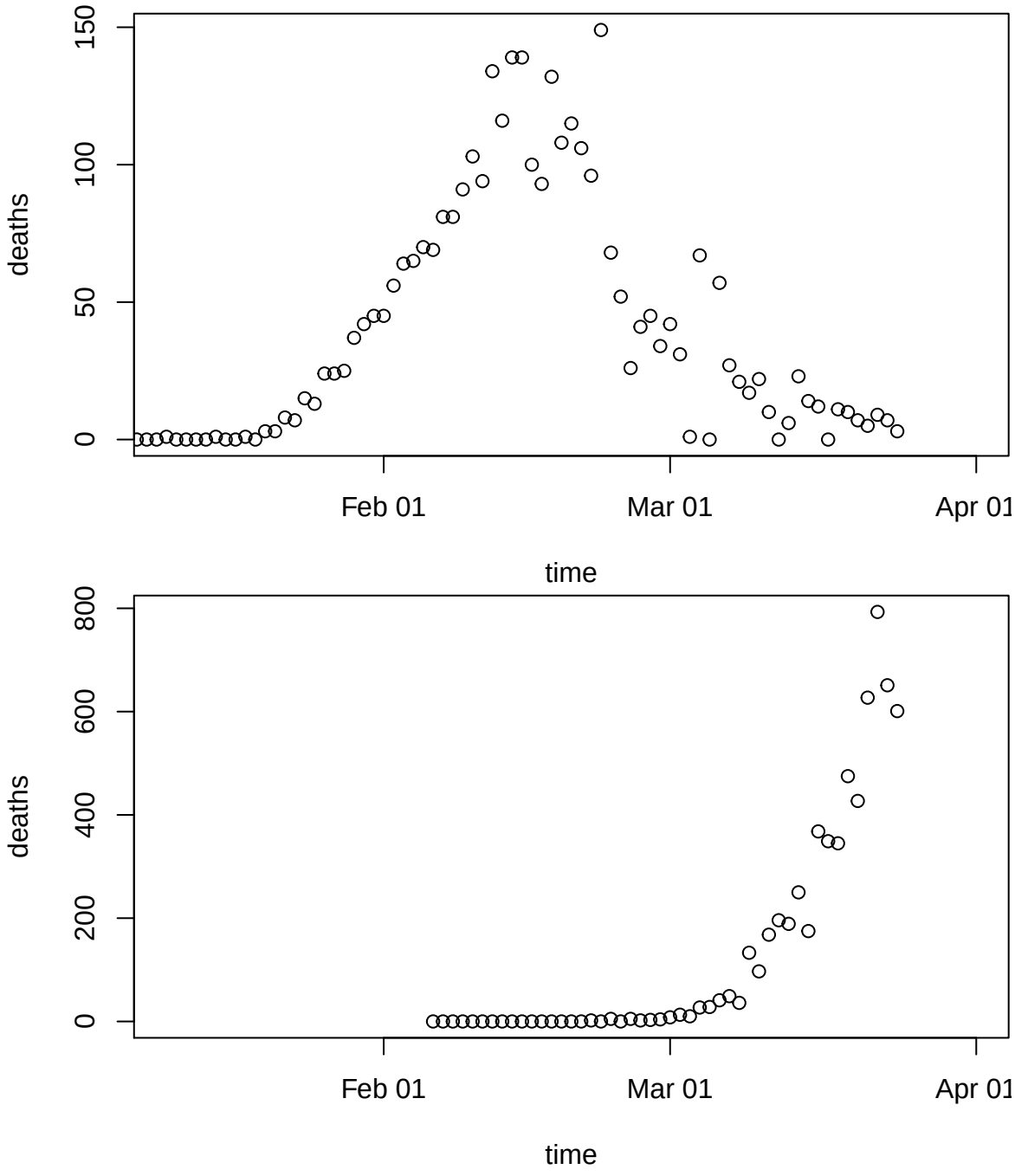
TODO

Question 2

```
if(!requireNamespace("nCov2019")) {
  devtools::install_github("GuangchuangYu/nCov2019")
}

x1 <- nCov2019::load_nCov2019(lang = 'en')

hubei = x1$province[which(x1$province$province == 'Hubei'), ]
hubei$deaths = c(0, diff(hubei$cum_dead))
```

```r
italy = x1$global[which(x1$global$country == 'Italy'), ]
italy$deaths = c(0, diff(italy$cum_dead))

x = list(Hubei = hubei, Italy = italy)
```

```r
for (D in names(x)) {
  plot(x[[D]][, c('time', 'deaths')], xlim = as.Date(c('2020/1/10', '2020/4/1')))
}
```



```
x$Hubei$weekday = format(x$Hubei$time, '%a')
x$Italy$weekday = format(x$Italy$time, '%a')
x$Italy$timeInt = as.numeric(x$Italy$time)
x$Hubei$timeInt = as.numeric(x$Hubei$time)
x$Italy$timeIid = x$Italy$timeInt
x$Hubei$timeIid = x$Hubei$time
```

```
gamItaly = gamm4::gamm4(
  deaths ~ weekday + s(timeInt, k = 40),
  random = ~ (1 | timeIid),
  data = x$Italy,
  family = poisson(link = 'log')
)
```

```
gamHubei = gamm4::gamm4(
  deaths ~ weekday + s(timeInt, k = 100),
  random = ~ (1 | timeIid),
  data = x$Hubei,
  family = poisson(link = 'log')
)
```

```
lme4::VarCorr(gamItaly$mer)
```

```
##  Groups  Name        Std.Dev.
##  timeIid (Intercept) 0.18447
##  Xr      s(timeInt)  1.07429
```

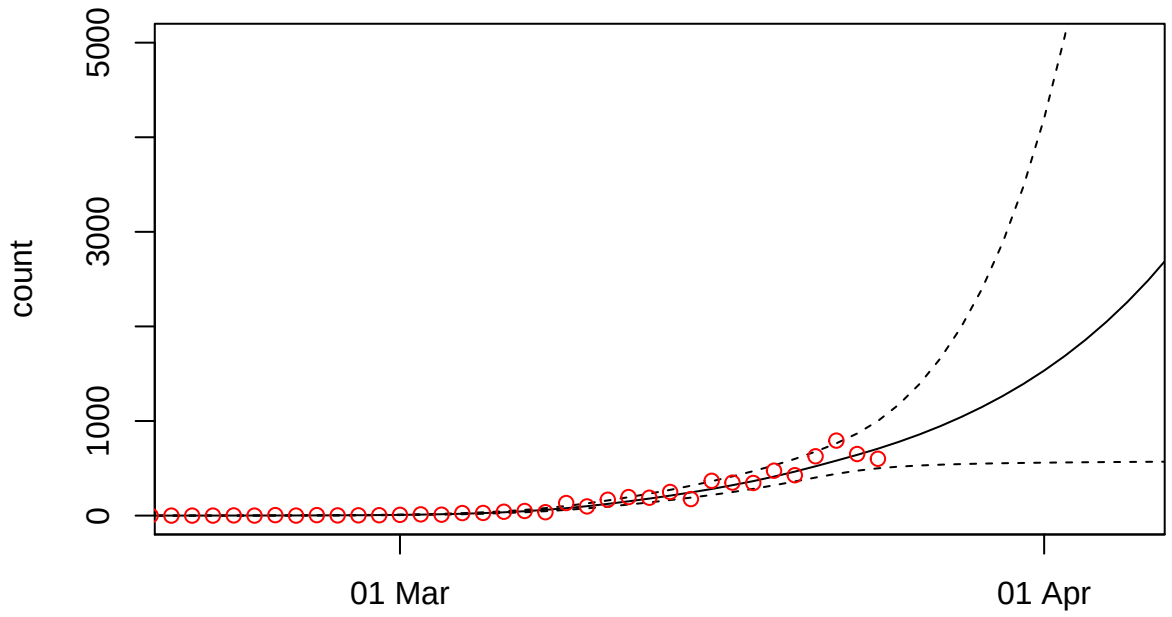
```
lme4::VarCorr(gamHubei$mer)
```

```
##  Groups  Name        Std.Dev.
##  timeIid (Intercept) 0.40748
##  Xr      s(timeInt)  3.84250
```

```
knitr::kable(cbind(summary(gamItaly$mer)$coef[, 1:2],
                   summary(gamHubei$mer)$coef[, 1:2]),
             digits = 3)
```

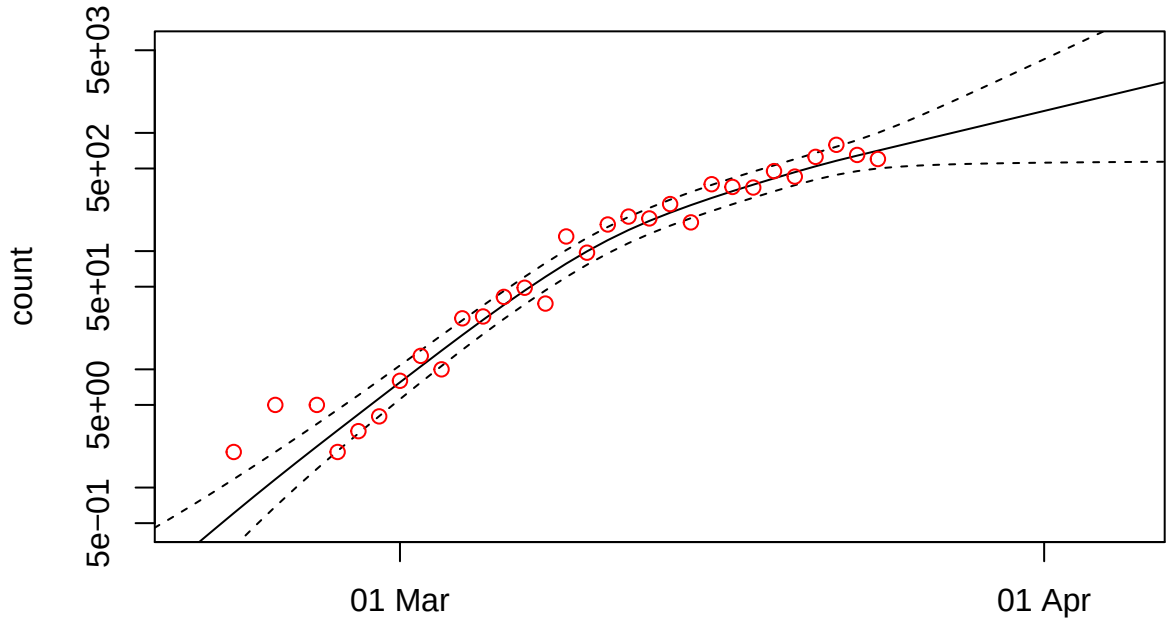
	Estimate	Std. Error	Estimate	Std. Error
X(Intercept)	1.356	0.321	-1.428	1.118
XweekdayMon	0.286	0.160	-0.126	0.210
XweekdaySat	0.133	0.163	-0.069	0.210
XweekdaySun	-0.120	0.165	0.003	0.209
XweekdayThu	0.144	0.167	-0.468	0.219
XweekdayTue	-0.088	0.164	-0.472	0.217
XweekdayWed	0.189	0.170	-0.039	0.212
Xs(timeInt)Fx1	3.062	0.901	4.731	3.978

```r
toPredict = data.frame(time = seq(as.Date('2020/1/1'), as.Date('2020/4/10'),
                                  by = '1 day'))
toPredict$timeInt = as.numeric(toPredict$time)
toPredict$weekday = 'Fri'
Stime = pretty(toPredict$time)

matplot(
  toPredict$time,
  exp(do.call(
    cbind,
    mgcv::predict.gam(gamItaly$gam, toPredict, se.fit = TRUE)
  )
  %*% Pmisc::ciMat()),
  col = 'black',
  lty = c(1, 2, 2),
  type = 'l',
  xaxt = 'n',
  xlab = '',
  ylab = 'count',
  ylim = c(0.5, 5000),
  xlim = as.Date(c('2020/2/20', '2020/4/5'))
)
axis(1, as.numeric(Stime), format(Stime, '%d %b'))
points(x$Italy[, c('time', 'deaths')], col = 'red')
```

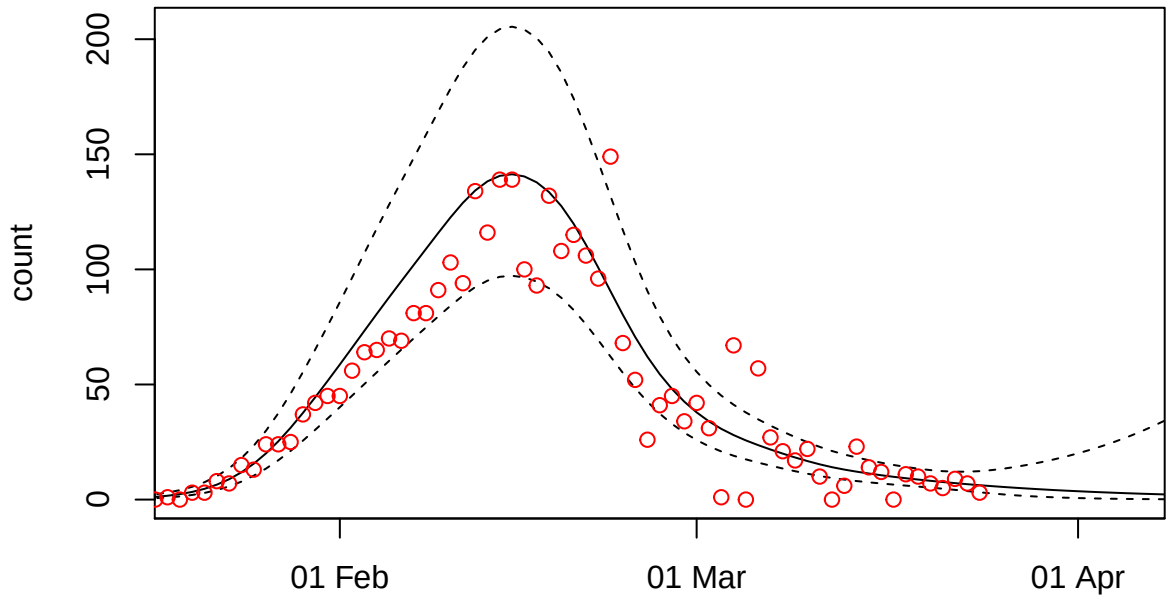



```r
matplot(
  toPredict$time,
  exp(do.call(
    cbind,
    mgcv::predict.gam(gamItaly$gam, toPredict, se.fit = TRUE)
  )
  %*% Pmisc::ciMat()),
  col = 'black',
```

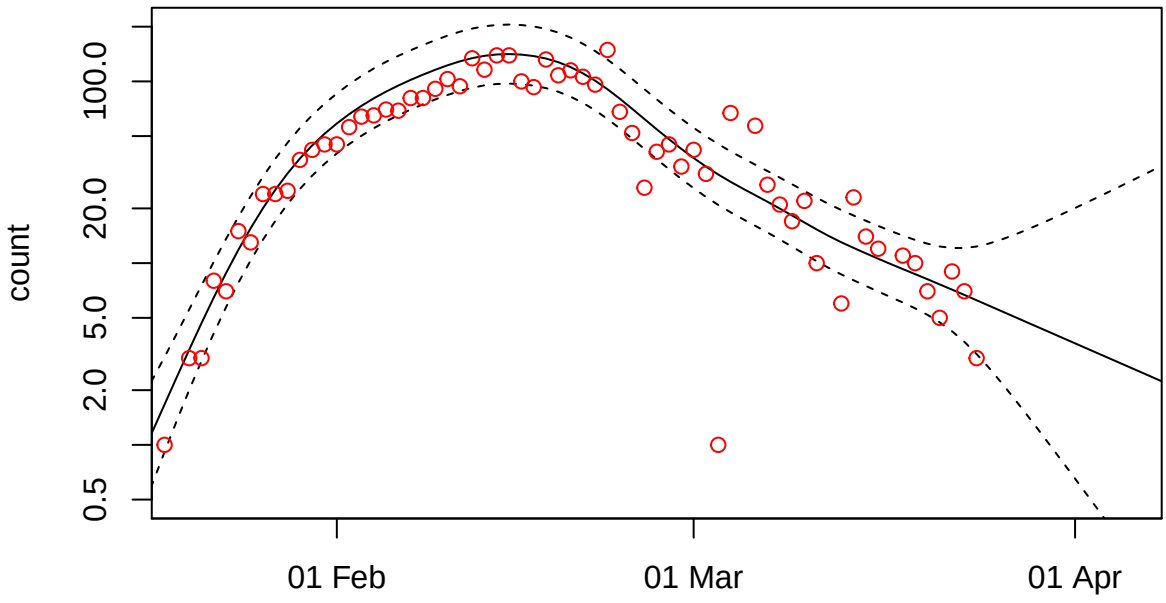
```
  lty = c(1, 2, 2),
  type = 'l',
  xaxt = 'n',
  xlab = '',
  ylab = 'count',
  ylim = c(0.5, 5000),
  xlim = as.Date(c('2020/2/20', '2020/4/5')),
  log = 'y'
)
axis(1, as.numeric(Stime), format(Stime, '%d %b'))
points(x$Italy[, c('time', 'deaths')], col = 'red')
```




```
matplot(
  toPredict$time,
  exp(do.call(
    cbind,
    mgcv::predict.gam(gamHubei$gam, toPredict, se.fit = TRUE)
  ) %*% Pmisc::ciMat()),
  col = 'black',
  lty = c(1, 2, 2),
  type = 'l',
  xaxt = 'n',
  xlab = '',
  ylab = 'count',
  xlim = as.Date(c('2020/1/20', '2020/4/5'))
)
axis(1, as.numeric(Stime), format(Stime, '%d %b'))
points(x$Hubei[, c('time', 'deaths')], col =
         'red')
```




```
matplot(
  toPredict$time,
  exp(do.call(
    cbind,
    mgcv::predict.gam(gamHubei$gam, toPredict, se.fit = TRUE)
  ) %*% Pmisc::ciMat()),
  col = 'black',
  lty = c(1, 2, 2),
  type = 'l',
  xaxt = 'n',
  xlab = '',
  ylab = 'count',
  xlim = as.Date(c('2020/1/20', '2020/4/5')),
  log = 'y',
  ylim = c(0.5, 200)
)
axis(1, as.numeric(Stime), format(Stime, '%d %b'))
points(x$Hubei[, c('time', 'deaths')], col = 'red')
```



1.

Write a down the statistical model corresponding to the gamm4 calls above, explaining in words what all of the variables are.

Answer:

TODO

2.

Write a paragraph describing, in non-technical terms, what information the data analysis presented here is providing. Write text suitable for a short 'Research News' article in a University of Toronto news publication, assuming the audience knows some basic statistics but not much about non-parametric modelling.

Answer:

TODO

3.

Explain, for each of the tests below, whether the test is a valid LR test and give reasons for your decision.

```
Hubei2 = gamm4::gamm4(
  deaths ~ 1 + s(timeInt, k = 100),
  random = ~ (1 | timeIid),
  data = x$Hubei,
  family = poisson(link = 'log'),
  REML = FALSE
)
Hubei3 = mgcv::gam(
  deaths ~ weekday + s(timeInt, k = 100),
  data = x$Hubei,
```

```
  family = poisson(link = 'log'),
  method = 'ML'
)
Hubei4 = lme4::glmer(
  deaths ~ weekday + timeInt + (1 | timeIid),
  data = x$Hubei,
  family = poisson(link = 'log')
)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 3.55896 (tol = 0.001, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```
lmtest::lrtest(Hubei2$mer, gamHubei$mer)
```

```
## Likelihood ratio test
##
## Model 1: y ~ X - 1 + (1 | Xr) + (1 | timeIid)
## Model 2: y ~ X - 1 + (1 | Xr) + (1 | timeIid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -298.89
## 2  10 -293.44  6 10.897    0.09162 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
nadiv::LRTest(logLik(Hubei2$mer), logLik(gamHubei$mer), boundaryCorrect = TRUE)
```

```
## $lambda
## 'log Lik.' -10.89679 (df=4)
##
## $Pval
## 'log Lik.' 0.5 (df=4)
##
## $corrected.Pval
## [1] TRUE
```

```
lmtest::lrtest(Hubei3, gamHubei$mer)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "gam", updated model is of class "glmerMod"

## Likelihood ratio test
##
## Model 1: deaths ~ weekday + s(timeInt, k = 100)
## Model 2: y ~ X - 1 + (1 | Xr) + (1 | timeIid)
##      #Df  LogLik      Df  Chisq Pr(>Chisq)
## 1 24.198 -304.74
## 2 10.000 -293.44 -14.198 22.611     0.0669 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
nadiv::LRTest(logLik(Hubei3), logLik(gamHubei$mer), boundaryCorrect = TRUE)
```

```
## $lambda
## 'log Lik.' -22.61127 (df=24.19792)
##
## $Pval
## 'log Lik.' 0.5 (df=24.19792)
##
## $corrected.Pval
## [1] TRUE
```

```
lmtest::lrtest(Hubei4, gamHubei$mer)
```

```
## Likelihood ratio test
##
## Model 1: deaths ~ weekday + timeInt + (1 | timeIid)
## Model 2: y ~ X - 1 + (1 | Xr) + (1 | timeIid)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -385.05
## 2  10 -293.44  1 183.22  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
nadiv::LRTest(logLik(Hubei4), logLik(gamHubei$mer), boundaryCorrect = TRUE)
```

```
## $lambda
## 'log Lik.' -183.219 (df=9)
##
## $Pval
## 'log Lik.' 0.5 (df=9)
##
## $corrected.Pval
## [1] TRUE
```

```
lmtest::lrtest(Hubei2$mer, Hubei3)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "glmerMod", updated model is of class "gam"
```

```
## Likelihood ratio test
##
## Model 1: y ~ X - 1 + (1 | Xr) + (1 | timeIid)
## Model 2: deaths ~ weekday + s(timeInt, k = 100)
##      #Df  LogLik     Df  Chisq Pr(>Chisq)
## 1  4.000 -298.89
## 2 24.198 -304.74 20.198 11.714     0.9256
```

```
nadiv::LRTest(logLik(Hubei2$mer), logLik(Hubei3), boundaryCorrect = TRUE)
```

```
## $lambda
## 'log Lik.' 11.71448 (df=4)
##
## $Pval
## 'log Lik.' 0.0003100782 (df=4)
##
## $corrected.Pval
## [1] TRUE
```

Answer:

TODO