

STA303 Assignment 3

Patrick Brown and Liza Bolton, University of Toronto

Winter 2020

1 Q1 Birth

The story in the Globe and Mail titled “Fewer boys born in Ontario after Trump’s 2016 election win, study finds” at www.theglobeandmail.com/canada/article-fewer-boys-born-in-ontario-after-trumps-2016-election-win-study refers to the paper by Retnakaran and Ye (2020). The hypothesis being investigated is that following the election of Donald Trump the proportion of babies born who are male fell. Women in the early stages of pregnancy are susceptible to miscarriage or spontaneous abortion when put under stress, and for biological reasons male fetuses are more at risk than female fetuses. Retnakaran and Ye (2020) use birth data from Ontario, and found the reduction in male babies was more pronounced in liberal-voting areas of the province than conservative-voting areas. Births in March 2017, which would have been 3 or 4 months gestation at the time of the November 2016 election, are shown to be particularly affected by the results of the election.

For testing the hypothesis that stress induced by Trump’s election is affecting the sex ratio at birth, the choice of Ontario as the study population by Retnakaran and Ye (2020) is an odd one. The dataset below considers was retrieved from wonder.cdc.gov, and contains monthly birth counts in the US for Hispanics and Non-Hispanic Whites, for rural and urban areas. Rural whites voted for Trump in large numbers, and would presumably not be stressed by the results of the election. Urban areas voted against Trump for the most part, and Americans of Hispanic origin had many reasons to be anxious following Trump’s election. 1 shows birth numbers and ratio of male to female births for rural Whites and urban Hispanics over time.

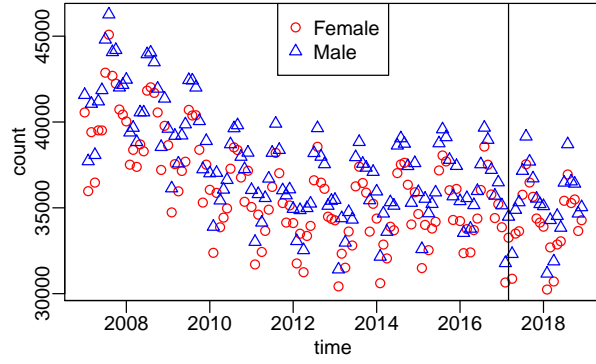
```
theFile = 'birthData.rds'
if(!file.exists(theFile)) {
  download.file('http://pbrown.ca/teaching/303/data/birthData.rds', theFile)
}
x = readRDS(theFile)
```

A Generalized Additive model was fit to these data by first defining some variables, and creating a ‘bygroup’ variable that’s a unique urban/hispanic indicator.

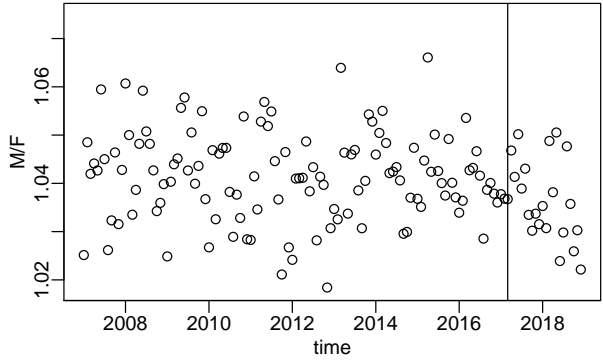
```
x$bygroup = factor(gsub("[:space:]", "", paste0(x$MetroNonmetro, x$MothersHispanicOrigin)))
x$timeInt = as.numeric(x$time)
x$y = as.matrix(x[,c('Male', 'Female')])
x$sin12 = sin(x$timeInt/365.25)
x$cos12 = cos(x$timeInt/365.25)
x$sin6 = sin(2*x$timeInt/365.25)
x$cos6 = cos(2*x$timeInt/365.25)
baselineDate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

The GAM model was fit as follows.

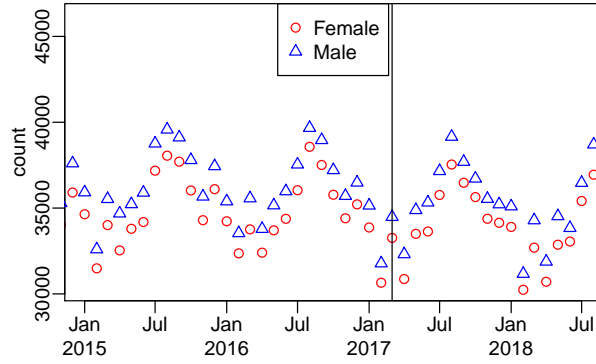
```
res = mgcv::gam(y ~ bygroup +
  cos12 + sin12 + cos6 + sin6 +
```



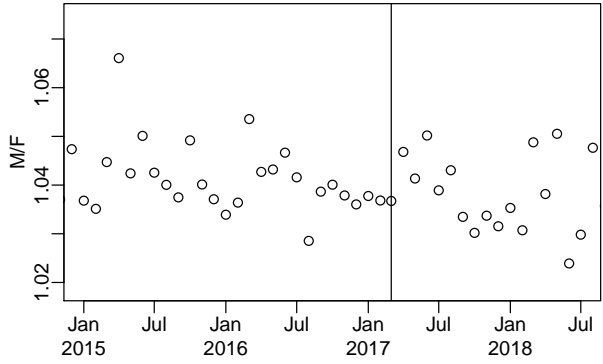
(a) Metro Hispanic or Latino



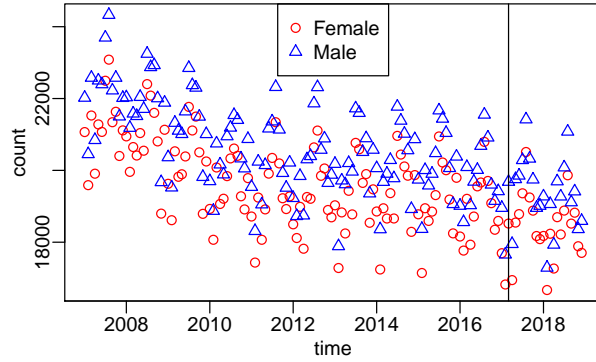
(b) Metro Hispanic or Latino



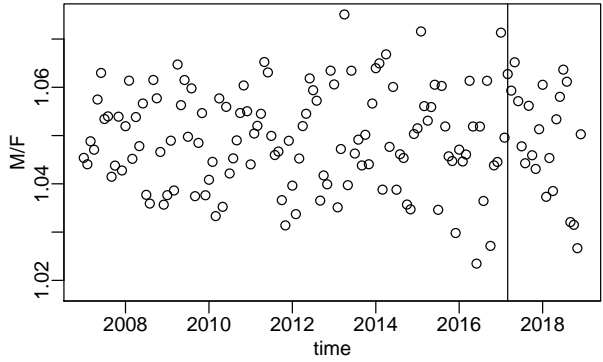
(c) Metro Hispanic or Latino



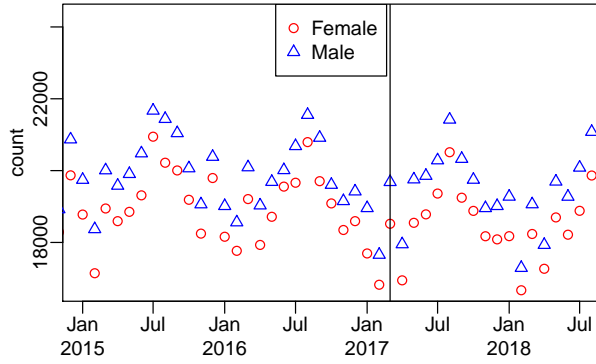
(d) Metro Hispanic or Latino



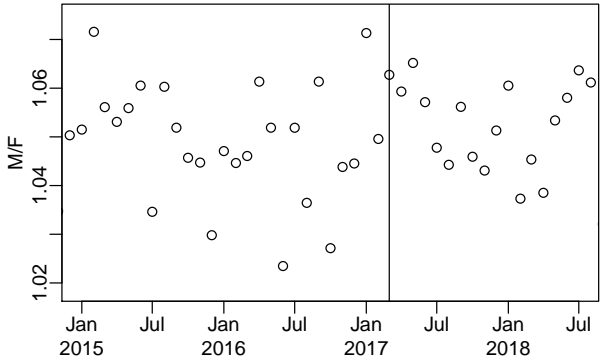
(e) Nonmetro Not Hispanic or Latino



(f) Nonmetro Not Hispanic or Latino



(g) Nonmetro Not Hispanic or Latino



(h) Nonmetro Not Hispanic or Latino

Figure 1: Monthly births of boys and girls in the US

```
s(timeInt, by=bygroup, k = 120, pc=baselineDateInt),
data=x, family=binomial(link='logit'))
```

A Generalized Linear Mixed Model was fit below.

```
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
```

Predict seasonally adjusted time trend (birth ratio assuming every month is January)

```
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)
```

these are shown in 2.

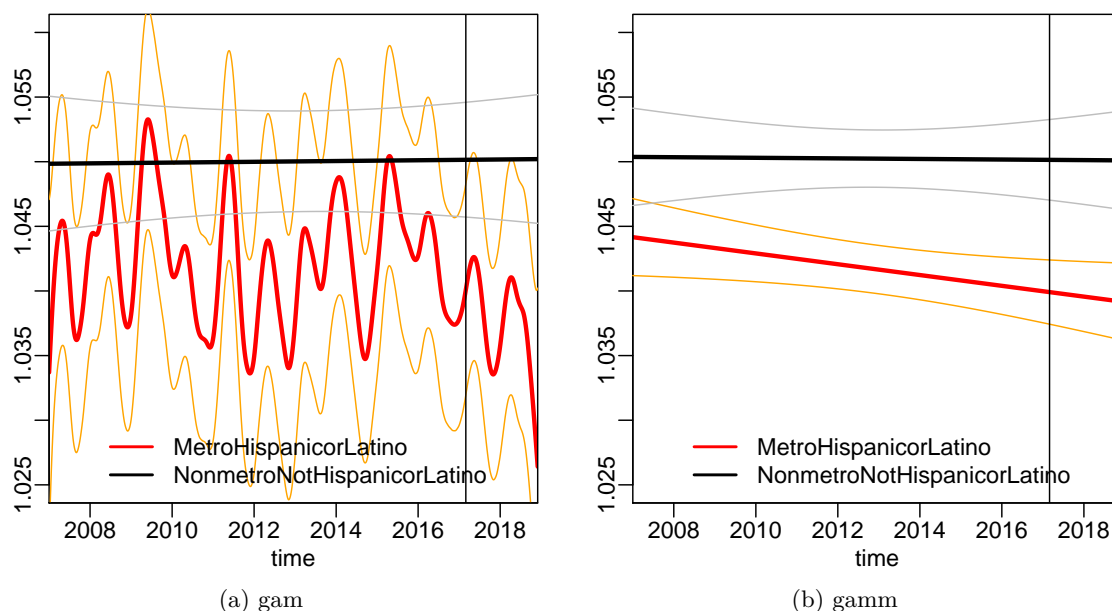


Figure 2: Predicted time trends

Predict independent random effects

```
ranef2 = lme4::ranef(res2$mer, condVar=TRUE, which1 = 'bygroup:timeInt')
ranef2a = exp(cbind(est=ranef2[[1]][[1]], se=sqrt(attributes(ranef2[[1]])$postVar))) %*% theCiMat)
```

These are shown in 3

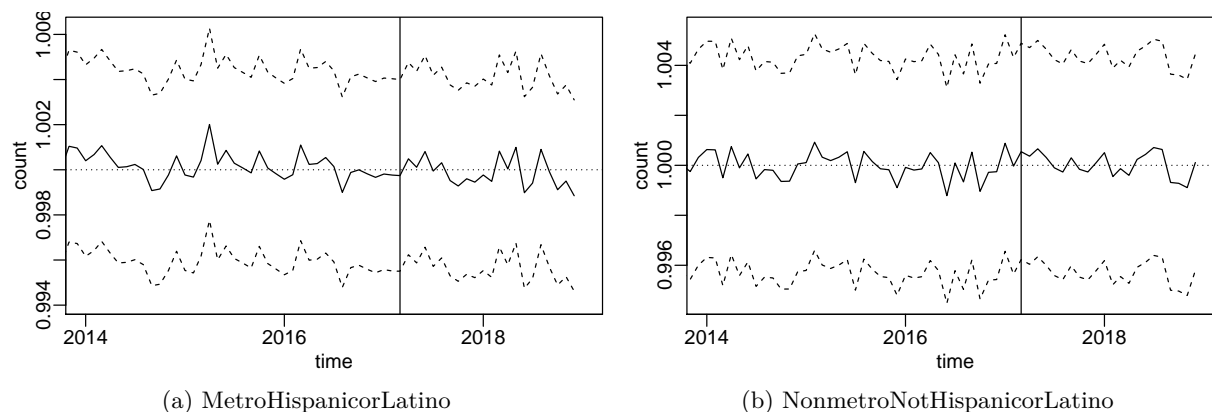


Figure 3: bygroup:timeInt random effects

1. Write down statistical models corresponding to `res` and `res2`
2. Which of the two sets of results is more useful for investigating this research hypothesis?
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.
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.

2 Q2 Death

This is the same data as you saw in the lab, but has been updated to 23 March.

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

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

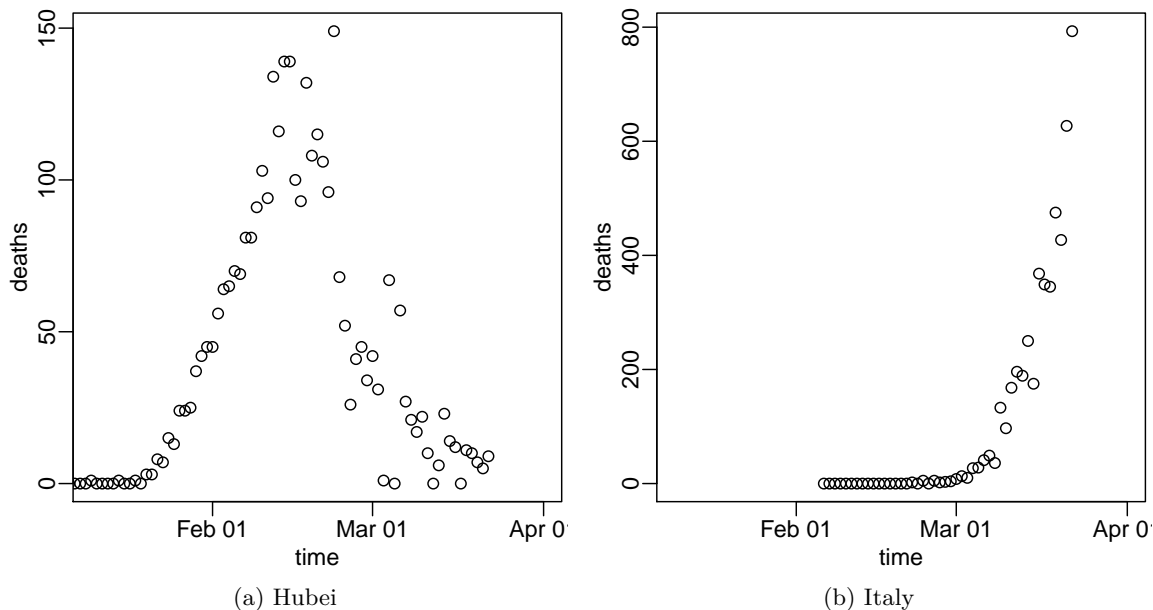


Figure 4: Covid 19 deaths

```
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$Italy$timeId = x$Italy$timeInt
x$Hubei$timeId = x$Hubei$time

gamItaly = gamm4::gamm4(deaths ~ weekday + s(timeInt, k=40), random = ~(1|timeId),
  data=x$Italy, family=poisson(link='log'))
gamHubei = gamm4::gamm4(deaths ~ weekday + s(timeInt, k=100), random = ~(1|timeId),
  data=x$Hubei, family=poisson(link='log'))

lme4::VarCorr(gamItaly$mer)

## Groups Name Std.Dev.
## timeId (Intercept) 0.10172
## Xr s(timeInt) 1.51127
```

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

## Groups Name Std.Dev.
## timeId (Intercept) 0.41303
## Xr s(timeInt) 3.44953

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)	1.000	0.422	-1.493	1.148
XweekdayMon	0.464	0.123	-0.137	0.216
XweekdaySat	0.105	0.109	-0.069	0.212
XweekdaySun	-0.171	0.121	0.002	0.212
XweekdayThu	0.210	0.117	-0.470	0.221
XweekdayTue	0.081	0.141	-0.458	0.223
XweekdayWed	0.216	0.117	-0.038	0.214
Xs(timeInt)Fx1	4.426	1.661	4.760	3.949

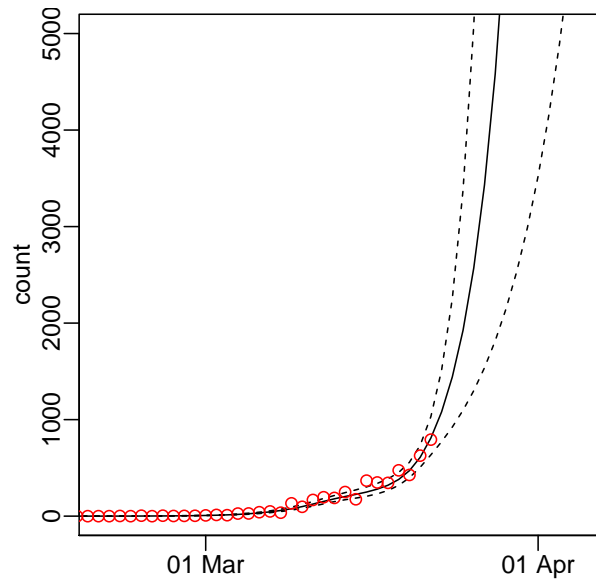
```
toPredict = data.frame(time = seq(as.Date('2020/1/1'), as.Date('2020/4/10'), by='1 day'))
toPredict$timeInt = as.numeric(toPredict$time)
toPredict$weekday = 'Fri'
Stime = pretty(toPredict$time)
matplot(toPredict$time,
  exp(do.call(cbind, mgcv::predict.gam(gamItaly$gam, toPredict, se.fit=TRUE)) %*% 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$Italy[,c('time','deaths')], col='red')
matplot(toPredict$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')))

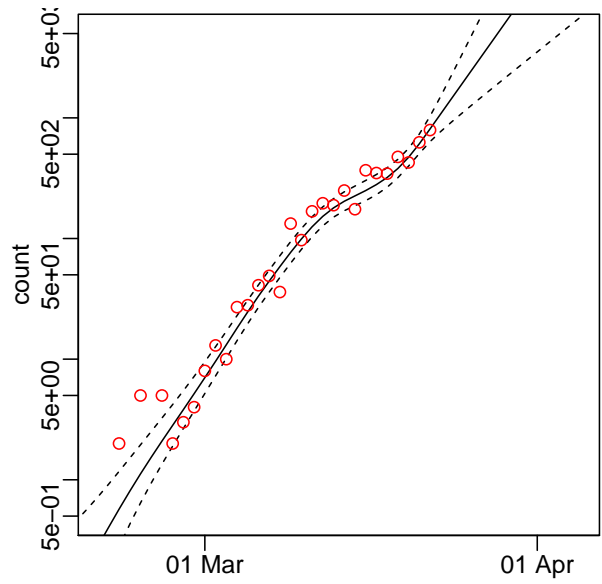
axis(1, as.numeric(Stime), format(Stime, '%d %b'))
points(x$Hubei[,c('time','deaths')], col='red')
matplot(toPredict$time,
  exp(do.call(cbind, mgcv::predict.gam(gamHubei$gam, toPredict, se.fit=TRUE)) %*% Pmisc::ciMat()),
  col='black', lty=c(1,2,2), type='l', xaxt='n', xlab='', ylab='count',
  xlim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

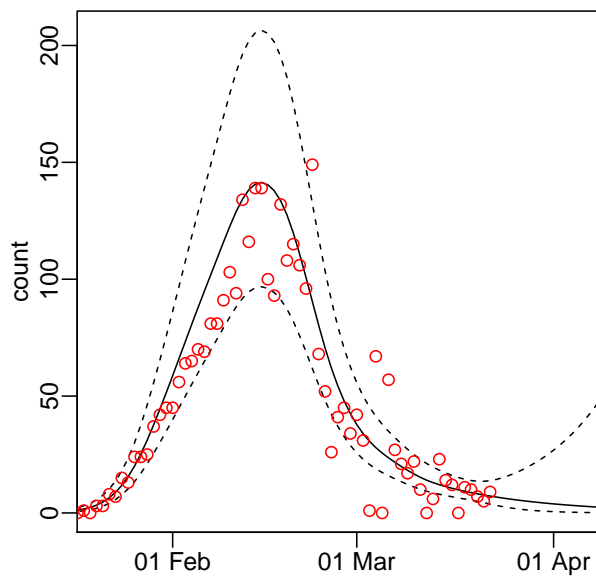
1. Write a down the statistical model corresponding to the `gamm4` calls above, explaining in words what all of the variables are.
2. Write a paragraph describing, in non-technical terms, what information the data analysis presented here is providing. Write text suitable for a short 'Research News' article in a University of Toronto news publication, assuming the audience knows some basic statistics but not much about non-parametric



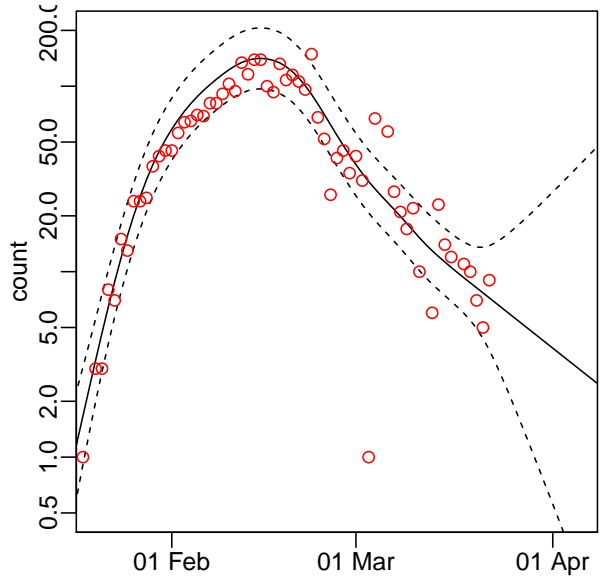
(a) Italy



(b) Italy



(c) Hubei



(d) Hubei

Figure 5: Predicted cases

modelling.

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|timeId),
  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|timeId),
  data=x$Hubei, family=poisson(link='log'))

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

## Likelihood ratio test
##
## Model 1: y ~ X - 1 + (1 | Xr) + (1 | timeId)
## Model 2: y ~ X - 1 + (1 | Xr) + (1 | timeId)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1    4 -294.10
## 2   10 -289.03  6 10.136    0.1191

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

## $lambda
## 'log Lik.' -10.13547 (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 | timeId)
##   #Df LogLik      Df Chisq Pr(>Chisq)
## 1 24.099 -300.45
## 2 10.000 -289.03 -14.099 22.84    0.06293 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

## $lambda
## 'log Lik.' -22.8398 (df=24.09876)
##
## $Pval
## 'log Lik.' 0.5 (df=24.09876)
##
## $corrected.Pval
## [1] TRUE

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



```

## Likelihood ratio test
##
## Model 1: deaths ~ weekday + timeInt + (1 | timeId)
## Model 2: y ~ X - 1 + (1 | Xr) + (1 | timeId)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1    9 -376.02
## 2   10 -289.03  1 173.98  < 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.' -173.984 (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 | timeId)
## Model 2: deaths ~ weekday + s(timeInt, k = 100)
##   #Df  LogLik      Df  Chisq Pr(>Chisq)
## 1   4.000 -294.10
## 2  24.099 -300.45 20.099 12.704    0.8897

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

## $lambda
## 'log Lik.' 12.70433 (df=4)
##
## $Pval
## 'log Lik.' 0.0001824052 (df=4)
##
## $corrected.Pval
## [1] TRUE

```

References

Retnakaran, Ravi and Chang Ye (2020). “Outcome of the 2016 United States presidential election and the subsequent sex ratio at birth in Canada: an ecological study”. In: *BMJ Open* 10.2. DOI: 10.1136/bmjopen-2019-031208.