Discrimination of American Hispanics after Trump's Election in 2016

Timothy Lee, credits to Prof Patrick Brown and Liza Bolton (University of Toronto) for starter code

April 2020

1 Context

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 @Retnakarane031208. 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. @Retnakarane031208 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.

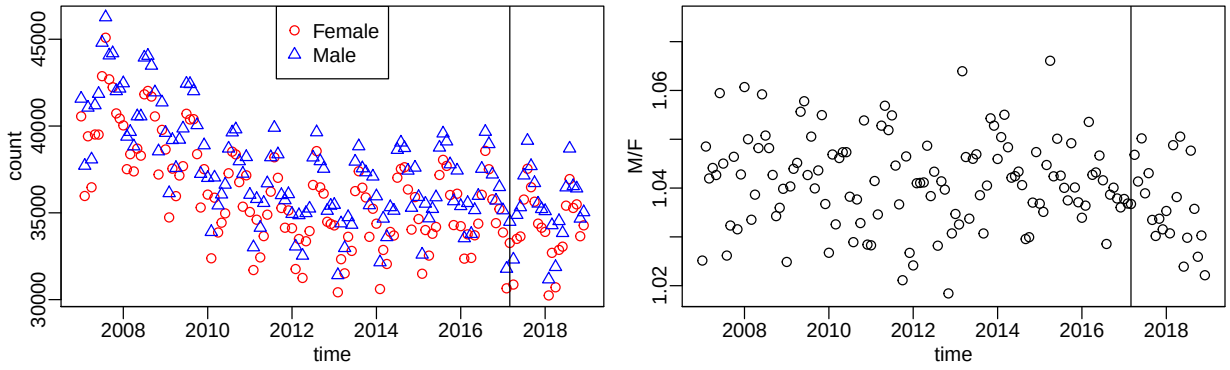
For testing the hypothesis that stress induced by Trump's election is affecting the sex ratio at birth, the choice of Ontario as the study population by @Retnakarane031208 is an odd one. The dataset below considers was retrieved from wonder.cdc.gov, and contains monthly birth counts in the US for Hispanics 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. Figure 1 below 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)
```

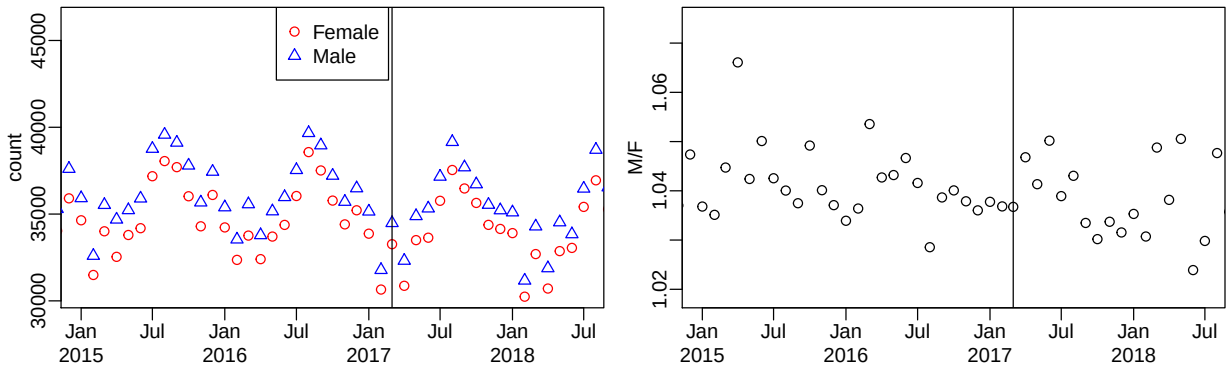
A Generalized Ardditive Model (GAM) was fit to these data by first defining some variables, and creating a 'bygroup' variable that's a unique urban/rural and White/Hispanic indicator.

```
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)
```

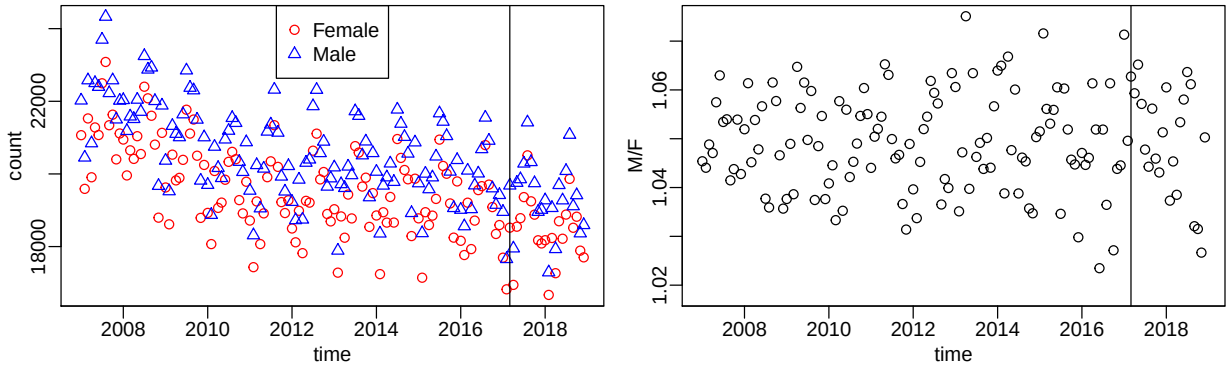
The GAM model was fit as follows, using cross-validation.



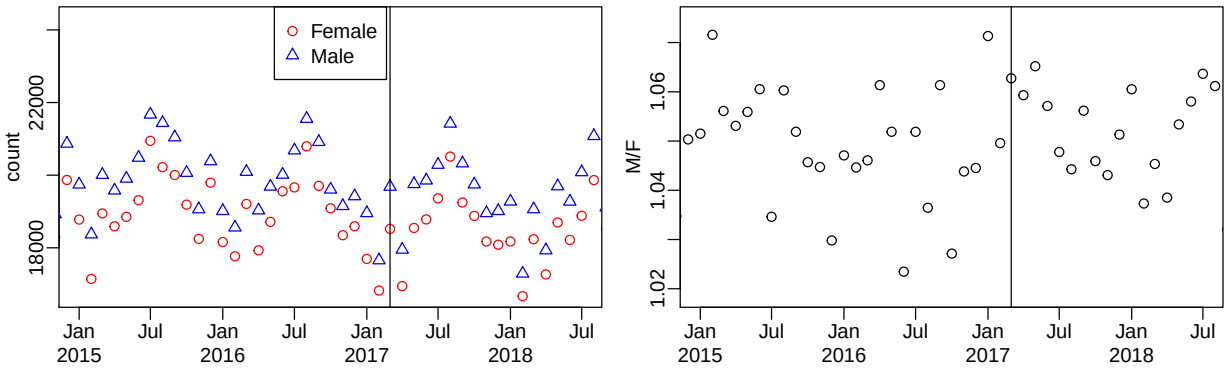
(a) Metro Hispanic or Latino (b) Metro Hispanic or Latino



(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

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

$$
\begin{aligned}
Y_{it} &\sim Binomial(N_{it}, \mu_{it}) \\
log(\frac{\mu_{it}}{1-\mu_{it}}) &= X_{it}\beta + f(t_i; \upsilon) \\
X_{it0} &= 1 \\
X_{it1} &= \cos(2\pi t_i/365.25) \\
X_{it2} &= \sin(2\pi t_i/365.25) \\
X_{it3} &= \cos(2\pi t_i/182.625) \\
X_{it4} &= \sin(2\pi t_i/182.625)
\end{aligned}
$$

, on time t of bygroup i.

μ_{it} is the proportion of male births and N_{it} is the total number of male and female births on time t of bygroup i.
We use logistic regression, where our Binomial response (log odds of the proportion of male births) is linked to a linear combination of our *bygroup* covariates of X_{it} with a logit link.
X_{it} are our *bygroup* covariates (urban Hispanics, rural Hispanics, urban White, or rural White indicator), $X_{it0}, ..., X_{it4}$ are our linear covariates (of sinusoids, to allow for potential seasonality), $f(t_i)$ is a smoothly-varying function of $timeInt$ and υ is its roughness parameter, where t_i is the unique $timeInt$ to represent time t of bygroup i.

A Generalized Linear Mixed Model was fit below, using MLE and accounted for the random effects of *bygroup* nested within each $timeInt$.

```
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'))
```

$$
\begin{aligned}
Y_{it} &\sim Binomial(N_{it}, \mu_{it}) \\
log(\frac{\mu_{it}}{1-\mu_{it}}) &= X_{it}\beta + f(t_i; \upsilon) + Z_{it} \\
Z_{it} &\sim N(0, \sigma^2) \\
X_{it0} &= 1 \\
X_{it1} &= \cos(2\pi t_i/365.25) \\
X_{it2} &= \sin(2\pi t_i/365.25) \\
X_{it3} &= \cos(2\pi t_i/182.625) \\
X_{it4} &= \sin(2\pi t_i/182.625)
\end{aligned}
$$

, on time t of bygroup i.

μ_{it} is the proportion of male births and N_{it} is the total number of male and female births on time t of bygroup i. We use logistic regression, where our Binomial response (log odds of the proportion of male births) is linked to a linear combination of our *bygroup* covariates of X_{it} and our overdispersion term of Z_{it} with a logit link. X_{it} are our *bygroup* covariates (urban Hispanics, rural Hispanics, urban White, or rural White indicator), $X_{it0}, ..., X_{it4}$ are our linear covariates (of sinusoids, to allow for potential seasonality), $f_i(t)$ is a smoothly-varying function and υ is its roughness parameter, where t_i is the unique $timeInt$ to represent time t of bygroup i.

Z_{it} is the overdispersion or the independent random effect (random intercept) for the interaction between the bygroup i and time t, i.e., it represents only the intercept varying among *timeInt* (nested) within each *bygroup* (it refers only to the interaction term and does NOT include the individual effect of *bygroup*).

Table of coefficients for both models.

```
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

Smoothing parameter for both models.

```
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
```

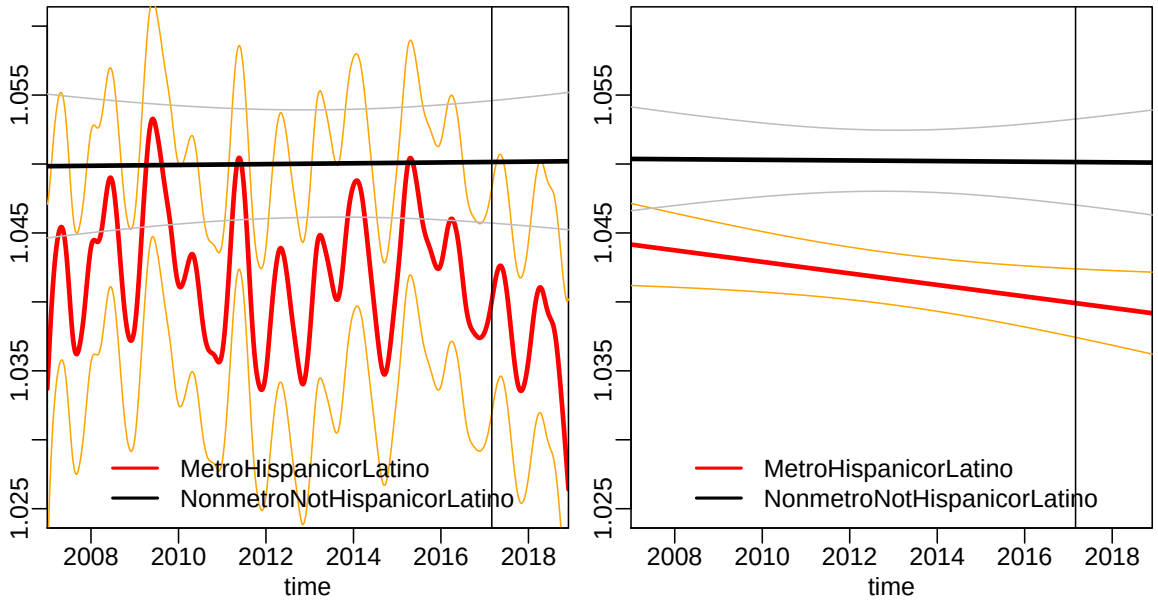
```
lme4::VarCorr(res2$mer)
```

```
##  Groups          Name                                          Std.Dev.
##  bygroup:timeInt (Intercept)                                   0.0022597
##  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 figure 2.



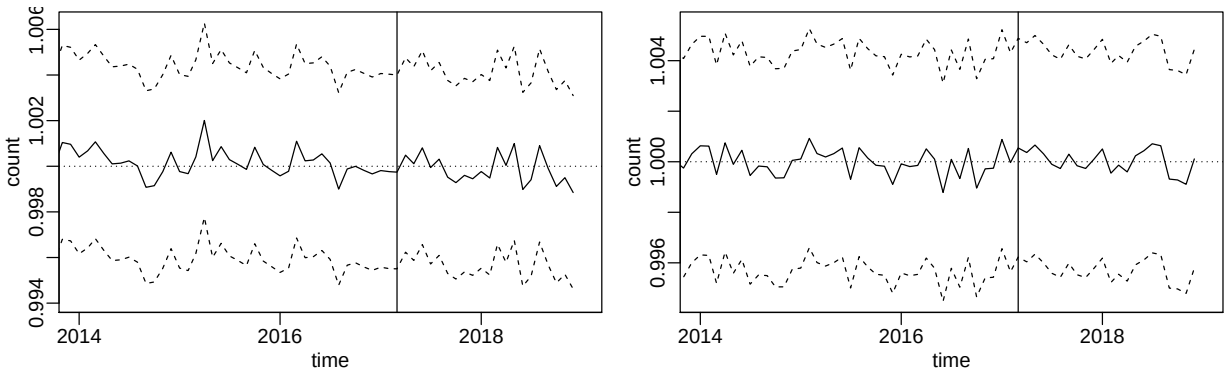
(a) gam (b) gamm

Figure 2: Predicted time trends

Predict independent random effects.

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

These are shown in figure 3.



(a) MetroHispanicorLatino (b) NonmetroNotHispanicorLatino

Figure 3: bygroup:timeInt random effects

2 Choice of model

From figure 2 (b), we can see that the gamm representing *res*2 has a much "smoother curve" compared to *res*'s gam, which has lots of extreme bumps and turns. I believe this is also more reflective of a real-world scenerio, where it is more reasonable to expect a more steady trend of the male to female birth ratio over the years compared to the gam with extreme ups and downs every year. Also, although from the gam, we

can see a sharp decreasing trend after 2016 (as per our hypothesis), it has failed to explain the sudden and unreasonable "spikes" from 2009 to 2010.

However, from figure 3, we can see that since the C.I.'s of $res2$ for both rural Whites and urban Hispanics include 1, it is possible to conclude that the overdispersion or the random effects added isn't significant. However, that is a tradeoff we will have to make as the reasons I mentioned above for figure 2 is a more serious problem. In other words, I think the discrepancy between these two models is mainly due to the methodology of the gam using cross-validation to pick the smoothing parameter (which could lead to the overfitting of the data, resulting in sudden dips and bumps) while gamm of $res2$ used maximum likelihood, resulting in a smoother and more realistic function. Hence, if we really have to pick a model from these two, I think the gamm represented by $res2$ (which also accounted for overdispersion) is the better and more realistic model to be used to investigate our research hypothesis.

3 Hypothesis 1: 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.

Firstly, as argued before, I have decided to use $res2$ (gamm) to investigate the hypothesis. Hence, based on figure 2(b), we are able to observe a trend of a steady and consistent decrease in male to female birth ratio for the urban Hispanics group, but a somewhat steady and flat maintenance for rural Whites from the period of 2008 to 2018. This is in support of our hypothesis, where we assume that discrimination against (urban) Hispanics will lead to a higher stress level for that population, and hence lower the ratio of male to female births for that group of population. Another thing to point out is that the confidence intervals for both groups in figure 2(b) didn't really overlap, suggesting that there might be a somewhat significant difference between the two groups of urban Hispanics and rural Whites (if C.I.'s overlap, then two lines could be in fact considered as "parallel". But since there is little or no overlap, we are able to reject the notion that the male to female ratios of these two groups are the same). However, it is also noteworthy to point out that there is little or no evidence that discrimination is "increasing in severity" for the urban Hispanics population, since there is only a very slight and gradual decreasing trend, indicating that the male to female birth ratios for urban Hispanics is only decreasing very slowly, yet somewhat consistently, over time. Of course, this hypothesis will only be valid under the assumption that there is a strong and definite causation that discrimination will lead to stress and thus a decrease in male to female birth ratio, however in reality, we do not know if "discrimination" is the sole factor that leads to the decrease (i.e., there might be other external/environmental factors that could lead to a decrease in male to female birth ratio for urban Hispanics).

4 Hypothesis 2: The election of Trump in November 2016 had a noticeable effect on the sex ratio of Hispanic-Americans roughly 5 months after the election.

Based on reasons listed above, the model we have chosen for this hypothesis is the gamm of $res2$. Hence, based on figure 2(b), we can see that there is somewhat a consistent and gradual decrease of male to female birth ratio starting from 2008 for the urban Hispanics group. Also, as discussed previously in Q1.3, since the confidence intervals for both groups in figure 2(b) didn't really overlap, it suggests that there might be a somewhat significant difference between the two groups of urban Hispanics and rural Whites and their corresponding male to female birth rates. However, our results are not consistent with the given hypothesis. Firstly, we don't have any data and evidence to claim that the decrease in sex ratios is caused purely by Trump's election in 2016, since the decrease in male to female birth rates are ongoing and consistently

decreasing already from year 2008. There is also not any "noticeable" drop 5 months after Trump election, which suggests that the decrease in male to female ratios for urban Hispanics is just part of the long-term trend all the way from 2008. This is also supported by the confidence intervals in figure 2 (b) from Trump's election in 2016 to 5 months after as the confidence intervals are more or less similar, suggesting that there might be no (significant/noticeable) change in male to female birth ratios at all. This is also somewhat reflected on figure 3 of the gamm, where we could see that "1" is including in the C.I. of both urban Hispanics and rural Whites, suggesting that the effect of bygroup race in a given specific time (or a particular year such as 2016 in our case) doesn't seem to have a (statistically) significant effect. The plots on figure 3 for both groups also show a similar trend, without any sudden fluctuations/dramatic changes after Trump's election in 2016 which again suggests that the decrease in ratio of male to female births for urban Hispanics is a long term effect.

Moreover, the hypothesis in Q1.4 is talking about the general group of "Hispanic-Americans", but figure 2 only includes the group of "urban" Hispanics and not rural Hispanics, which might potentially be different. Hence, it is logical to conclude that the given hypothesis isn't supported by our data and analysis using the gamm of $res2$.