STA303 A3

Manyi Luo - 1003799419

4/6/2020

Question 1

a.

Write down statistical models corresponding to res and res2.

The statistical model corresponding to res is called Generalized Additive Model (GAM), which can be expressed as following:

$$y_{ij} \sim Binomial(N, \pi_i)$$

$$logit(\pi_i) = x_i\beta + f_i(t)$$

whereas y_{ij} is the response variable including two levels (male, female), and it represents the jth observation from ith group;

x_i is a vector of covariates including bygroup and four basis functions defined already (fixed effect):

x_{i0} - bygroup, which is an unique urban/hispanic indicator variable representing four levels (Hispanics and Non-Hispanic Whites, for both rural and urban areas);

x_{i1} - $cos(\frac{\pi t_i}{356.25})$

x_{i2} - $sin(\frac{\pi t_i}{356.25})$

x_{i3} - $cos(\frac{2\pi t_i}{356.25})$

x_{i4} - $sin(\frac{2\pi t_i}{356.25})$

$f(t_i)$ is the smooth function for s(timeInt); i is determined by four levels in bygroup.

The statistical model corresponding to res2 is called Generalized Addictive Mixed Model (GAMM), which can be expressed by following:

$$y_{ij} \sim Binomial(N, \pi_i)$$

$$logit(\pi_i) = x_i\beta + f_i(t) + U_{it}$$

This model also include components in res, which has the same explanation as above. It also includes one more term, U_{it}, to describe independent random effect for each day, which comes from random = ~(1| bygroup:timeInt).

b.

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

The second set of results differentiates from the first set by adding random effect to capture the independent effect for each day. According to figure 2 part (b), it represents the model with random effects and it shows a consistent downward trend of male to female birth ratio for NonmetroNotHispanicorLatino group from 2008 to 2018. Since the data is dependent day by day, the addition of random effects results overdispersion, which can help to explain our data better.

Also, GAMM (second set of results) differentiates from GAM model (first set of results) by using the maximum likelihood to pick the smoothing parameter instead of using the cross validation. We tend to trust maximum likelihood, which can give us a more sensible result.

However, the results of lme4 :: VarCorr (res2$mer) shows that the variances of the random effect are nearly zero. This means it's also sufficient to use the first set of results if we want to avoid a more complicated model.

Therefore, both sets of results are useful in different aspects to investigate this research hypothesis and the second set of results is more preferred since it has two advantages.

c.

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.

The discrimination against Hispanics has been increasing in severity over time, and this can be illustrated through both table results and figure. First of all, the summary table with estimated coefficients indicates there exists a positive relationship between male to female births ratio and bygroupNonmetroNotHispanicorLatino, and there exists a negative relationship between male to female births ratio and bygroupNonmetroHispanicorLatino. Thus, we can conclude that Hispanic or Latino origin is significant in resulting negative births ratio.

Also, according to figure 2, the male to female ratio for NonmetroNotHispanicorLatin stays unchanged in 1.050 from 2008 to 2018, which can be illustrated by a horizontal line. However, the sex ratio for MetroHispanicorLatin stays between 1.040 and 1.045, and it also follows a significant downward trend, which indicates the long-term effect of discrimination. Also, the sex ratio for NonmetroNotHispanicorLatin group is consistently higher than MetroHispanicorLatin group and their confidence interval hardly intersect over the entire time scope. Therefore, we are able to conclude that the male to female ratio is experiencing increase severity due to discrimination against Hispanics over time.

d.

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.

It untrue to say that the election of Trump in November 2016 had a noticeable effect on the sex ratio of Hispanic-Americans roughly 5 months after the election. From figure 3, we can identify the exponential result of the change in sex ratio over the time. Figure 3 shows the estimated odd ratio for Hispanic or Latino group lies around 1 with some small fluctuations. So the effect before taking expoential should be around zero, which is a significantly small effect in short run.

Also, the confidence interval of random effects includes 1, which is insignificant. Therefore, we are unable to make a conclusion about Trump's election can create noticeable effects on the sex ratio for both of the two groups and the hypothesis is not convincible.

Question 2

a.

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

The statistical model corresponding to gamm4 is called Generalized Addictive Mixed Model (GAMM), which can be expressed as following:

$$y_{it} \sim Poisson(\lambda_t)$$

$$log(\lambda_t) = x_t\beta + f(t) + Z_t$$

whereas y_{it} is the response variable representing Covid-19 death for region i, which includes two levels (Italy, Hubei);

x_t is a vector of covariates representing weekday (fixed effect);

$f(t)$ is the smooth function for s(timeInt) and more specifically, 40 knots and 100 knots are used to soomth $f(t)$ in gamItaly and gamHubei seperately.

It also include Z_t to describe independent random effect for each day, which comes from random = ~(1|timelid) and

$$Z_t \sim N(0, \sigma^2)$$

.

b.

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.

The data analysis between Italy and Hubei can be conducted through both figure and table.

By reading figure 5 part (a) and part (c), we can notice that for Hubei, a sharp increase in deaths per day occurs before the middle of February, then the number of deaths per day decreases sharply afterward. However, in Italy, there is nearly no change in deaths per day before March, and the number of deaths per day increases sharply after the beginning of March and continues to grow through out the month.

Also, in order to see the death rate per day, we can further look into figure 5 part (b) and part (d). We can see that both Italy and Hubei experience a sharp increase in deaths around the first month of their outbreak. Hubei seems to have a faster rate of growth, as its slope is steeper. We can also conclude Hubei has already entered the last stage of virus outbreak due to the shape of rise and fall pattern. In contrast, Italy just enters the early stage of outbreak and it will follow the pattern in Hubei like what is predicted by non-linear model.

What's more, when looking into the summary table, we can find an opposite relationship between weekday and deaths in Italy and Hubei, as Italy has positive deaths number except Sunday in a week. In Huibei, there is only positive deaths number on Sunday. We can also see that weekday has more significant relationship with deaths number in Italy when conducting t-test to examine its significance.

c.

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

LR test 1: Hubei2 vs gamHubei

First of all, lmtest is not a valid test. Reading through code, gamHubei uses REML as default method, whereas Hubei2 uses REML = False. They use different methods and we should not use REML to conduct likelihood ratio tests.

Nadiv is not a valid test. Hubei2 differentiates from gamHubei by having one more fixed covariate, so Nadiv can not be applied in this case, since random effect and boundary correction is not necessary. And the boundary correction here is invalid too. Thus, it's not a valid LR test.

LR test 2: Hubei3 vs gamHubei

First of all, lmtest is not a valid test. This is because gamHubei and Hubei3 use different methods. Specifically, gamHubei uses REML as default and Hubei3 uses ML. ML is used when we are interested in fixed effect, and in contrast, REML is more likely to be used when we are interested in random effect. Similarily, we should not use REML to conduct likelihood ratio tests.

On the other hand, Nadiv is a valid test. Hubei3 and gamHubei have nested relationship, as gamHubei has one more random effect (random = ~(1|timeId)) comparing to Hubei3. The result from Nadiv shows that the random effect is insignificant in model gamHubei; Specifically, p-value is greater than 0.05, which fails to reject the null hypothesis $\sigma^2 = 0$. So the simple model should be selected at a 95% confidence level. What's more, since components from two models only different in random effect, the boundary correction here is valid, making it a LR test. However, the sequence of models are wrong, so the p-value will be useless. Thus, it's not a valid LR test.

LR test 3: Hubei4 vs gamHubei

First of all, lmtest is not a valid test. Since gamHubei uses REML as default, we shouldn't be testing REML with likelihood ratio tests. Nadiv test is valid, since Hubei4 and gamHubei are nested models. Hubei4 is a special case of gamHubei and gamHubei is larger than model Hubei4, since timeInt is a special form of smooth function f(timeInt). But due to REML, it's not a valid LR test.

LR test 4: Hubei2 vs Hubei3

First of all, lmtest is not a valid test. Obviously, Hubei2 and Hubei3 are not nested models: eventhough both of the models have the same soomth function (s(timeInt, k=100)), but Hubei2 has one more random effect (random = ~(1|timeId)) comparing to Hubei3 and Hubei3 has one more fixed effect term (weekday) comparing to Hubei2. What's more, Hubei3 uses ML method, which is different from Hubei2. Nadiv test is not valid, since they are not even nested. Thus, it's not a valid LR test.