A3 Soln

Xiangyu Kong, 1002109620

03 April, 2020

Question 1 Birth

1.

Write down statistical models corresponding to res and res2

Answer:

The statistical model for `res` is

$$
\begin{gathered}
Y_i \sim Binomial(\lambda_i, N_i) \\
h(\lambda_i) = log(\frac{\lambda_i}{1-\lambda_i}) = X_i\beta + f(W_j; v) + \epsilon_i
\end{gathered}
$$

Where

- Y_i is the response variable. It represents the number of babies that are males for group i.
- λ_i is the proportion of male babies in group i.
- N_i is the total number of babies in group i.
- $h(\lambda_i)$ is the logit link function.
- X_i, W_i are the covariates.
  - X_i contain indicator variable `bygroup`, numerical variables representing 12 months frequency: `cos12`, `sin12`, and 6 months frequency: `cos6`, `sin6`
  - W_i contain numeric variable `timeInt`, category variable `bygroup` and their interactions.
- β are the parameters.
- $f(w; v)$ are the smoothing functions of `timeInt` interacting with `bygroup`, with smoothness parameter v.
- ϵ_i are residuals for group i.

The statistical model for `res2` is

$$
\begin{gathered}
Y_{ij} \mid A_i, B_{ij} \sim Binomial(\lambda_{ij}, N_{ij}) \\
h(\lambda_{ij}) = log(\frac{\lambda_{ij}}{1-\lambda_{ij}}) = X_{ij}\beta + A_i + B_{ij} + f(W_{ij}; v) + \epsilon_{ij} \\
A_i \sim N(0, \sigma_A^2) \\
B_{ij} \sim N(0, \sigma_B^2)
\end{gathered}
$$

Where

- Y_{ij} is the response variable. It represents the number of babies that are males for group ij.
- λ_{ij} is the proportion of male babies in group ij.
- N_{ij} is the total number of babies in group ij.
- $h(\lambda_{ij})$ is the logit link function.
- X_{ij}, W_i are the covariates.
  - X_i contain indicator variable `bygroup`, numerical variables representing 12 months frequency: `cos12`, `sin12`, and 6 months frequency: `cos6`, `sin6`
  - W_i contain numeric variable `timeInt`, category variable `bygroup` and their interactions.
- β are the parameters.
- A_i is the ith `bygroup`'s deviation from the population average
- B_{ij} is the ith `bygroup`'s j's `timeInt`'s deviation from the population average.
- $f(w; v)$ are the smoothing functions of `timeInt` interacting with `bygroup`, with smoothness parameter v.
- ϵ_i are residuals for group i.

2.

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

Answer:

The results for `res2` is more useful for investigating the hypothesis that stress induced by Trump's election is affecting the sex ratio at birth.

The difference between the models `res` and `res2` is that `res2` contains random effects of `timeInt` nested within `bygroup` giving the model random intercepts. This accounts for the grouped effect introduced by race and areas and the time.

Considering the statement that Rural whites voted for Trump in large numbers, and would presumably not be stressed by the results of the election, and Urban areas voted against Trump for the most part, and Americans of Hispanic origin had many reasons to be anxious following Trump's election, it seems appropriate to use the grouping effect to explain the variations caused by region and race.

This is also confirmed according to the prediction graphs (Figure 2: Predicted time trends). For `res`, the predicted lines fluctuate too much. The plot for `res2` is smoother and illustrates the trend better.

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

In order to address the hypothesis that discrimination against Hispanics, while present in the full range of the dataset, has been increasing in severity over time, we look at results from the model `res2`.

By looking at the prediction graphs (Figure 2: Predicted time trends), the predictions for `res2` presents a smoother curve compared to that of `res`. From the graph, we see that rural Whites have a relatively flat curve while urban Hispanics has a downward trend. This indicates that over the timespan of 2007 to 2019, the ratio of male to female babies remains relatively the same for rural Whites and the ratio of male to female babies decreases for urban Hispanic. The random effects graph (Figure 3: bygroup:timeInt random effects) indicates that the variability explained by `bygroup:timeInt` has remained the same from 2013 to 2019.

Combining the two graphs, we conclude that the ratio of male to female babies decreases for urban Hispanic and remain roughly the same for rural Whites. The long-term trend 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. Thus we agree with the hypothesis.

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

In order to address the hypothesis 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, we look at results from the model `res`.

`res2` explains the general trend of the data, but does not explain the effect precisely to months. `res` explains month to month differences in more detail.

5 months after November 2016 (including November), is March 2017. From the prediction graphs (Figure 2: Predicted time trends), after the vertical line at March 2017,

Question 2

1.

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

Answer: TODO

The model corresponding to `gamItaly` is

$$
\begin{aligned}
Y_i \mid A_i &\sim Poisson(\lambda_i) \\
h(\lambda_i) = log(\lambda_i) &= X_i\beta + A_i + f(W_i; v) + \epsilon_i \\
A_i &\sim N(0, \sigma_A^2)
\end{aligned}
$$

Where

- Y_i is the response variable. It represents the number of deaths for group i.
- λ_i is the mean number of deaths for group i.
- $h(\lambda_i)$ is the log link function.
- X_i, W_i are the covariates.
  - X_i contain indicator variable `weekday`.
  - W_i contain numeric variable `timeInt`.
- β are the parameters.
- A_i is the ith `timeIid`'s deviation from the population average. In this case, every day has its own random intercept.
- $f(w; v)$ are the smoothing functions of `timeInt` interacting with `bygroup`, with smoothness parameter v, and 40 knots.
- ϵ_i are residuals for group i.

The model corresponding to `gamHubei` is

$$
\begin{aligned}
Y_i \mid A_i &\sim Poisson(\lambda_i) \\
h(\lambda_i) = log(\lambda_i) &= X_i\beta + A_i + f(W_i; v) + \epsilon_i \\
A_i &\sim N(0, \sigma_A^2)
\end{aligned}
$$

Where

- Y_i is the response variable. It represents the number of deaths for group i.
- λ_i is the mean number of deaths for group i.
- $h(\lambda_i)$ is the log link function.
- X_i, W_i are the covariates.
  - X_i contain indicator variable `weekday`.
  - W_i contain numeric variable `timeInt`.
- β are the parameters.
- A_i is the ith `timeIid`'s deviation from the population average. In this case, every day has its own random intercept.
- $f(w; v)$ are the smoothing functions of `timeInt` interacting with `bygroup`, with smoothness parameter v, and 100 knots.
- ϵ_i are residuals for group i.

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

The log standard deviation for `timeInt` random effect in Italy is 0.10172. The log standard deviation for `timeInt` random effect in Hubei is 0.41303.

On a typical Friday in Italy, it is likely to have 2.7182818 death cases.

On a typical Friday in Hubei, it is likely to have 0.2246976 death cases.

3.

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

Answer: TODO

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

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

lmtest::lrtest(Hubei3, gamHubei$mer)

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

lmtest::lrtest(Hubei4, gamHubei$mer)

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

lmtest::lrtest(Hubei2$mer, Hubei3)

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