A3 Soln

Xiangyu Kong, 1002109620

09 April, 2020

Question 1 Birth

Question 1.1

Write down statistical models corresponding to res and res2

Answer:

The statistical model for `res` is

$$Y_i \sim Binomial(N_i, p_i)$$
$$h(p_i) = log(\frac{p_i}{1-p_i}) = X_i\beta + s(t_i) + f(W_i; v) + \epsilon_i$$

Where

- Y_i is the response variable. It represents the number of babies that are males for group i.
- p_i is the proportion of male babies in group i.
- N_i is the total number of babies in group i.
- $h(p_i)$ is the logit link function.
- X_i, W_i are the covariates.
  - X_i contains indicator variable `bygroup` (combination of `MetroNonmetro` and `MothersHispanicOrigin`, 4 levels)
  - W_i contains numeric variable `timeInt` (the date of birth represented in numeric), interacting with indicator variable `bygroup`.
- β are the parameters.
- $f(w; v)$ are the smoothing functions of `timeInt` interacting with `bygroup`, with smoothness parameter v, up to 120 knots.
- t_i is the numeric variable `timeInt` in group i.
- $s(t_i)$ is seasonal cycle function of `timeInt`. It represents two frequencies: 12-month ($sin(\pi t_i/365.25)$, $cos(\pi t_i)$ and a 6-month ($sin(2\pi t_i/365.25)$, $cos(2\pi t_i/365.25)$).
- ϵ_i are residuals for group i.

The statistical model for `res2` is

$$Y_{it} \mid U_{it} \sim Binomial(p_{it}, N_{it})$$
$$h(p_{it}) = log(\frac{p_{it}}{1-p_{it}}) = X_{it}\beta + U_{it} + s(t_i) + f(W_{it}; v) + \epsilon_{it}$$
$$U_{it} \sim N(0, \sigma_U^2)$$

Where

- Y_{it} is the response variable. It represents the number of babies that are males for time t in bygroup level i.
- p_{it} is the proportion of male babies for time t in bygroup level i.
- N_{it} is the total number of babies for time t in bygroup level i.
- $h(p_{it})$ is the logit link function.
- X_{it}, W_{it} are the covariates.
  - X_{it} contains indicator variable `bygroup` (combination of `MetroNonmetro` and `MothersHispanicOrigin`, 4 levels).
  - W_{it} contains numeric variable `timeInt` (the date of birth represented in numeric), interacting with indicator variable `bygroup`.
- β are the parameters.
- U_{it} is the ith `bygroup` level's tth `timeInt`'s deviation from the population average.

- $f(w; v)$ are the smoothing functions of `timeInt` interacting with `bygroup`, with smoothness parameter v, up to 120 knots.
- t_{it} is the numeric variable `timeInt` for time t in bygroup level i.
- $s(t_{ij})$ is seasonal cycle function of `timeInt`. It represents two frequencies: 12-month ($sin(\pi t_{ij}/365.25)$, $cos(\pi t_{ij})$ and a 6-month ($sin(2\pi t_{ij}/365.25)$, $cos(2\pi t_{ij}/365.25)$).
- ϵ_{it} are residuals for group it.

Question 1.2

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 `res` and `res2` is that `res` is smoothed with generalized cross validation, where as `res2` is smoothed with maximum likelihood, and contains a random effect of `timeInt` interacting with `bygroup`.

From Figure 3, we can see that the random effects for `res2` are very similar. Both the exponentiated effects are scattered around 1.00 level, so the effects are almost 0. This suggests that we potentially do not need the random effects.

From Figure 2 we can see that the predicted time trends for `res` is more wiggly. The red fitted line for `NonmetroNotHispanicorLatino` fluctuates a great deal and does not provide a clear indication of the trend. The predicted time trend for `res2` is smoother and straighter, indicating that the model is more suited for explaining the trend.

The hypothesis focuses on the trend before and after Trump's election. Although we potentially do not need the random effects, `res2` explains the trend better than `res`, so `res2` is more suited to investicating this research hypothesis.

Question 1.3

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:

Using the model selected from part 2, we look at results produced by `res2`.

By looking at the prediction graphs (Figure 2: Predicted time trends), the predictions for `res2` presents two smooth curves. 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 two lines diverge as time progress, meaning that the difference of male to female ratio between rural Whites and urban Hispanic increases over time.

The 95% confidence intervals only barely overlapp on the left side, and do not overlap for most of the time, indidcating that this increasing difference is significant.

Combining the two satements that stress during pregnancy reduces the number of male babies, and racial discrimination increases stress, we can say that the increasing difference of male to female ratio between rural Whites and urban Hispanic (urban Hispanic having the lower ratio) suggests that the stress for urban Hispanic might be higher over time, and could further indicate that the discrimination against Hispanics has been increasing in severity over time.

Question 1.4

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:

Using the model selected from part 2, we look at results produced by `res2`.

By looking at the prediction graphs (Figure 2: Predicted time trends), the predictions for `res2` presents two smooth curves. From the graph, we see that urban Hispanics has a downward trend. This indicates that over the timespan of 2007 to 2019, the ratio of male to female babies decreases for urban Hispanic. The line is relatively straight, meaning from 2007 to 2019, the rate of decreasing (coresponding to the slope of the graph) is constant.

If Trump's election had a noticable effect on the sex ratio of Hispanic-Americans, the rate of the decreasing of male to female ratio would be changed, resulting in a different slope and a more wiggly line after November 2016. However, this was not shown in the prediction graph. The straight line after November 2016 with constant slope suggests that Trump's election did not have a noticable effect on the sex ratio of Hispanic-Americans. Thus we can reject 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.

Question 2

Question 2.1

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

Answer:

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 in Italy.
- 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` the day of the week (7 levels).
  - W_i contain numeric variable `timeInt` a numeric representation of date.
- β are the parameters.
- A_i is the ith `timeIid`'s (numeric representation of date) deviation from the population average. In this case, every day has its own random intercept.
- $f(w; v)$ are the smoothing functions of `timeInt`, with smoothness parameter v, up to 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 in Hubei.
- 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` the day of the week (7 levels).
  - W_i contain numeric variable `timeInt` a numeric representation of date.
- β are the parameters.
- A_i is the ith `timeIid`'s (string representation of date) deviation from the population average. In this case, every day has its own random intercept.
- $f(w; v)$ are the smoothing functions of `timeInt`, with smoothness parameter v, up to 100 knots.
- ϵ_i are residuals for group i.

The difference between the two models are

- Y_i corresponds to death cases in different regions: `gamItaly` for Italy, and `gamHubei` for Hubei.
- $f(w; v)$ has different number of knots. `gamItaly`'s smoothing function has up to 40 knots where as `gamHubei`'s smoothing function has up to 100 knots.

Question 2.2

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:

On a typical Friday in Italy, it is likely to have $exp(1.000) = 1$ death cases. Using a 95% confindence interval ($effect \pm 2 \times se$), we can see that only Monday's confidence interval does not include 0. This means that with 95% confidence, we can say that comparing to Friday, monday is more likely to have higher number of death cases in Italy, where as other days do not have a significant effect on the number of death cases in Italy.

On a typical Friday in Hubei, it is likely to have $exp(-1.493) = 0.2247$ death cases. Using a 95% confindence interval ($effect \pm 2 \times se$), we can see that all confidence intervals include 0. This means that the days of the week do not have a signifancant effect on the number of death cases in Hubei.

According to the prediction graphs in Figure 5, Italy's death cases seems to be increasing in the future, because the predicted lines and its 95% confidence interval both indicate an upward growth. Hubei's death cases seems to be decreasing in the future. However, the 95% confidence interval seems to diverge. This indicates that although we predict Hubei's death cases will decrease, we cannot make this claim with great confidence, and it is possible that Hubei's death cases will increase in the future.

If the COVID19 spread trend for Italy and Hubei are similar, we can say that Italy has not yet reached the peak of death cases, and there might be more in the future. As for Hubei, it seems that it has already reached its peak, but we can not make a confident claim about what will happen in the future.

Question 2.3

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

Answer:

- `lmtest::lrtest(Hubei2$mer, gamHubei$mer)` is not a valid LR test because `gamHubei` is fitted with REML instead of ML, and we shouldn't test REML models with likelihood ratio tests. If it was fitted with ML, the test would be valid because `Hubei2` is nested within `gamHubei`. `Hubei2` is the special case when `gamHubei` removes the `weekday` covariate and uses mean instead.
- `nadiv::LRTest(logLik(Hubei2$mer), logLik(gamHubei$mer), boundaryCorrect = TRUE)` is not a valid LR test because `gamHubei` is fitted with REML instead of ML, and we shouldn't test REML models with likelihood ratio tests. Also, although `Hubei2` is nested within `gamHubei`, we are not testing for the random effect. `gamHubei` and `Hubei2` contain the same random effects (smothing on `timeInt` and random effect of `timeIid`). `boundaryCorrect` should only be set to TRUE when we are testing for random effects.
- `lmtest::lrtest(Hubei3, gamHubei$mer)` is not a valid LR test because `gamHubei` is fitted with REML instead of ML, and we shouldn't test REML models with likelihood ratio tests. Also, although `Hubei3` is nested within `gamHubei`, the difference between the two model is the random effect of `timeIid`. To test for the random effect being significant, it is better to use `nadiv::LRTest`.
- `nadiv::LRTest(logLik(Hubei3), logLik(gamHubei$mer), boundaryCorrect = TRUE)` is a not valid LR test because `gamHubei` is fitted with REML instead of ML, and we shouldn't test REML models with likelihood ratio tests. If it was fitted with ML, the test would be valid because `Hubei3` is nested within `gamHubei`. It is equivalent to setting the random effect `timeIid` to 0 in gamHubei. By setting `boundaryCorrect = TRUE`, we are testing the significance for the random effect of `timeIid`.
- `lmtest::lrtest(Hubei4, gamHubei$mer)` is not a valid LR test because `gamHubei` is fitted with REML instead of ML, and we shouldn't test REML models with likelihood ratio tests. Also, although `Hubei4` is nested within `gamHubei`. It is equivalent to using straight lines as `f(timeInt)` in `gamHubei`. By testing the two models, we will be testing the significance of the smoothing, which is a random effect. To test for the random effect being significant, it is better to use `nadiv::LRTest`.
- `nadiv::LRTest(logLik(Hubei4), logLik(gamHubei$mer), boundaryCorrect = TRUE)` is not a valid LR test because `gamHubei` is fitted with REML instead of ML, and we shouldn't test REML models with likelihood ratio tests. If it was fitted with ML, the test would be valid because `Hubei4` is nested within `gamHubei`. It is equivalent to using straight lines as `f(timeInt)` in `gamHubei`. by setting `boundaryCorrect = TRUE`, we are testing the significance for smoothing on `timeInt`.
- `lmtest::lrtest(Hubei2$mer, Hubei3)` is not a valid LR test because `Hubei2` and `Hubei3` are not nested. `Hubei2` contains a random effect of `timeIid`, which `Hubei3` does not. `Hubei3` has `weekday` as one of its covariates, but `Hubei2` does not. Thus the two models are not nested, and the LR test is inappropriate.
- `nadiv::LRTest(logLik(Hubei2$mer), logLik(Hubei3), boundaryCorrect = TRUE)` is not a valid LR test because `Hubei2` and `Hubei3` are not nested. `Hubei2` contains a random effect of `timeIid`, which `Hubei3` does not. `Hubei3` has `weekday` as one of its covariates, but `Hubei2` does not. Thus the two models are not nested, and the LR test is inappropriate.