STA2201 Lab 11!

Timothy Jordan Regis

03/04/2023

Overview

In this lab you'll be fitting a second-order P-Splines regression model to foster care entries by state in the US, projecting out to 2030.

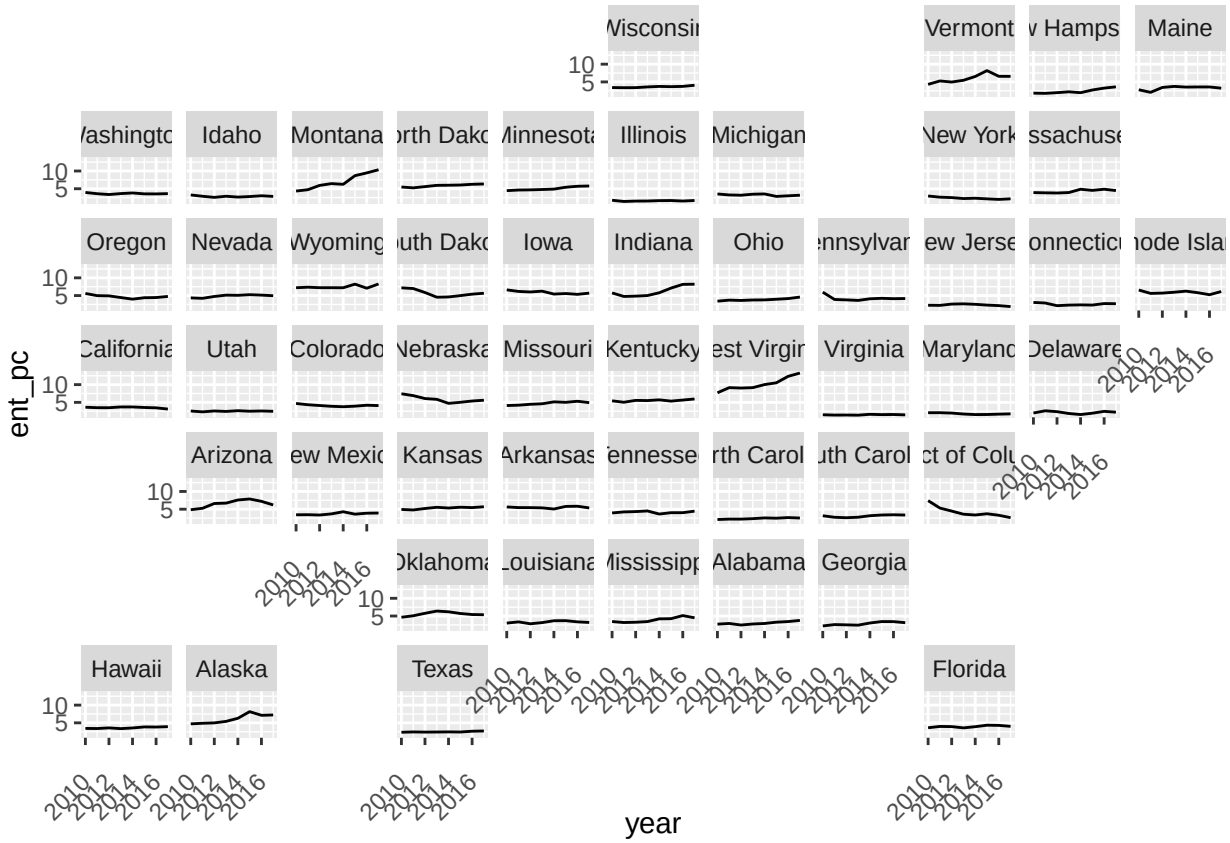
Here's the data

```
## # A tibble: 6 x 6
##    fips state    year   ent child_acs ent_pc
##   <dbl> <chr>   <dbl> <dbl>     <dbl>  <dbl>
## 1     1 Alabama  2010  3063   1131261   2.71
## 2     1 Alabama  2011  3257   1120773   2.91
## 3     1 Alabama  2012  2763   1122353   2.46
## 4     1 Alabama  2013  3041   1105933   2.75
## 5     1 Alabama  2014  3192   1101149   2.90
## 6     1 Alabama  2015  3605   1104332   3.26
```

Question 1

Make a plot highlighting trends over time by state. Might be a good opportunity to use `geofacet`. Describe what you see in a couple of sentences.

```
df11 %>% ggplot(aes(year, ent_pc)) +
  geom_line() +
  facet_geo(~state) + theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```



As we can see, when left on the same scales, variation by year is quite similar across states. The few exceptions include Montana, West Virginia, and Vermont, which show somewhat significant increasing trends over time, while the District of Columbia tends to see a decreasing relationship over time. We can also notice some states have higher baseline levels of foster care entries, such as Vermont and Arizona, however, it becomes difficult to predict why this may be the case without further investigating the characteristics from each state. Alternatively, we can also let the y-axis be free and have different scales for each plot. In doing so, we do see many different relationships, but this factor massively increases the complexity and length of interpretations, requiring us to finely explain the results of many states that would clutter up this assignment.

Question 2

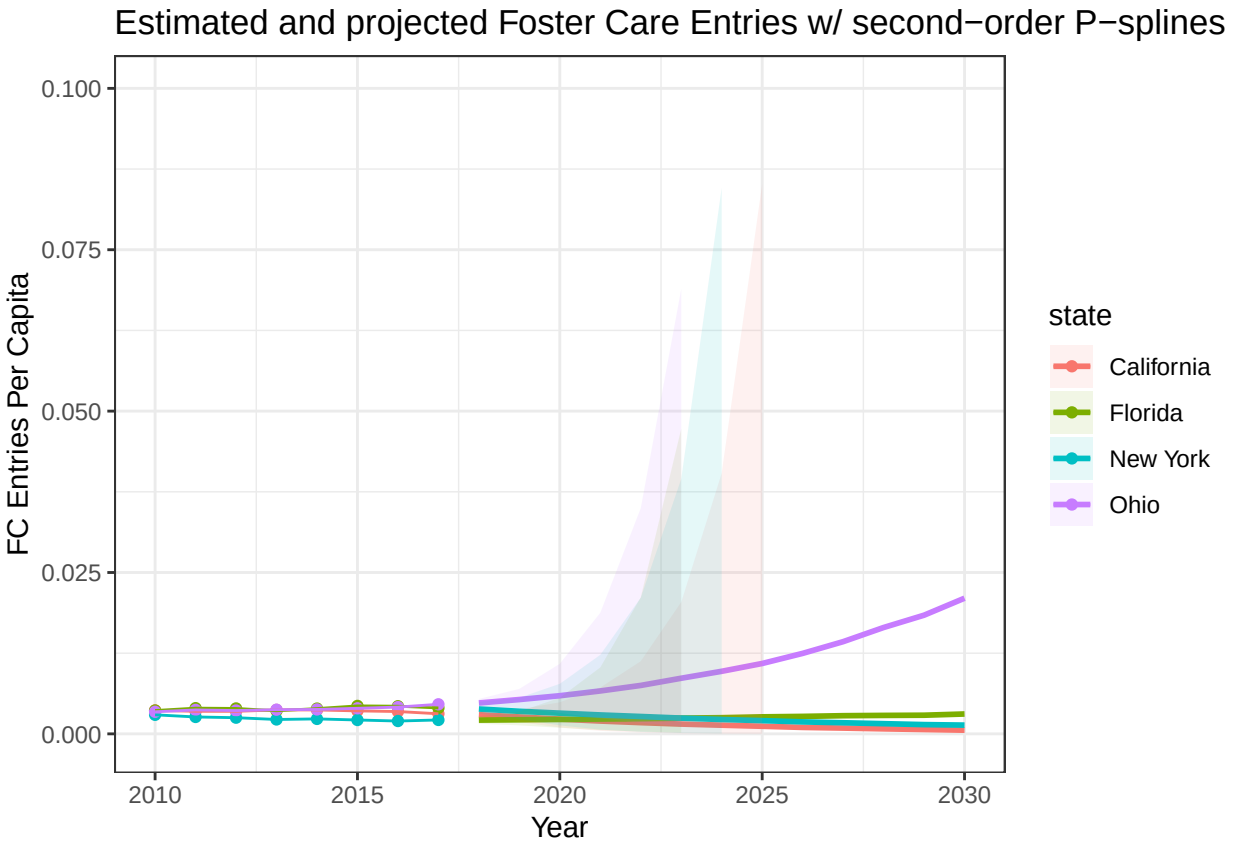
Fit a hierarchical second-order P-Splines regression model to estimate the (logged) entries per capita over the period 2010-2017. The model you want to fit is

$$\begin{aligned} y_{st} &\sim N(\log \lambda_{st}, \sigma^2_{y,s}) \\ \log \lambda_{st} &= \alpha_k B_k(t) \\ \Delta^2 \alpha_k &\sim N(0, \sigma^2_{\alpha,s}) \\ \log \sigma_{\alpha,s} &\sim N(\mu_\sigma, \tau^2) \end{aligned}$$

Where $y_{s,t}$ is the logged entries per capita for state s in year t. Use cubic splines that have knots 2.5 years apart and are a constant shape at the boundaries. Put standard normal priors on standard deviations and hyperparameters.

Question 3

Project forward entries per capita to 2030. Pick 4 states and plot the results (with 95% CIs). Note the code to do this in R is in the lecture slides.



For our plot we have chosen some of the most popularly known states including California, Florida, New York, and Ohio. As we can see, Ohio has a significantly higher predicted trend than the other states, which remain quite flat, but the high standard error on our plots suggest that we cannot be 100% certain about this conclusion.

Note: I'm not sure why the CIs behave the way they do (cutting off before reaching the end). I think this may be due to them simply being too large for R to plot, but I do believe these suggest an increasing level of variation as the predictions stretch out further as expected. I also needed to cut the y-axis off to ensure that the trends were at least somewhat visible, I think this may be causing the problem, but it is uninterpretable otherwise (Included at the end).

Question 4 (bonus)

P-Splines are quite useful in structural time series models, when you are using a model of the form

$$f(y_t) = \text{systematic part} + \text{time-specific deviations}$$

where the systematic part is model with a set of covariates for example, and P-splines are used to smooth data-driven deviations over time. Consider adding covariates to the model you ran above. What are some potential issues that may happen in estimation? Can you think of an additional constraint to add to the model that would overcome these issues

With any stan model, we can run the risk of seeing high correlations between our predictors, making it hard for the model to search the sample space and eventually converge. The common method to alleviate this issue is through the centering or standardization of our variables.

Full Plot from Q3