Week 11: Splines

04/04/23

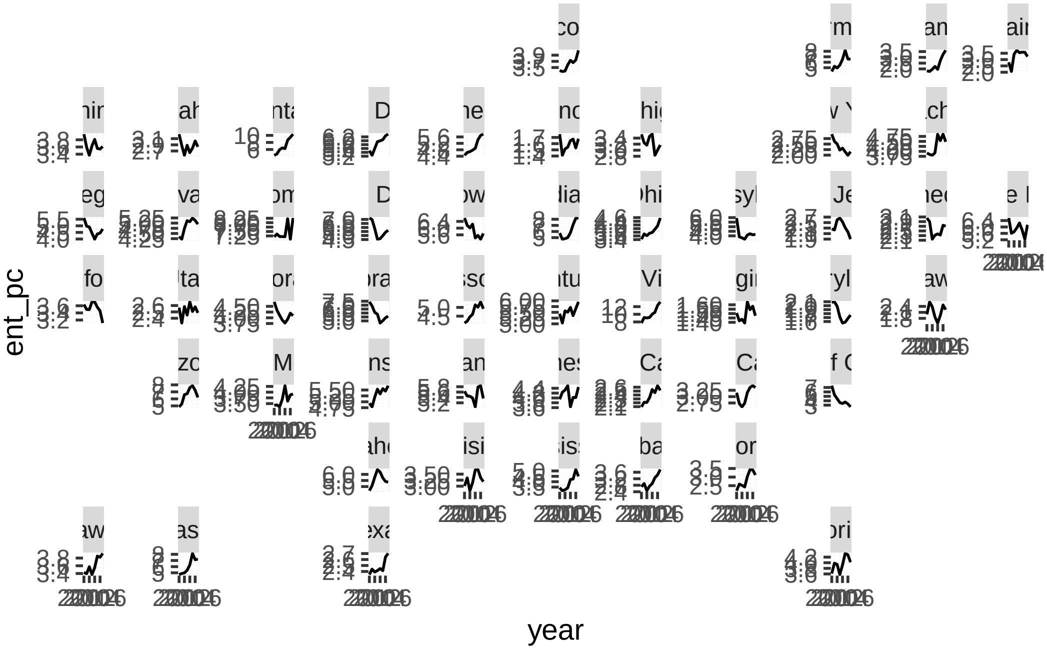
Overview

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

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.



- More states have upward trends instead of downward trends.
- States at left or right are more likely to have a downward trend, most states at middle have a upward trend.
- States at bottom are more likely to have a upward trend, states at top are more likely to have downward trends.

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

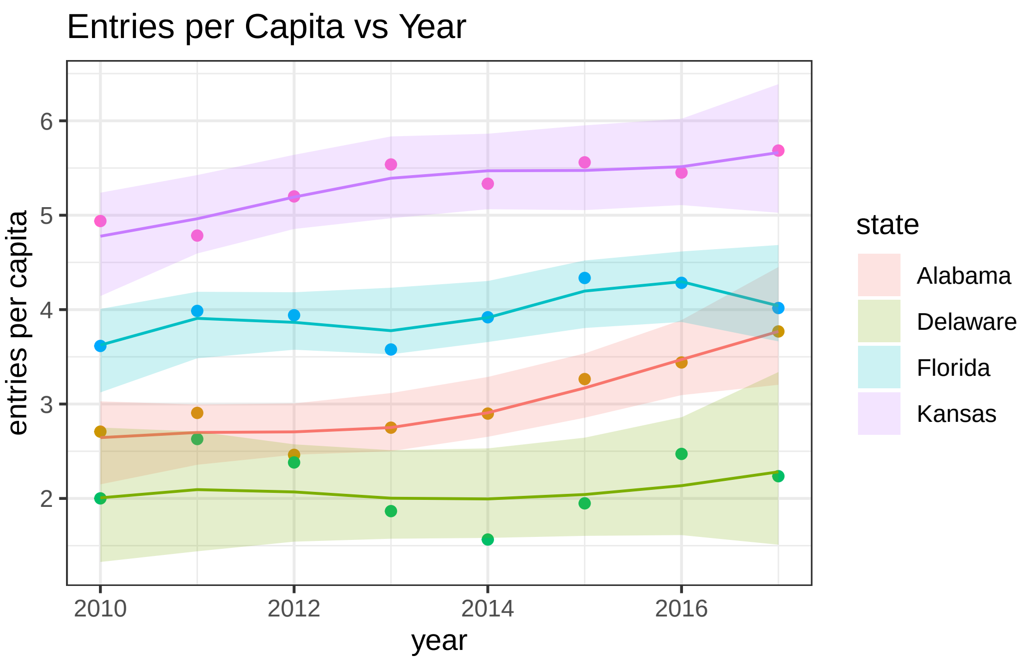
$$\begin{aligned} y_{st} &\sim N(\log \lambda_{st}, \sigma_{y,s}^2) \\ \log \lambda_{st} &= \alpha_k B_k(t) \\ \Delta^2 \alpha_k &\sim N(0, \sigma_{\alpha,s}^2) \\ \log \sigma_{\alpha,s} &\sim N(\mu_\sigma, \tau^2) \end{aligned}$$

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.

- Summary of the first few alphas:

```
                 mean     se_mean        sd       2.5%       25%       50%
alpha[1,1] 0.5557081 0.006332811 0.4050655 -0.1923000 0.3121625 0.5358444
alpha[1,2] 0.7654402 0.009898476 0.5583258 -0.2740913 0.4199565 0.7190606
alpha[1,3] 0.6754424 0.007816542 0.4907999 -0.2881593 0.3905222 0.6705946
                  75%    97.5%    n_eff      Rhat
alpha[1,1] 0.7772310 1.453646 4091.265 0.9998345
alpha[1,2] 1.0789606 1.972397 3181.550 1.0004573
alpha[1,3] 0.9531422 1.693589 3942.571 0.9998710
```

- Plot a few states:



Question 3

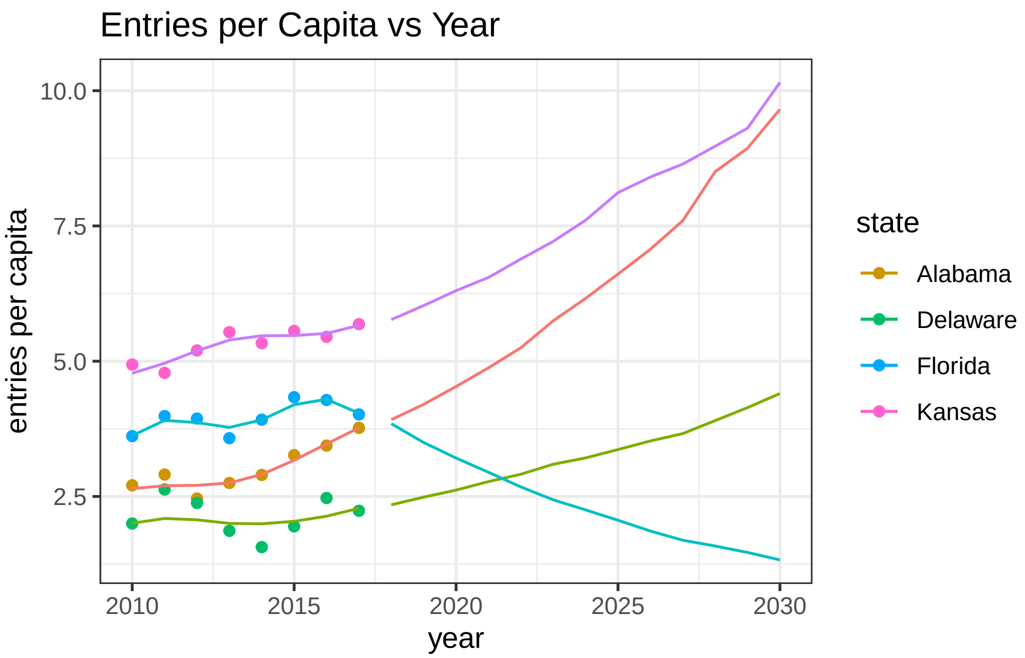
Project forward entries per capita to 2030. Pick 4 states and plot the results (with 95% CIs). Note the code to do this in R is in the lecture slides.

- Pick Alabama, Delaware, Florida and Kansas:

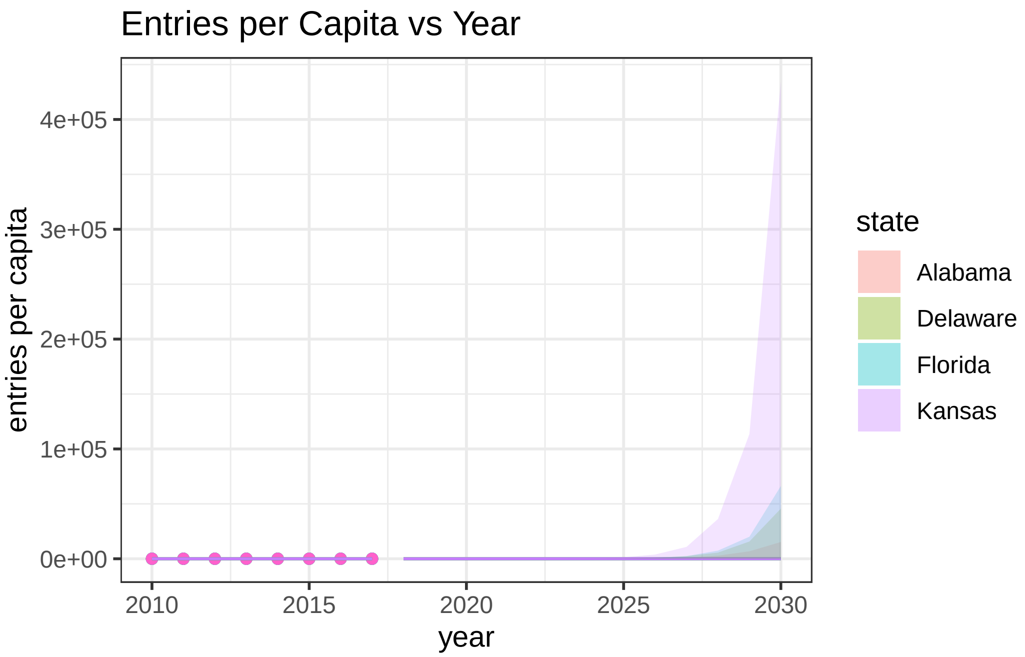
```
Index of  Alabama : 1
Index of  Delaware : 8
Index of  Florida : 10
Index of  Kansas : 17
```

```
# A tibble: 6 x 5
   year state   val   upr  lowr
  <int> <chr> <dbl> <dbl> <dbl>
1  2018 al     3.92  5.34 2.87
2  2018 de     2.35  4.62 1.14
3  2018 fl     3.85  5.03 3.18
4  2018 ka     5.77  7.22 4.56
5  2019 al     4.20  6.90 2.52
6  2019 de     2.49  6.43 0.975
```

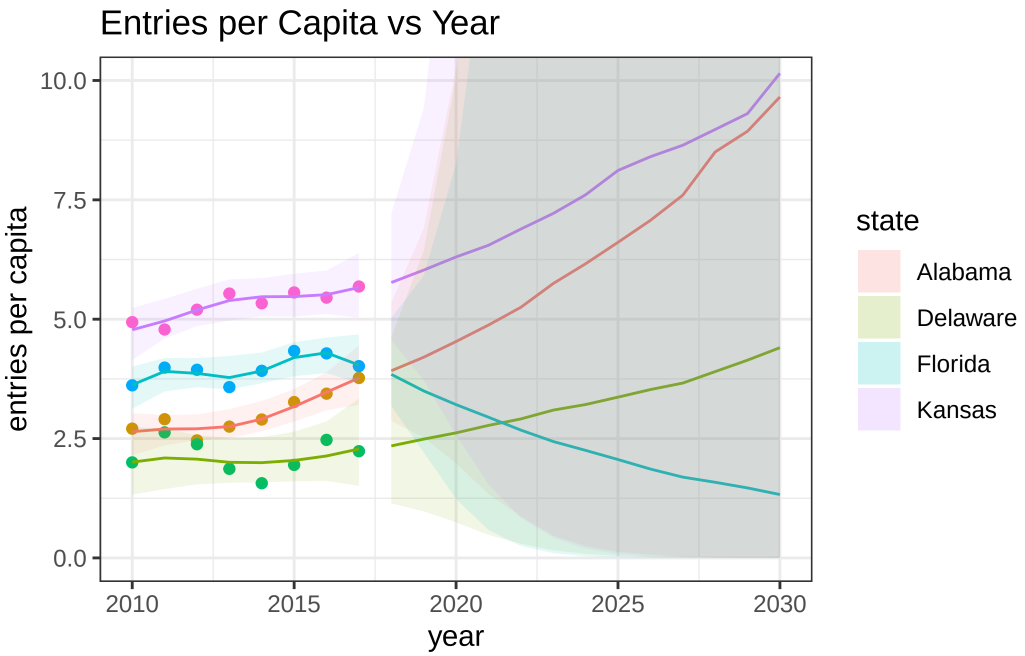
- Projection without CI:



- Projection with CI:



- Projection with CI zoomed in:



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?

- One potential issue is that the model may be over-fitting the data, the time-specific part might be just modeling the unnecessary noise left by the systematic part.
- Another potential issue is the model can have unidentifiable issue that can't tell the difference between the systematic part and the time-specific part.
- Constraint the sum of the parameters to be fixed may fix this.