STA2201H Methods of Applied Statistics II

Monica Alexander

Week 10: Temporal data

Roadmap

- A2 being graded
- Research proposal and A3 due Friday 29 March
- Feedback by Monday 1 April
- Presentations Wednesday 3 April
  - Short (5 min)
  - 3-4 slides: Introduction/Motivation, Data, Proposed model

Overview

Shifting our focus to thinking about models when we have time series of data

- Temporal models to estimate, smooth and project:
    - AR(1)
    - Random walks
    - Hierarchical smoothing

Reading for this week: Congdon (2006). Bayesian statistical modeling. Chapter 8

Measurement error

Measurement error

- We have talked about hierarchical models and how to fit them in Stan
- Often consider something like

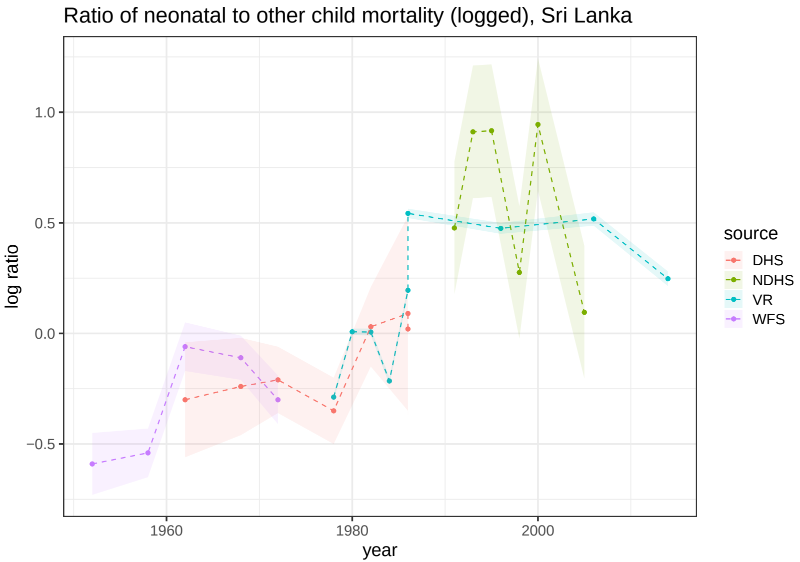
$$y_i | \mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$$

  with a model on μ_i. But what is σ^2?
- So far, have assumed it's constant for all observations i and estimated in the model
- But often have some info based on how the data were collected
- Can incorporate info about measurement error into our models

Motivating example

- Interested in estimating and projecting neonatal mortality in Sri Lanka over time
- Data available are from different sources which have differing degrees of error
- Best: vital registration systems (VR) but this hasn't always existed in Sri Lanka
- Also rely on survey data

Motivating example



Child mortality in Sri Lanka

Goals: estimate expected level of ratio over time

Issues:

- overlapping observations
- missing years
- different data sources
- different errors

Let's start off simple

For starters with the Sri Lankan data, let's model just a linear function over time

$$y_t \sim N(\mu_t, \sigma^2)$$

with

$$\mu_t = \alpha + \beta(t - t_c)$$

t_c is the mid-year of the study period.

But there's an issue with the indexes here!

Allowing for overlapping observations and missing data

A pretty straightforward extension:

$$y_i \sim N(\mu_{t[i]}, \sigma^2)$$

with

$$\mu_t = \alpha + \beta(t - t_c)$$

where $t[i]$ is the same indexing as in hierarchical case: the year t which observation i relates to.

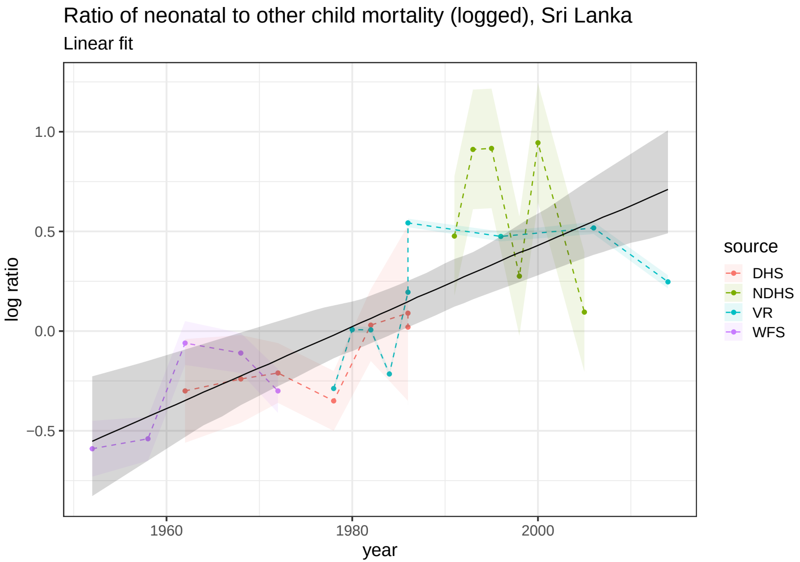
Fit in Stan

```
data {
  int<lower=0> N; //number of observations
  int<lower=0> T; //number of years
  int<lower=0> mid_year;
  vector[N] y; //log ratio
  vector[T] years; // vector of unique years
  int<lower=0> year_i[N]; // year index for observation i

}

parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;

}
```

Fit in Stan

```
transformed parameters{
  vector[T] mu;

  for(t in 1:T){
    mu[t] = alpha + beta*(years[t] - mid_year);
  }
}

model {
  vector[N] y_hat;

  y ~ normal(mu[year_i], sigma);

  alpha ~ normal(0, 1);
  beta ~ normal(0,1);
  sigma ~ normal(0, 1);
}
```

Results


Ratio of neonatal to other child mortality (logged), Sri Lanka
Linear fit
log ratio
1.0
0.5
0.0
−0.5
1960
1980
2000
year
source
DHS
NDHS
VR
WFS

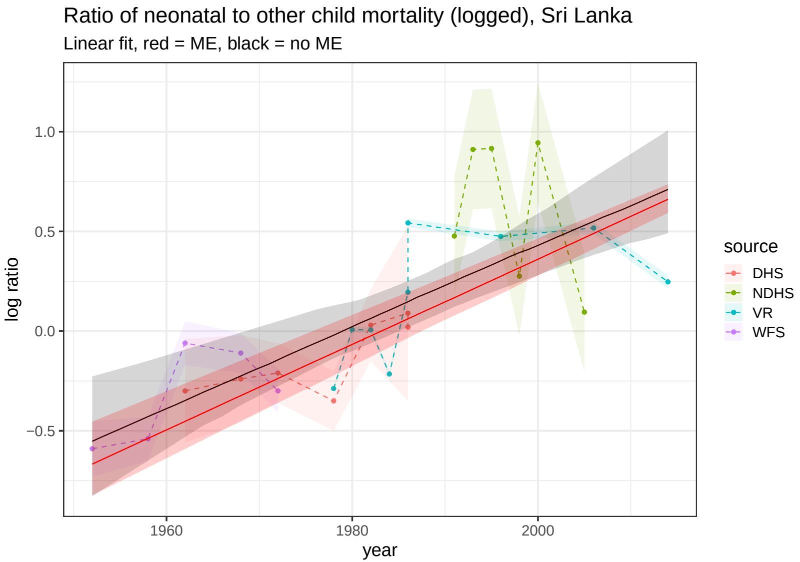
Incoporating measusrement error

Adding in the measurement error (standard errors based on sampling in this case) involves swapping out the estimated σ^2 with data:

$$y_i \sim N(\mu_{t[i]}, s_i^2)$$

```
model {

  y ~ normal(mu[year_i], se);

  alpha ~ normal(0, 1);
  beta ~ normal(0,1);
}
```

Result



Ratio of neonatal to other child mortality (logged), Sri Lanka
Linear fit, red = ME, black = no ME
log ratio
1.0
0.5
0.0
−0.5
1960
1980
2000
year
source
DHS
NDHS
VR
WFS

Take-aways

- Easy to account for missing data with right index set-up
- Accounting for measurement error useful when have data from different sources

This was a pretty simple linear model, can we do better?

Time series

Goals of time series modeling

We observe outcome of interest at particular time points t, y_t.

- y_t may have additional indexes e.g. y_{st} (e.g. deaths in state s year t)
- y_t may be related to covariates X_t
- y_t may have missing observations in the period

Some potential goals:

- forecasting
- back projecting
- reconstruction missing points
- smoothing

Goals of time series modeling

What you might be used to: Box-Jenkins approach.

Focus on the outcome:

- Start with y_t
- Remove anything systematic (trend, seasonality)
- Find an appropriate ARIMA specification
- Stationarity or death (differencing, transformations etc)

Perspective for this lecture

Strutural time series

Think about the outcome as:

$$y_t = \text{systematic part} + \text{fluctuations}$$

- The systematic part is potentially Trend + Seasonal Effects + Regression Term
- The errors/ fluctuations are likely to be autocorrelated because we're dealing with time
- We could model the systematic effects is by a set of fixed coefficients
- Or we could model them to vary over time, allowing for forecasts to place more weight on recent observations
- Intuitively: can model time dependency in outcome through time dependency in other parts
- We care less about stationarity, although still important for model specification and projections

Road map

- Simple AR(1) for y_t
  - how to run in Stan
  - how to forecast
- What if we have missing observations?
- What if the mean is non-zero?

Example: foster care populations

- Linear trend models
- Random walk models
- Hierarchical extensions

AR(1) process

A zero-mean autoregressive process y_t of order 1, referred to here as an $AR(1)$ process $y_t \sim AR(1)$ for $t = 0, \pm 1, \pm 2, \ldots$ is given by

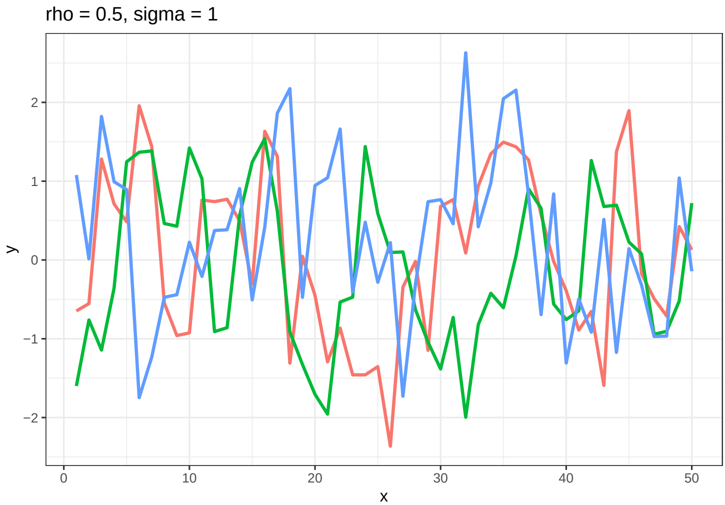
$$y_t = \rho y_{t-1} + \varepsilon_t$$
$$\varepsilon_t | \sigma \sim N\left(0, \sigma^2\right), \text{ independent}$$

- An $AR(1)$ process with normally distributed innovations ε_t and we assume that ε_t is indep. of y_{t-k} for $k > 0$

AR(1)

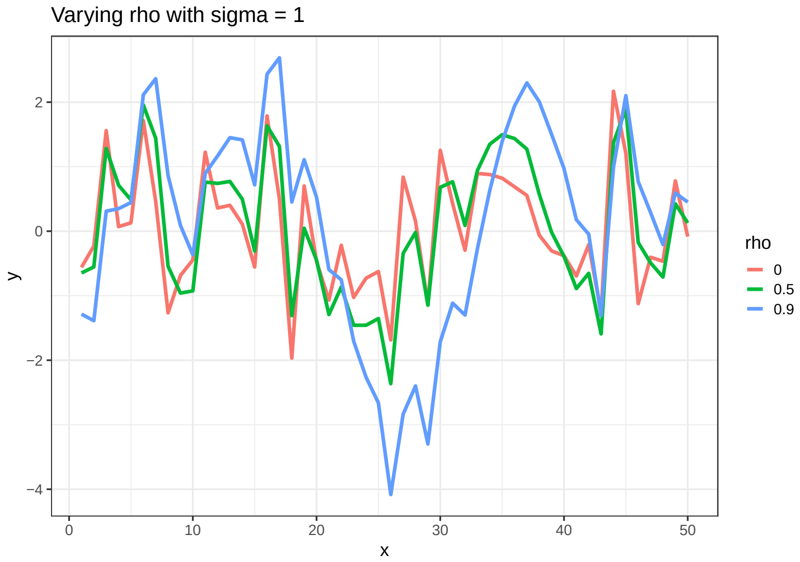
- An $AR(1)$ process is an example of a stochastic process: a sequence of random variables indexed by time.

Three different simulations:



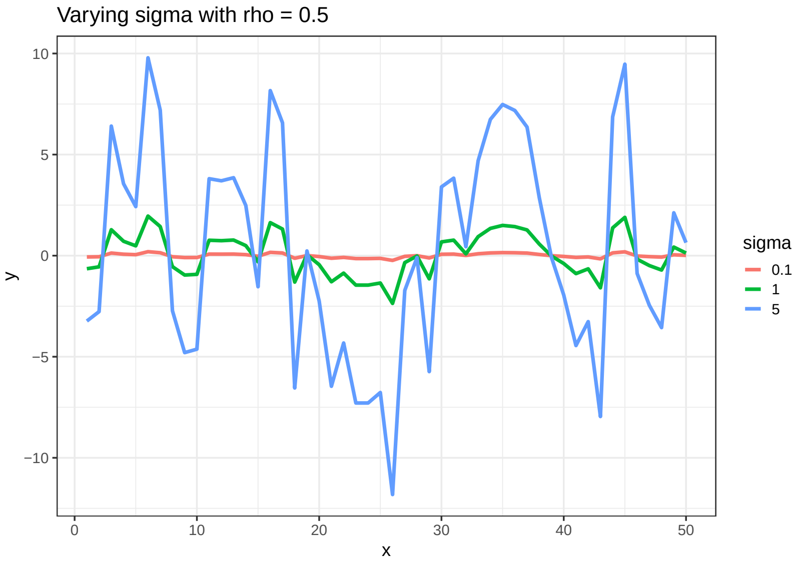
AR(1)

Interpretation of ρ?



AR(1)

Interpretation of σ?



AR(1)

$$y_t = \rho y_{t-1} + \varepsilon_t$$

$$\varepsilon_t | \sigma \sim N\left(0, \sigma^2\right), \text{ independent}$$

- For fixed ρ, σ controls magnitude of series
- ρ determines strength of autocorrelation

Stationarity for time series processes

A time series process is weakly (or second order) stationary if

- Mean $E(y_t)$ is constant with time t
- Covariance function $\gamma_{t,t+k} = \text{Cov}(y_t, y_{t+k})$ for any time t and time lag k depends on lag k only (is constant with time t).
- An AR(1) process is stationary if and only if $|\rho| < 1$

If the AR(1) is stationary then

$$\text{Var}(y_t) = \rho^2 \, \text{Var}(y_{t-1}) + \text{Var}(\varepsilon_t)$$

which implies stationary variance

$$\text{Var}(y_t) = \sigma^2 / \left(1 - \rho^2\right)$$

Stationarity

More general form, for $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$

$$\boldsymbol{y}|\rho, \sigma \sim N_n(\boldsymbol{0}, \Sigma)$$

with $\Sigma_{t,s} = \text{Cov}\left(y_t, y_s|\rho, \sigma\right) = \sigma^2 / \left(1 - \rho^2\right) \cdot \rho^{|t-s|}$.

Fit and forecast in a Bayesian setting

Suppose we have time series $y_1, \ldots, y_n$ and want to fit a Bayesian zero-mean AR(1) model to it, to construct forecasts.

Proposed model

$$y_t \sim AR(1)$$

$$\rho \sim U(-1, 1)$$

$$\sigma \sim N_+(0, 1)$$

How to fit in Stan? What's the likelihood of the y_i's?

How to fit in Stan?

We wrote that $\boldsymbol{y}|\rho, \sigma \sim N_n(\mathbf{0}, \Sigma)$, so could fit based on that. But this is slow! Generally good to avoid Multivariate normals is possible.

Faster option: decompose the likelihood function

$p(\boldsymbol{y}) = p\left(y_1\right) p\left(y_2|y_1\right) p\left(y_3|y_2, y_1\right) \cdot \ldots \cdot p\left(y_n|y_{n-1}, \ldots, y_1\right)$

where

$$
\begin{aligned}
y_t &= \rho y_{t-1} + \varepsilon_t \\
y_t|y_{t-1}\rho, \sigma &\sim N\left(\rho y_{t-1}, \sigma^2\right) \\
p\left(y_t|y_{t-1}, \ldots, y_1, \rho, \sigma\right) &= p\left(y_t|y_{t-1}, \rho, \sigma\right)
\end{aligned}
$$

Model block

```
model {

   y[2:N] ~ normal(rho * y[1:(N - 1)], sigma);

  // equivalent, but slower
  //for (n in 2:N)
    //  y[n] ~ normal(rho * y[n-1], sigma);
}
```

AR(1) in Stan

Fine, but what happened to y_1?

- Could just not model, condition on y_1, so leave out of data (what is done in Stan manual!)
- Loss of data in likelihood, hence less preferable (but ok if you are working with long time series).

Other option: use stationary distribution for y_1:

$$y_1 \sim N\left(0, \sigma^2 / \left(1 - \rho^2\right)\right)$$

Model

```
data {
  int<lower=0> N;
  int<lower=0> P;
  vector[N] y;
}
parameters {
  real<lower = -1, upper = 1> rho;
  real<lower=0> sigma;
}
model {
   //likelihood
   y[1] ~ normal(0, sigma/sqrt((1-rho^2)));
   y[2:N] ~ normal(rho * y[1:(N - 1)], sigma);

   //priors
   rho ~ uniform(-1, 1);
   sigma ~ normal(0,1);
}
```

Fitting to simulated data with $\rho = 0.5$ and $\sigma = 0.1$

```r
y <- GetAR(nyears = 100, rho = 0.5, sigma = 0.1) #from distortr package
N <- length(y)

mod1 <- stan(data = list(y = y, N = N),
             file = "ar1_1.stan", iter = 100)
```

```
##         mean   Rhat
## rho   0.4949 1.0023
## sigma 0.0945 0.9915
```

How to get projections?

- Given and posterior sample $\rho^{(s)}$ and $\sigma^{(s)}$ one can forecast trajectory $y_{n+p}^{(s)}$ with $p \geq 1$ as

$$y_{n+p}^{(s)}|y_{n+p-1}^{(s)}, \rho^{(s)}, \sigma^{(s)} \sim N\left(\rho^{(s)} y_{n+p-1}^{(s)}, \left(\sigma^{(s)}\right)^2\right)$$

where $y_n^{(s)} = y_n$. Once we have set of posterior samples, $y_{n+p}^{(1)}, y_{n+p}^{(2)}, \ldots, y_{n+p}^{(S)}$ point forecasts and 95% CIs can be constructed.

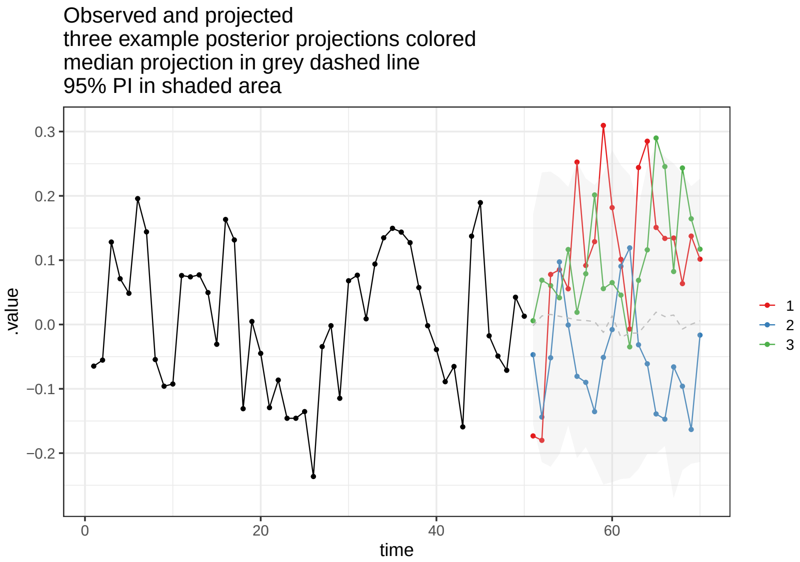
- Can do in R or in Stan
- Note: can also back-project in the same way

Projections in Stan using the generated quantities block

The full model

```
data {
  int<lower=0> N;
  int<lower=0> P;
  vector[N] y;
}
parameters {
  real<lower = -1, upper = 1> rho;
  real<lower=0> sigma;
}
model {
   //likelihood
   y[1] ~ normal(0, sigma/sqrt((1-rho^2)));
   y[2:N] ~ normal(rho * y[1:(N - 1)], sigma);

   //priors
   rho ~ uniform(-1, 1);
   sigma ~ normal(0,1);
}
generated quantities {
  //project forward P years
  vector[P] y_p;
  y_p[1] = normal_rng(rho*y[N], sigma);
  for( i in 2:P){
    y_p[i] = normal_rng(rho*y_p[i-1], sigma);
  }
}
```

Results



Observed and projected
three example posterior projections colored
median projection in grey dashed line
95% PI in shaded area
0.3
0.2
0.1
0.0
−0.1
−0.2
.value
0
20
40
60
time
1
2
3

Missing data

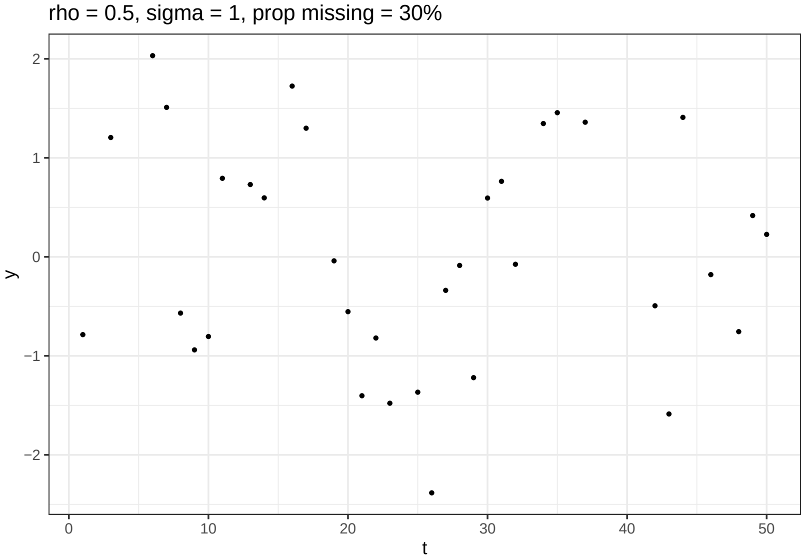
Missing data

- Now imagine we have observations y_t but some t's are missing
- e.g. if we observe $y_1, y_2, \ldots, y_n$ from time points $t_1, t_2, \ldots, t_n$ with $t_i \neq t$.
- As above, keep process model the same but change the data model
- Need to create an indexing vector $t[i]$ which tells you what t the ith observation refers to
- Just like `year_i` in the Sri Lanka example

Missing data Stan model

```
data {
  int<lower=0> N;
  int<lower=0> N_obs;
  vector[N_obs] y;
  int t_i[N_obs];
}
parameters {
  real<lower = -1, upper = 1> rho;
  vector[N] mu;
  real<lower=0> sigma;
  real<lower=0> sigma_y;
}
model {

   y ~ normal(mu[t_i], sigma_y);
   mu[1] ~ normal(0, sigma/sqrt((1-rho^2)));
   mu[2:N] ~ normal(rho * mu[1:(N - 1)], sigma);

   //priors
   rho ~ uniform(-1,1);
   sigma ~ normal(0,1);
   sigma_y ~ normal(0,1);
}
```

Missing data: simulation



Results

Non-zero means

Non-zero means

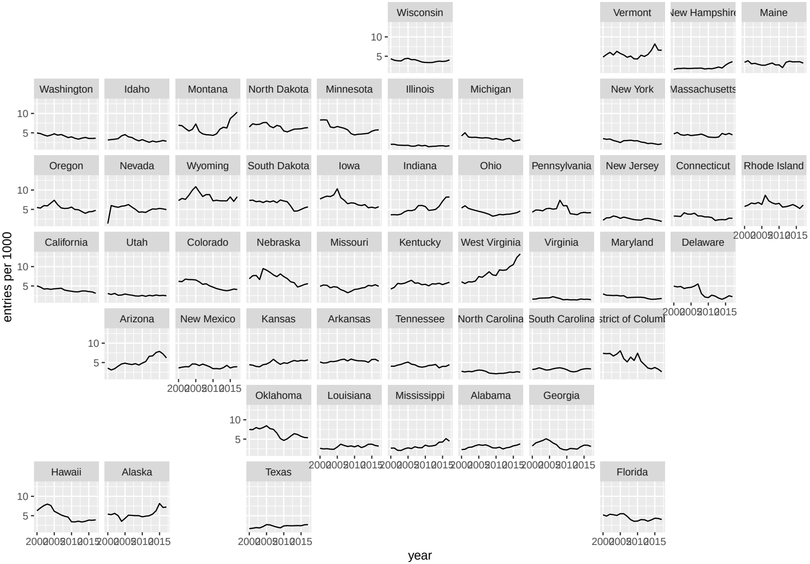
Suppose we have the following candidate model for y_t

$$y_t|\gamma_t, \delta \sim N\left(\gamma_t, \delta^2\right)$$
$$\gamma_t = \kappa_t + \mu_t, \text{ with } \mu_t \sim AR(1)$$

- μ_t is zero-mean AR(1) model
- κ_t could be
  - a constant α
  - related to covariate e.g. $\kappa_t = x_t\beta$
  - ...
- Fit as before but add in mean term
- Easy in theory, in practice model specification often hard

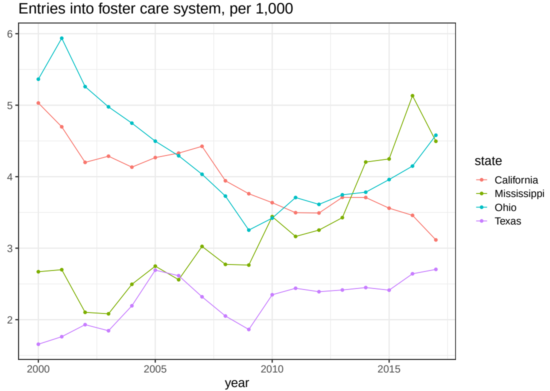
Example from this point

Goal: Project foster care populations by state in the US



Projecting foster care populations

- There's a number of different outcomes of interest, but let's look at entries into system (children aged 0-17)
- Let's use population of children as exposure variable, alternatively, think of modeling entries per captia
- Ignore issues of population age structure for now



Foster care populations

- Goal is projection, but understanding is important
  - why are things going up or down?
  - Are there driving factors that are modifiable or can be planned for?
- Uncertainty around projections is important

How to approach problem?

Data model

- y_{st} is number of entries into foster care system in state s and year t
- P_{st} is child population in same state and year

$$y_{st} \sim \text{Poisson}(\lambda_{st} P_{st})$$

λ_{st} is rate of entries, the outcome of interest. Model for λ_{st}?

Model for λ_{st}?

Start with no covariates (apart from time!)

Possibilities:

- Simplest would be

$$\log \lambda_{st} = \alpha + \beta t + \varepsilon_{st}$$

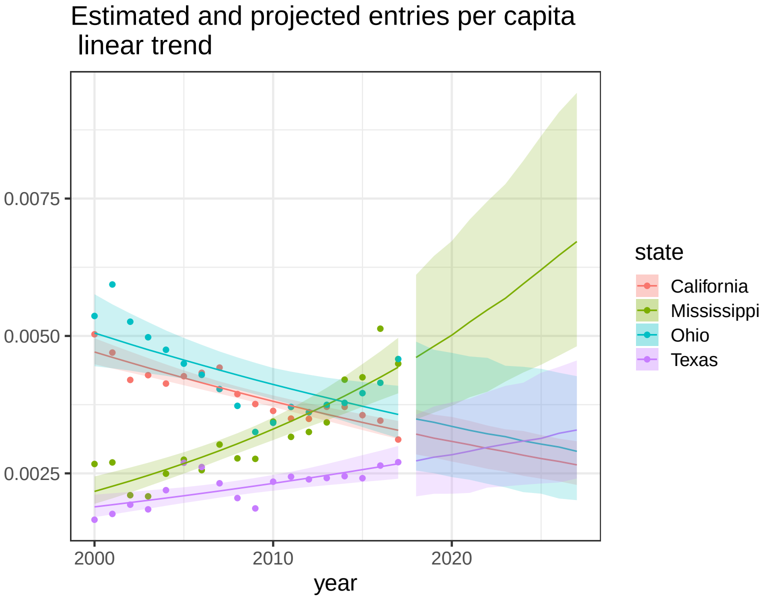
with $\varepsilon_{st} \sim N(0, \sigma^2)$

- What about autocorrelated errors

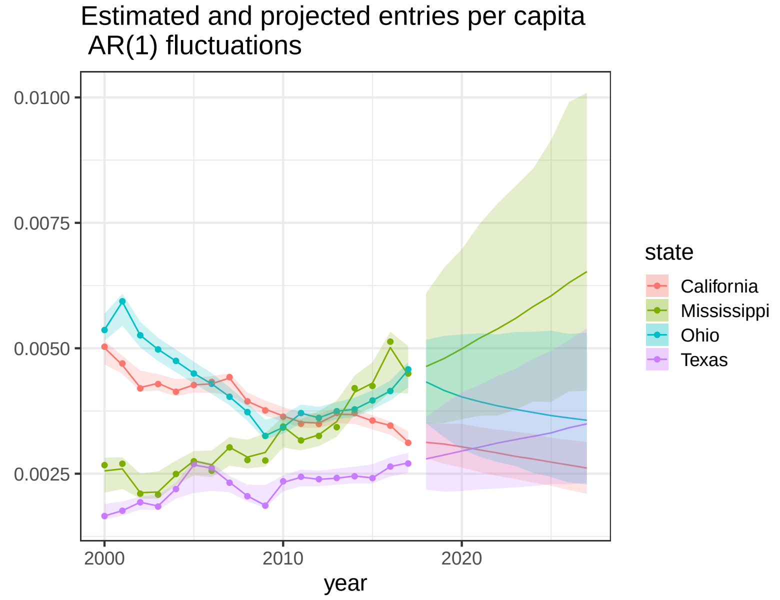
$$\log \lambda_{st} = \alpha + \beta t + \varepsilon_{st}$$

with $\varepsilon_{st} \sim AR(1)$

Linear in time



Linear in time with AR(1) errors



Moving away from non-linear trends

- Linear trend + AR(1) wasn't terrible, but probably want to put more weight on more recent observations
- Simplest option here is a random walk:

$$\log \lambda_{st} = \alpha_{st}$$

with $\alpha_{st} \sim N(\alpha_{s,t-1}, \sigma_s^2)$ or equivalently $\Delta\alpha_{st} \sim N(0, \sigma_s^2)$.

Random walk

Now we've lost stationary. The α's have the form

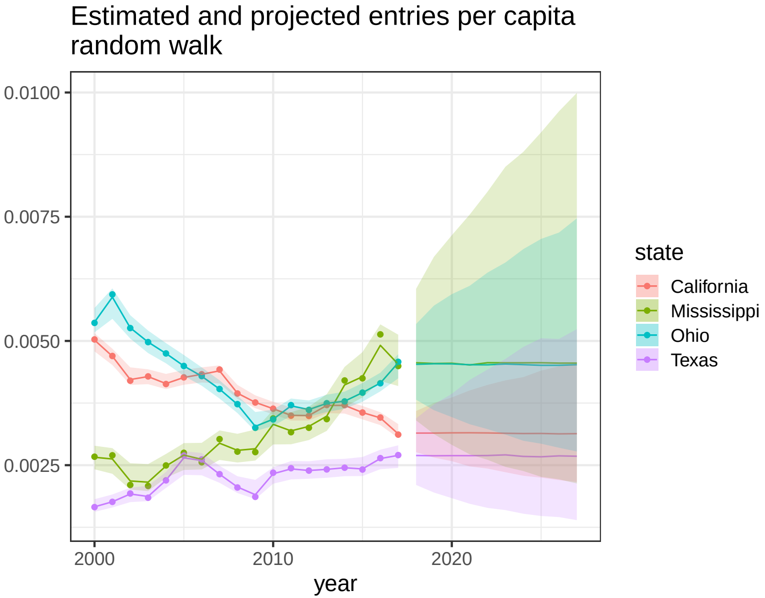
$$\alpha_t = \alpha_{t-1} + \varepsilon_t$$
$$\varepsilon_t | \sigma \sim N\left(0, \sigma^2\right)$$

Suppose $\alpha_1 = 0$, and that σ is known, then for $t > 0$, then

$$E\left(\alpha_t\right) = E\left(\alpha_{t-1}\right) + E\left(\varepsilon_t\right) = 0 \text{ and}$$
$$\mathrm{Var}\left(\alpha_t\right) = \mathrm{Var}\left(\alpha_{t-1}\right) + \mathrm{Var}\left(\varepsilon_t\right) = (t-1)\sigma^2$$

In practice what does this mean for our projections?

Random walk



Estimated and projected entries per capita
random walk
0.0100
0.0075
0.0050
0.0025
2000
2010
2020
year
state
California
Mississippi
Ohio
Texas

Random walk

- We've gone from our projections to caring about all years to just caring about the last year
- Projections in RW are based on the last observed level
- Uncertainty increases forever with time (c.f. stationary AR(1))

Higher-order random walks

We can increase the random walk's memory by moving to higher order random walks. E.g. a second-order random walk is

$$\log \lambda_{st} = \alpha_{st}$$

with

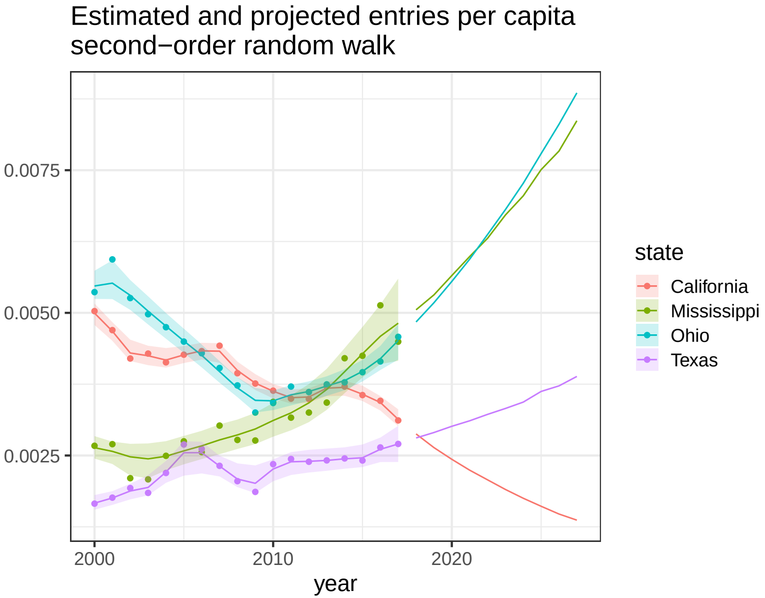
$\alpha_{st} - \alpha_{s,t-1} \sim N(\alpha_{s,t-1} - \alpha_{s,t-2}, \sigma_s^2)$ or equivalently

$\alpha_{st} \sim N(2\alpha_{s,t-1} - \alpha_{s,t-2}, \sigma_s^2)$ or equivalently

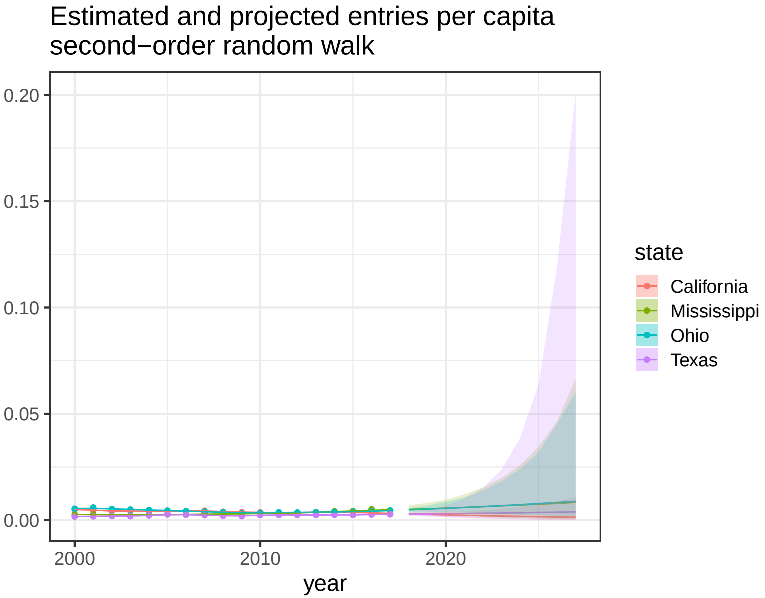
$\Delta^2 \alpha_{st} \sim N(0, \sigma_s^2)$.

If a first-order RW projects the level, what does a second-order RW project?

Second order RW



Oh no


Estimated and projected entries per capita
second−order random walk
0.20
0.15
0.10
0.05
0.00
2000
2010
2020
year
state
California
Mississippi
Ohio
Texas

Moving forward: hierarchical model

- Second order random walk gives 'reasonable' point estimates but unrealistic and unusable uncertainty intervals
- But we are working with hierarchical data: states within regions within the US
- Currently we are fitting a separate time series to each state
- Could model hierarchically such that information about the variability in the random walks (i.e. the σ^2 term) could be shared across states

Hierachical model for σ_s^2

A plausible set-up:

$$\log \lambda_{st} = \alpha_{st}$$

with

$\alpha_{st} \sim N(2\alpha_{s,t-1} - \alpha_{s,t-2}, \sigma_s^2)$

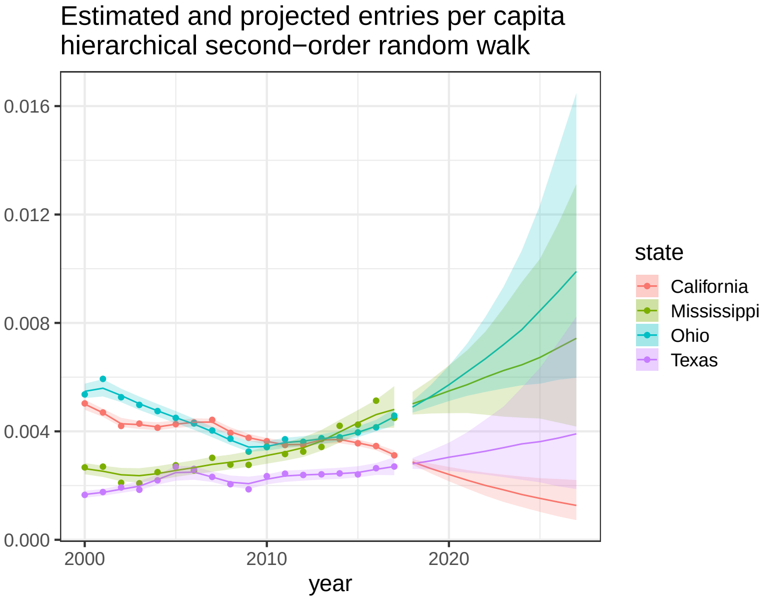
and

$\log \sigma_s \sim N(\mu_\sigma, \tau^2)$

with the usual prior on $\tau \sim N_+(0, 1)$.

- model the log of the σ's to ensure positive
- Make sure you can see that this is hierarchical. For reference, the non-hierarchical model just has $\sigma_s \sim N_+(0, 1)$

Looking better



Foster care: summary

- Second order RW shows promise in picking up characteristics of time series
- But of little use for understanding **why** changes are happening, and whether they are likely to happen in future

What I ended up doing: Bayesian hierarchical state-space model

- a whole suite of candidate covariates
- association between child welfare outcomes and covariates is allowed to vary by geography and over time (in a smooth way)
  - i.e. we can put a time series model on the regression coefficients!
- covariates chosen through consultation with domain knowledge experts and shrinkage priors

Post-script: Bayesian state-space (dynamic linear) models

The linear Gaussian state-space model, also called a dynamic linear model, assumes Normal errors and can be written in a general form as

$$y_t = F_t x_t + v_t, \quad v_t \sim N(0, V_t)$$
$$x_t = G_t x_{t-1} + w_t, \quad w_t \sim N(0, W_t)$$

- State-space models describe how a particular process or state x_t evolves over time, and how those states relate to data we observe, y_t.
- Developed in the context of modeling underlying physical processes (where we are interested in x_t), but useful in understanding changes in observed outcomes, too, in a regression framework

State-space (dynamic linear) models

A simple dynamic linear regression would have the form

$$\begin{aligned}
y_t &= \boldsymbol{X}'_t \boldsymbol{\beta}_t + \epsilon_t \\
\boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \\
\epsilon_t &\sim N\left(0, \sigma^2_\epsilon\right) \\
\boldsymbol{\eta}_t &\sim N\left(\mathbf{0}, \Sigma_\eta\right)
\end{aligned}$$

- The first line here is our usual linear regression set-up, with the only difference being the regression coefficients $\boldsymbol{\beta}_t$ vary over time.
- Different way of estimating these models, but we can go full Bayes and use MCMC

Foster care model

$$\log y_{s,t} \sim N(\mu_{s,t}, s_y^2)$$

$$\mu_{s,t} = \alpha_s + \mathbf{X_{s,t}}'\beta_{r,t} + \delta_{s,t}$$

$$\alpha_s \sim N(\mu_\alpha[r], \sigma_\alpha^2[r])$$

$$\beta_{r,t} \sim N(2 \cdot \beta_{r,t-1} - \beta_{r,t-2}, \sigma_\beta^2)$$

$$\delta_{s,t} \sim N(\rho_s \delta_{s,t-1}, \sigma_\delta^2)$$

Summary

Hierarchical take-aways:

- Up until today we have been putting hierarchical structures on regression coefficients (slopes, intercepts)
- Can also put hierarchical model on variance terms!
- Interpretation: the variability in a series in a particular state tells us something about the variability in another state
- Has the effect of shrinking the variance towards a global mean

Model checking?

- In general, you can't use LOO-CV to compare time series models in the same way we have been doing, because of the time dependence in the data

Possibilities:

- As usual: residual plots, where residual = observation - estimate
- Out-of-sample validation, e.g. leaving out data at random (if reconstruction of missing values is of interest) or the most recent observations (if forecasting is of interest).
- In-sample validation (depending on context): construct 1 (or more)-step ahead forecasts and compare observation to that forecast.
- There is a 'future' version of LOO discussed here: https://mc-stan.org/loo/articles/loo2-lfo.html