

# STA2201H Methods of Applied Statistics II

Monica Alexander

Week 10: Temporal data

# Roadmap

- ▶ A2 being graded
- ▶ Research proposal and A3 due Friday 31 March
- ▶ Feedback by Monday 3 April
- ▶ Presentations Wednesday 5 April
  - ▶ Short (5 min)
  - ▶ 3-4 slides: Introduction/Motivation, Data, Proposal model
- ▶ Research project write up due at earliest 19 April

# 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

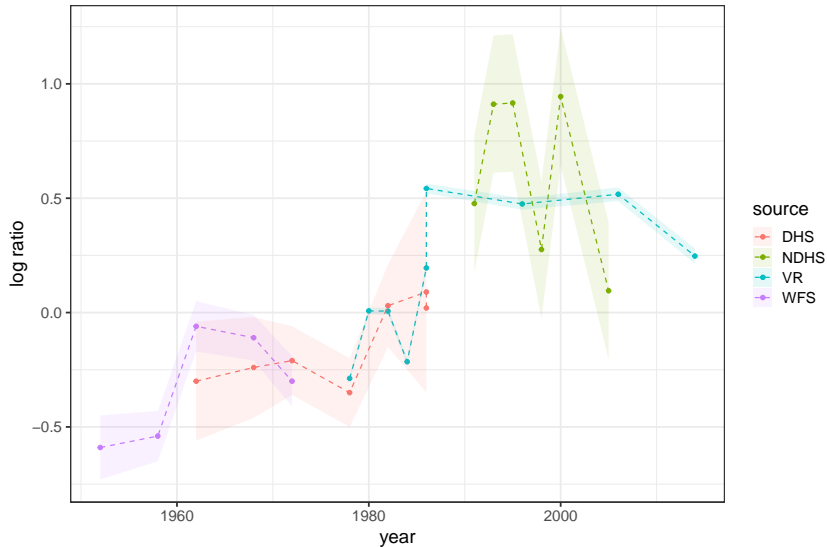
Reading for this week: Congdon (2006). Bayesian statistical modeling. Chapter 8

## Last week

- ▶ 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

Ratio of neonatal to other child mortality (logged), Sri Lanka



# 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;  
  
}
```

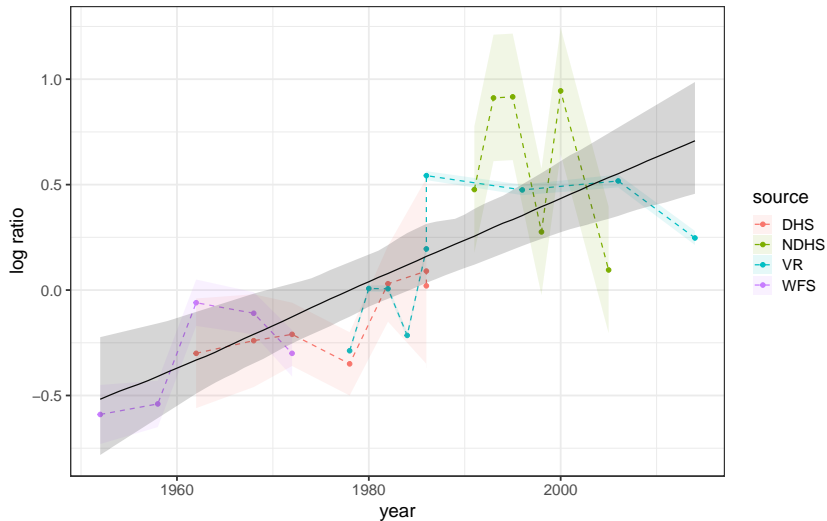
## Fit in Stan

```
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



## Incorporating measurement error

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

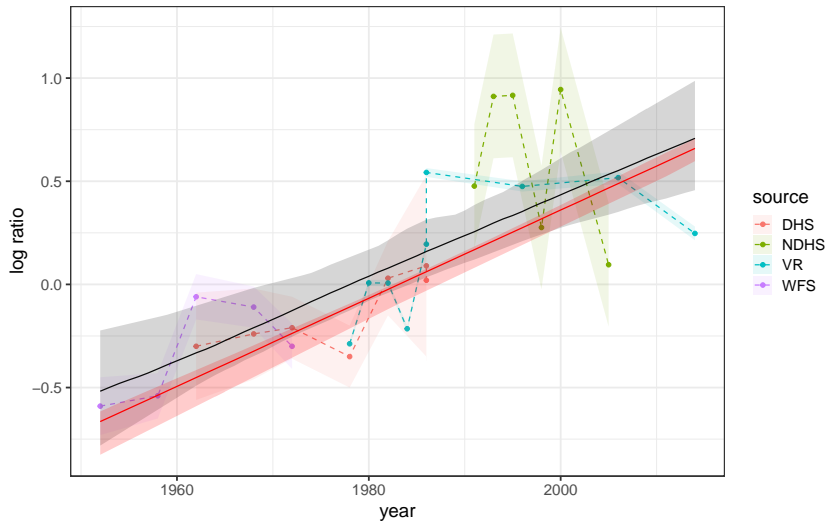
$$y_i \sim N(\mu_{t[i]}, s_i^2)$$

```
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



## 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

## Structural 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  $t = 0, \pm 1, \pm 2, \dots$  is given by

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

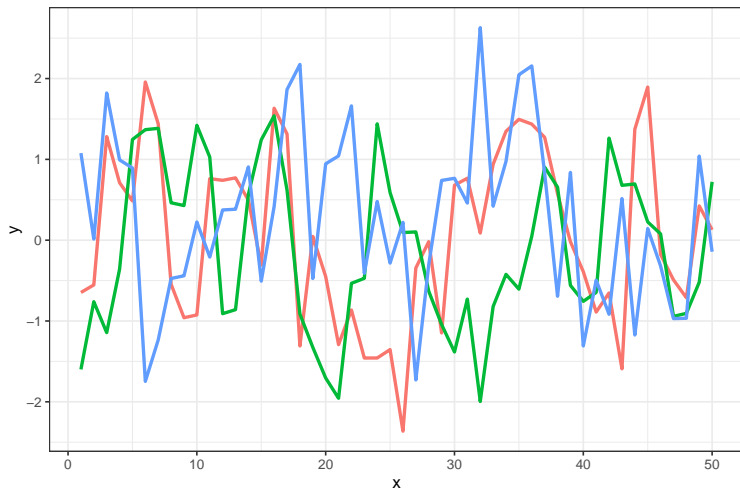
- An  $AR(1)$  process with normally distributed innovations  $\varepsilon_t$  and we assume that  $\varepsilon_t$  is indep. of  $y_{t-k}$  for  $k > 0$

# AR(1)

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

Three different simulations:

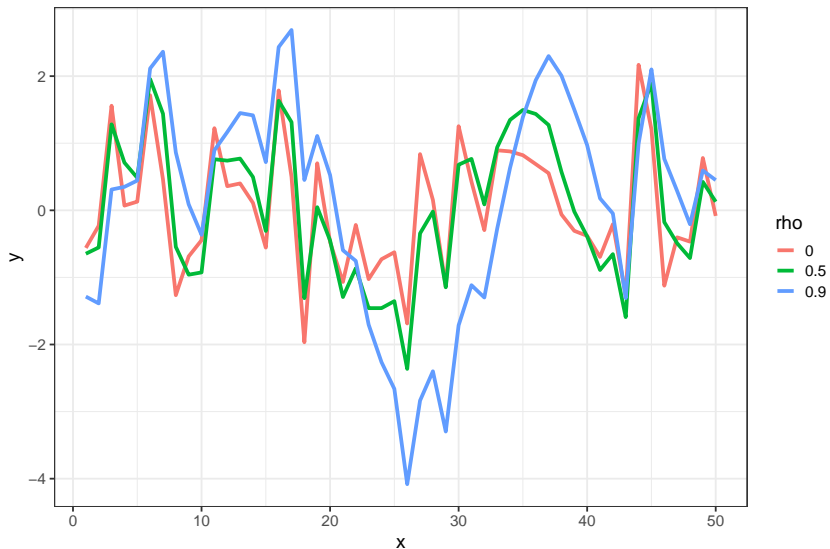
$\rho = 0.5$ ,  $\sigma = 1$



# AR(1)

Interpretation of  $\rho$ ?

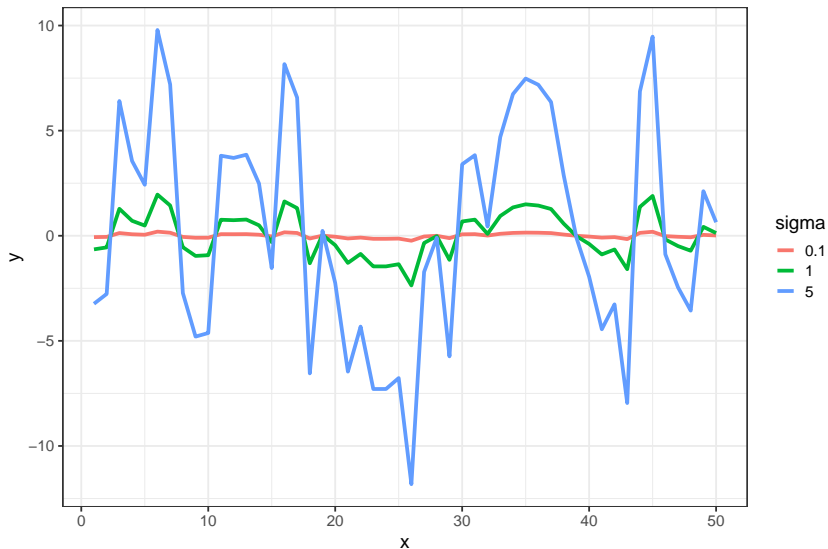
Varying  $\rho$  with sigma = 1



# AR(1)

Interpretation of  $\sigma$ ?

Varying sigma with rho = 0.5



# AR(1)

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

- ▶ For fixed  $\rho$ ,  $\sigma$  controls magnitude of series
- ▶  $\rho$  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

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

# Stationarity

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

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

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

## Fit and forecast in a Bayesian setting

Suppose we have time series  $y_1, \dots, y_n$  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  $\mathbf{y}|\rho, \sigma \sim N_n(\mathbf{0}, \Sigma)$ , so could fit based on that. But this is slow! Generally good to avoid Multivariate normals is possible.

Faster option: decompose the likelihood function

$$p(\mathbf{y}) = p(y_1) p(y_2|y_1) p(y_3|y_2, y_1) \cdot \dots \cdot p(y_n|y_{n-1}, \dots, y_1)$$

where

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

$$y_t|y_{t-1}\rho, \sigma \sim N(\rho y_{t-1}, \sigma^2)$$

$$p(y_t|y_{t-1}, \dots, y_1, \rho, \sigma) = p(y_t|y_{t-1}, \rho, \sigma)$$

## 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  $y_1$ :

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

# 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 $\sigma = 0.1$

```
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      0.4792 1.0127  
## sigma    0.0938 1.0097
```



## 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,  $y_{n+p}^{(1)}, y_{n+p}^{(2)}, \dots, y_{n+p}^{(S)}$  point forecasts and 95% CIs can be constructed.

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

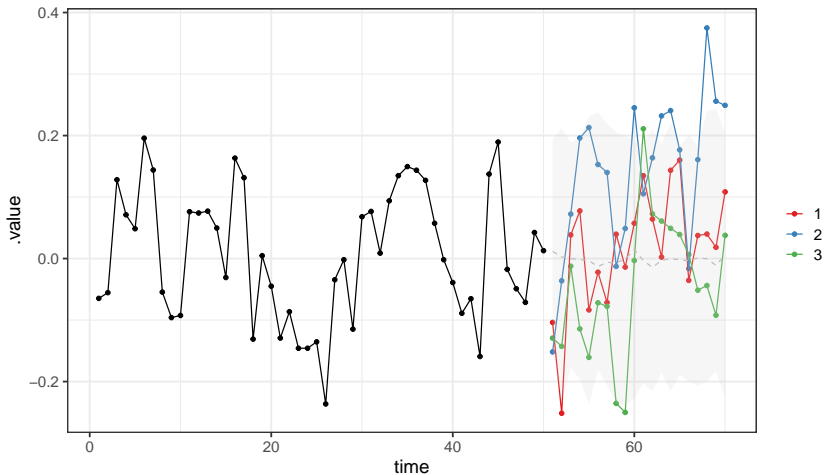
# Projections in Stan using the generated quantities block

## 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



Missing data

## Missing data

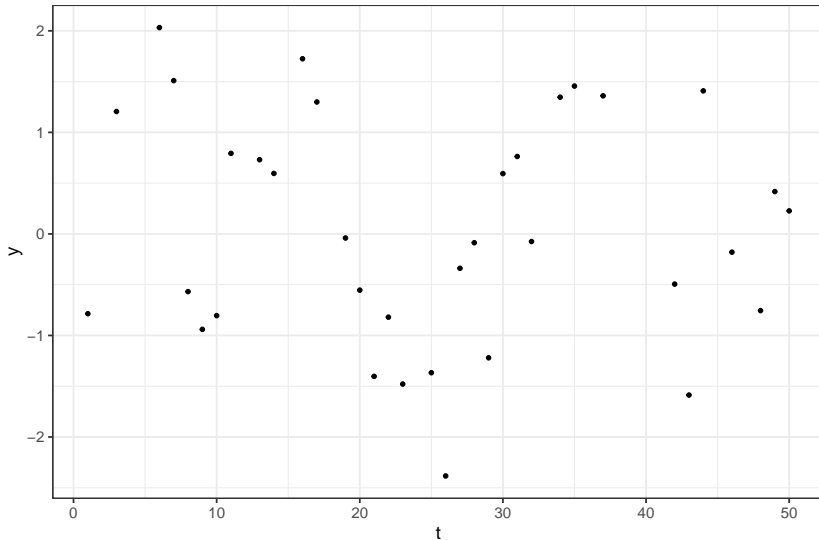
- ▶ Now imagine we have observations  $y_t$  but some  $t$ 's are missing
- ▶ e.g. if we observe  $y_1, y_2, \dots, y_n$  from time points  $t_1, t_2, \dots, t_n$  with  $t_i \neq t$ .
- ▶ As above, keep process model the same but change the data model
- ▶ Need to create an indexing vector  $t[i]$  which tells you what  $t$  the  $i$ th 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);  
}
```

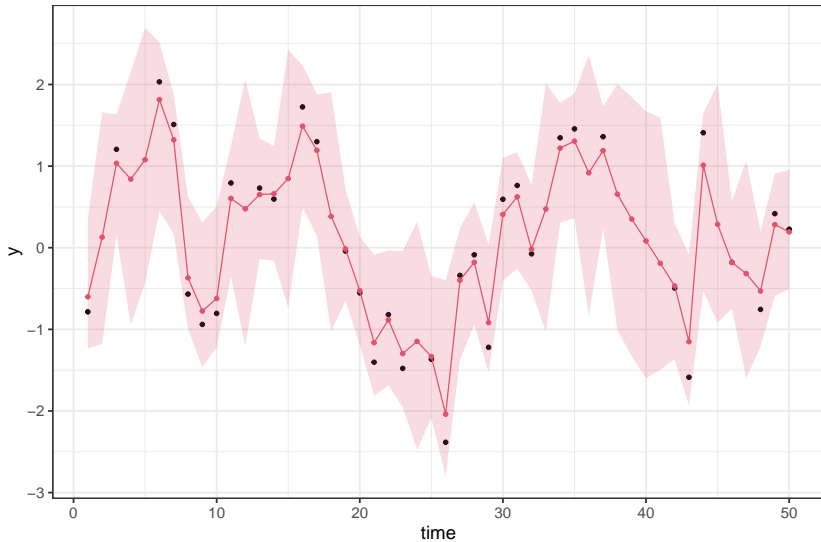
# Missing data: simulation

$\rho = 0.5$ ,  $\sigma = 1$ , prop missing = 30%



# Results

$\rho = 0.5$ ,  $\sigma = 1$ , 30% of data missing





## Missing data

Suppose we want to get  $\mu_t$  given  $\mu_{t-1}$  and  $\mu_s$ , where  $s > t$ . What is the conditional mean and variance of  $\mu_t$ ?

It turns out that

$$E(\mu_t | \mu_{t-1}, \mu_s) = \frac{1}{1-A} \left( \rho \cdot (1 - \rho^{2(s-t)}) \cdot \mu_{t-1} + \rho^{s-t} (1 - \rho^2) \cdot \mu_s \right)$$

$$\text{Var}(\mu_t | \mu_{t-1}, \mu_s) = \frac{\sigma^2}{1 - \rho^2} \left( 1 - \frac{\rho^2 - 2A + \rho^{2(s-t)}}{1 - A} \right)$$

$$A = \rho^{2(s-t+1)}$$

- ▶ Conditional mean is weighted average of two points, where weights depend on how far away  $s$  is
- ▶ Variance increases with  $s$

Non-zero means

## Non-zero means

Suppose we have the following candidate model for  $y_t$

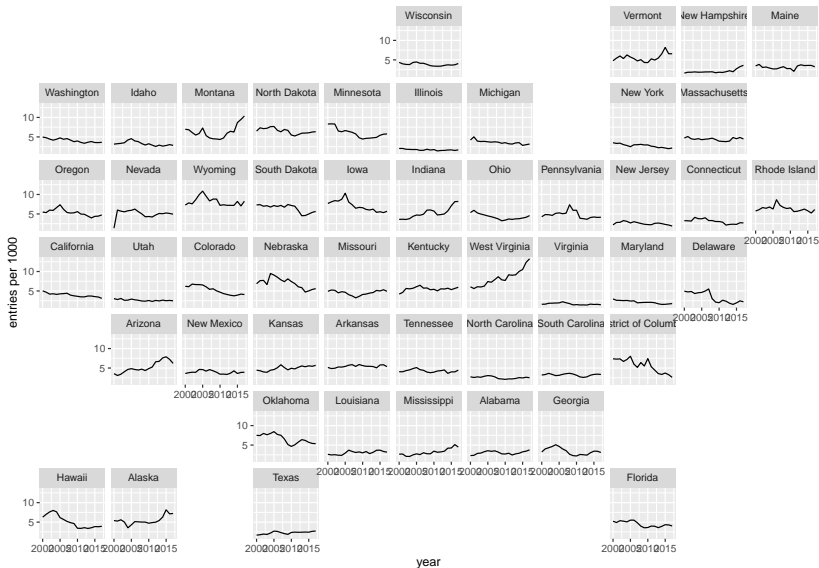
$$y_t | \gamma_t, \delta \sim N(\gamma_t, \delta^2)$$

$$\gamma_t = \kappa_t + \mu_t, \text{ with } \mu_t \sim AR(1)$$

- ▶  $\mu_t$  is zero-mean AR(1) model
- ▶  $\kappa_t$  could be
  - ▶ a constant  $\alpha$
  - ▶ 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

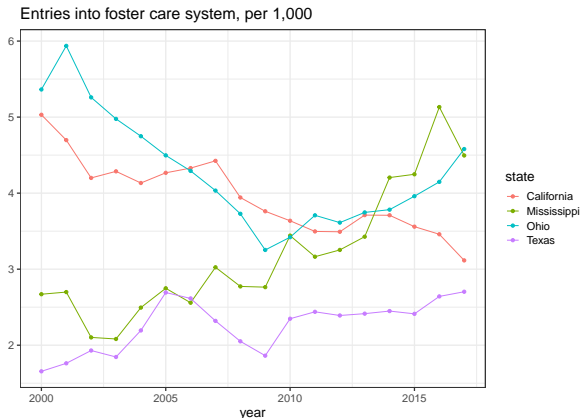
# Example from this point

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 capita
- ▶ 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

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

$\lambda_{st}$  is rate of entries, the outcome of interest. Model for  $\lambda_{st}$ ?

## Model for $\lambda_{st}$ ?

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

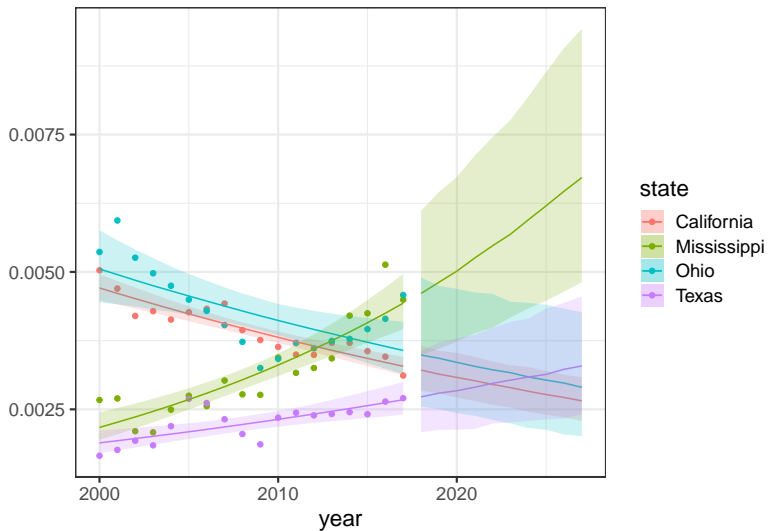
$$\log \lambda_{st} = \alpha + \beta t + \varepsilon_{st}$$

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



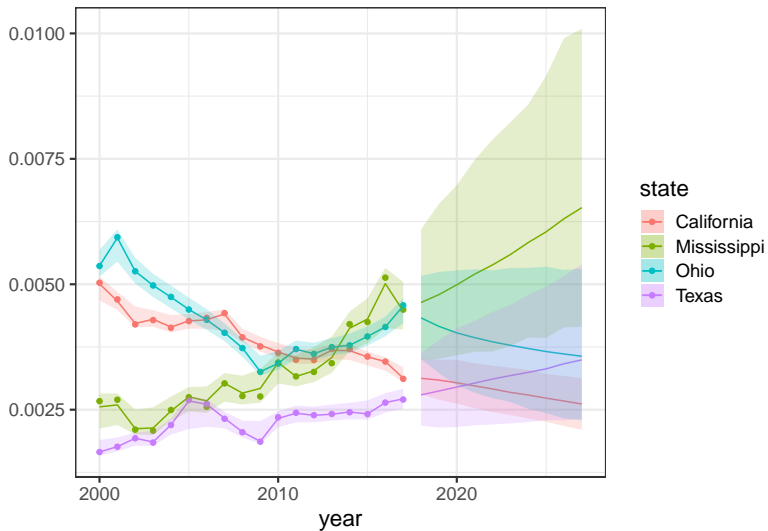
# Linear in time

Estimated and projected entries per capita  
linear trend



# Linear in time with AR(1) errors

Estimated and projected entries per capita  
AR(1) fluctuations



## 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  $\alpha$ 's have the form

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

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

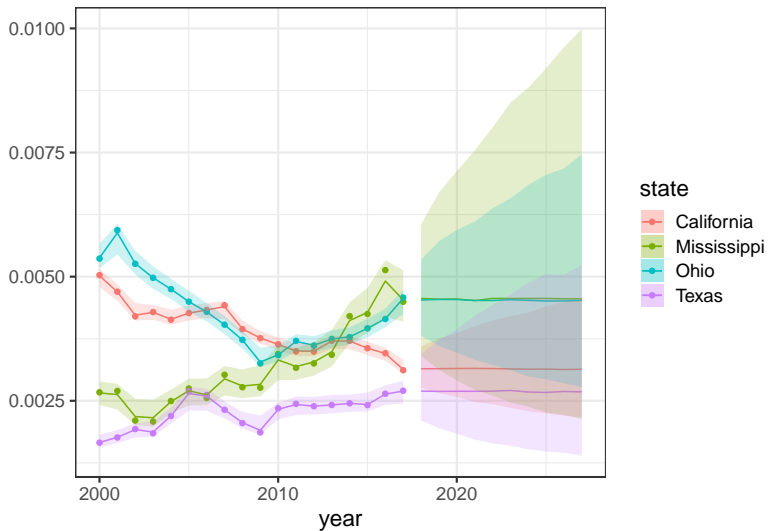
$$E(\alpha_t) = E(\alpha_{t-1}) + E(\varepsilon_t) = 0 \text{ and}$$

$$\text{Var}(\alpha_t) = \text{Var}(\alpha_{t-1}) + \text{Var}(\varepsilon_t) = (t-1)\sigma^2$$

In practice what does this mean for our projections?

# Random walk

Estimated and projected entries per capita random walk



## 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

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

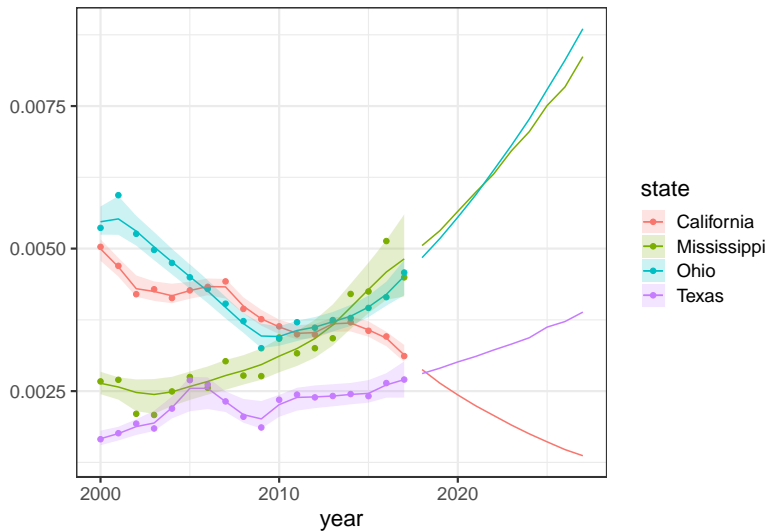
$$\alpha_{st} \sim N(2\alpha_{s,t-1} - \alpha_{s,t-2}, \sigma_s^2) \text{ or equivalently}$$

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

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

## Second order RW

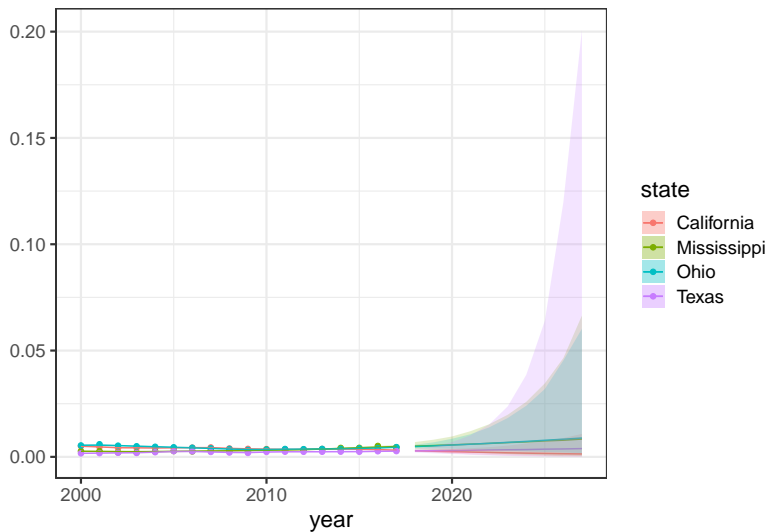
Estimated and projected entries per capita  
second-order random walk





Oh no

Estimated and projected entries per capita  
second-order random walk



## 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  $\sigma^2$  term) could be shared across states

## Hierarchical model for $\sigma_s^2$

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

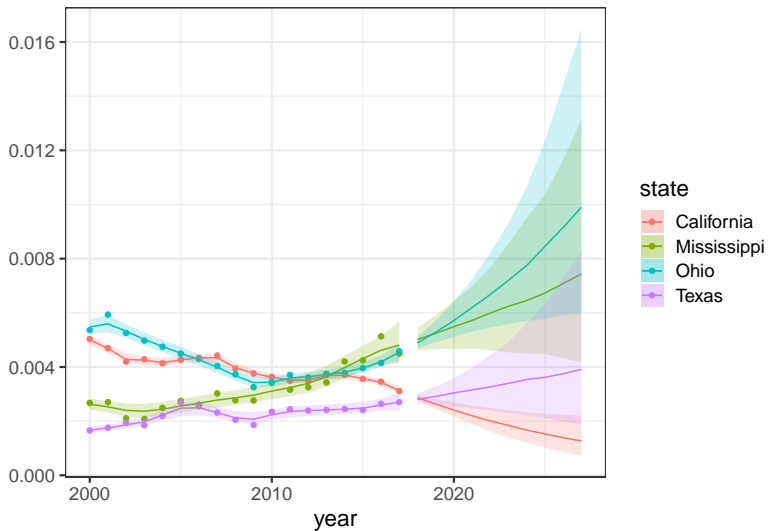
$$\log \sigma_s \sim N(\mu_\sigma, \tau^2)$$

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

- ▶ model the log of the  $\sigma$ '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

Estimated and projected entries per capita  
hierarchical second-order random walk



## 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

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

- ▶ 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

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

- ▶ 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

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

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

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

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

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



# 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>