Shinystan demo

- install.packages(“shinystan”)

Space and Time in Regression Models

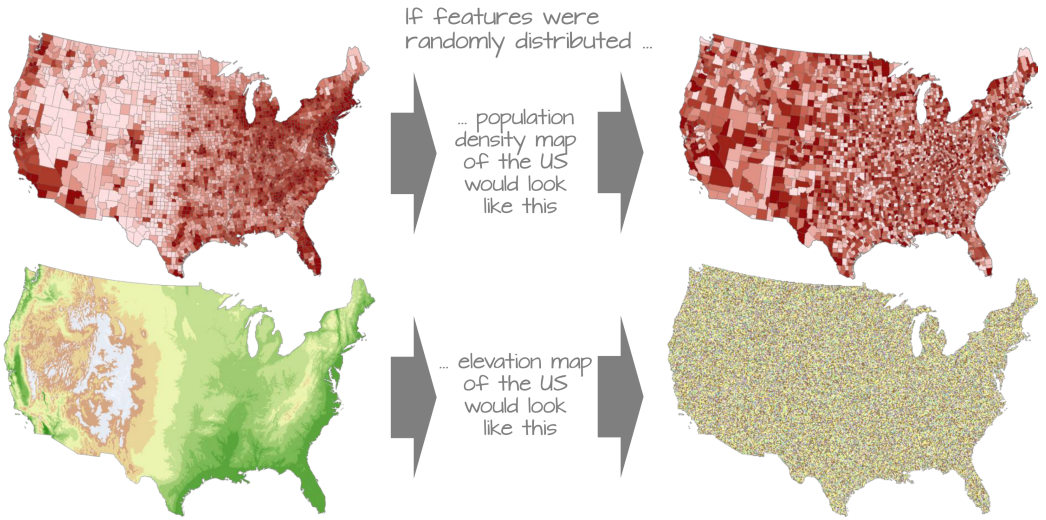
Lecture 28: EEB 622; April 28, 2020

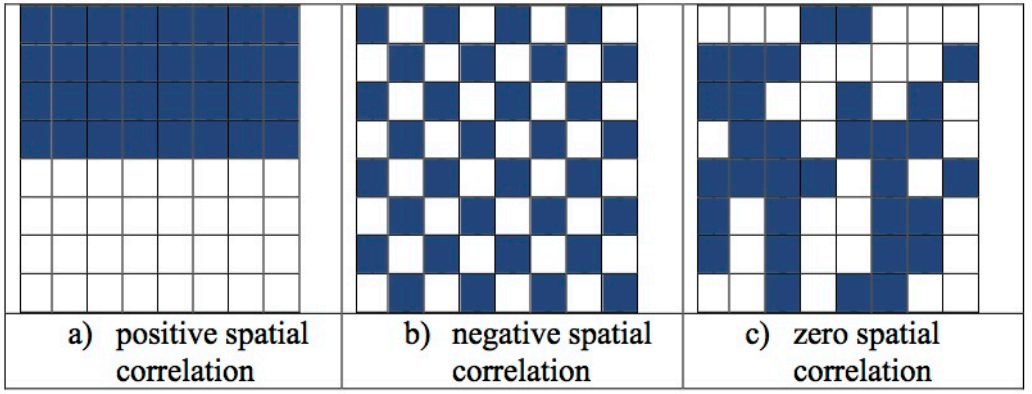
Learning objectives

- What is autocorrelation and why is it an issue in analyses?
- How can we account for spatial autocorrelation?
- How can we model repeated measures or data with temporal structure?

Autocorrelation

- A measure of similarity (correlation) between "nearby" observations
- **Spatial autocorrelation:**
  - Sites that are closer together tend to be more similar than sites that are far apart.
  - E.g. Are biodiversity patterns driven by elevation?
- **Temporal autocorrelation:**
  - Observations that are temporally proximate are more likely to be similar.
  - E.g. Repeated measures of a single subject
    - How do truck drivers respond to sleep deprivation over a 10 day period?
- **Evolutionary autocorrelation**:
  - Species with a more recent common ancestor, tend to be similar in many traits
- Autocorrelation can be negative too!
  - E.g. Tree diversity in the tropics?



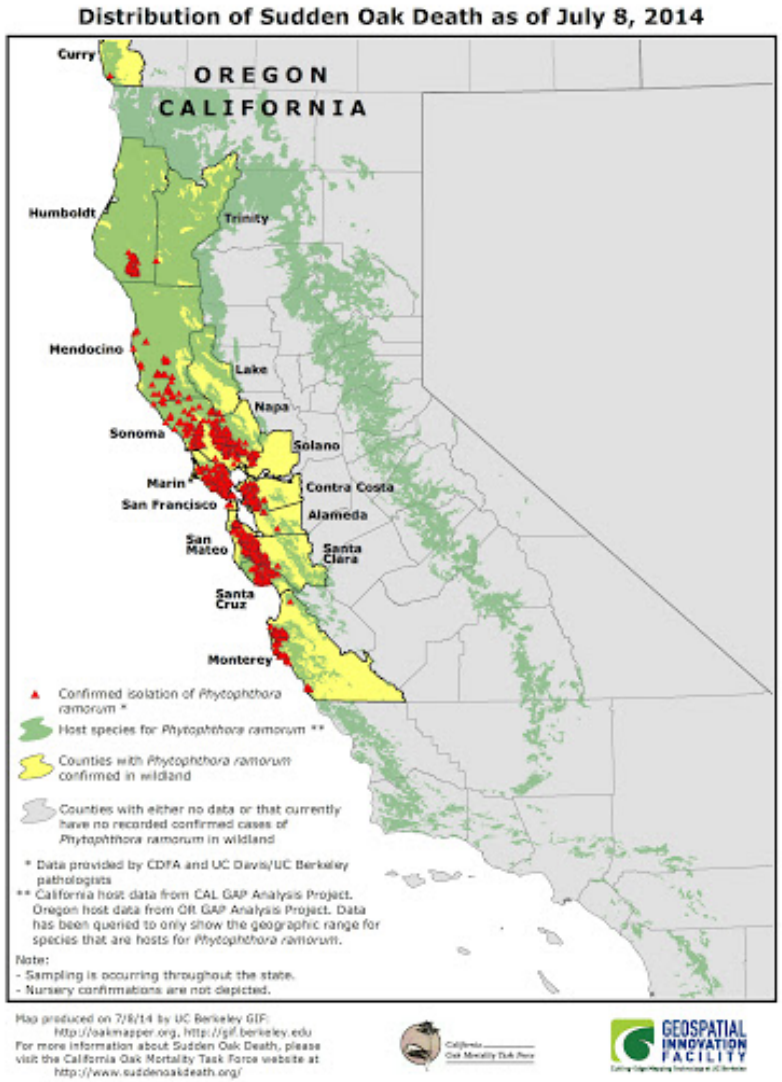


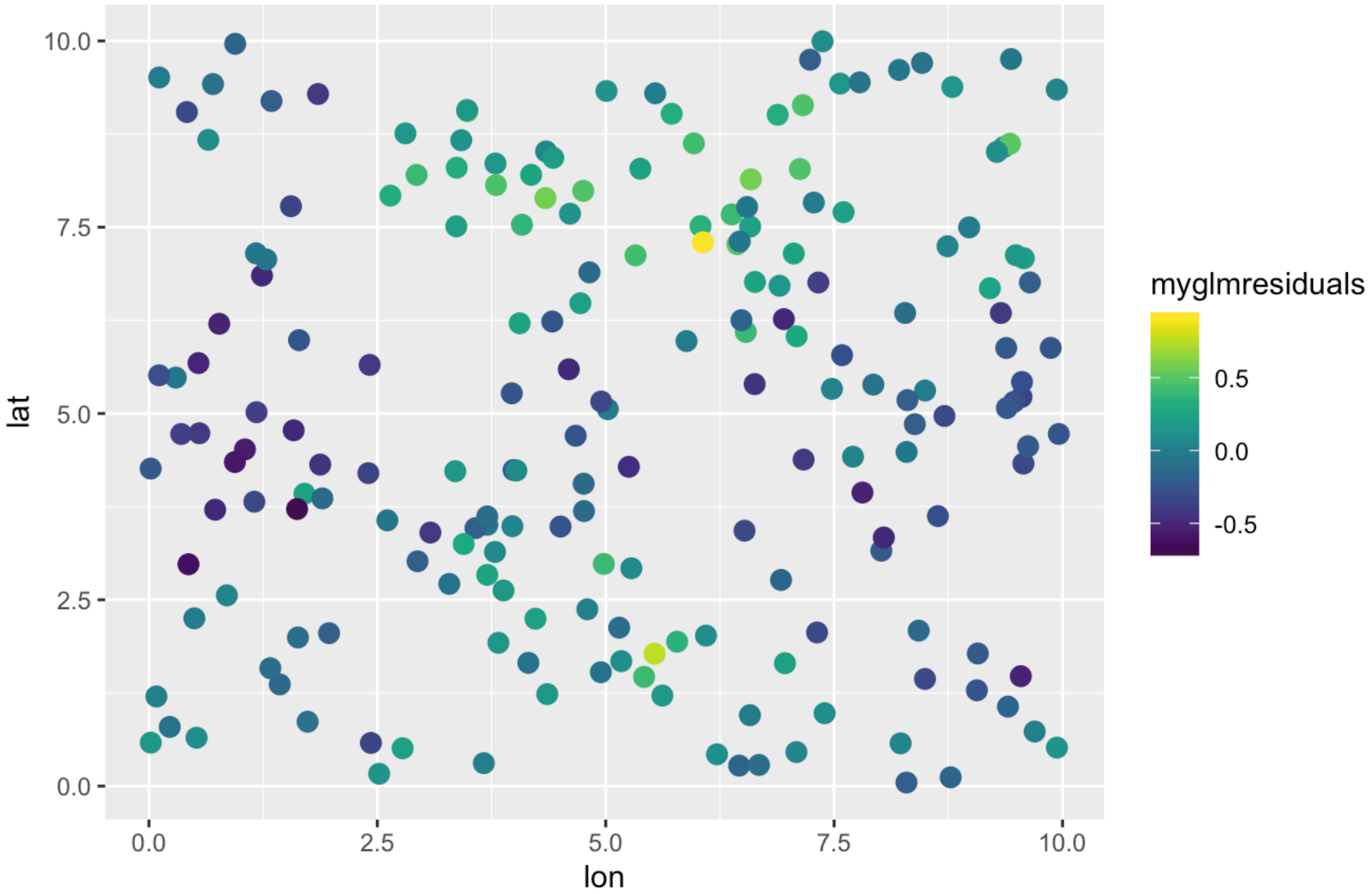
Why is autocorrelation an issue?

- Independent variable actually confounded with other factors.
- Violates the assumption that the residuals in regression are independent
- Accounting for AC can reveal patterns that otherwise would be hidden.

Spatial autocorrelation in regression

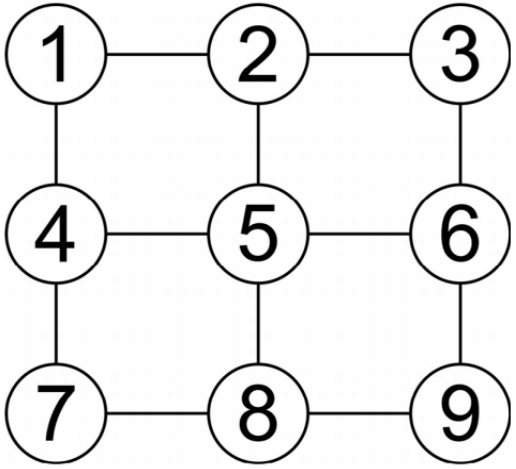
- Examples in:
  - Epidemiology:
    - Which factors drive susceptibility to a disease?
  - Invasion biology/species distribution modeling:
    - Which factors influence suitable habitat for this species?
      - Occurrence is mixture of both dispersal AND survival.
  - Remote sensing:
    - Data from nearby polygons are more similar than those farther apart.
    - Similar habitat conditions, biological processes such as migration or dispersal, and human impacts or management interventions




10.0
7.5
5.0
2.5
0.0
lat
0.0
2.5
5.0
7.5
10.0
lon
myglmresiduals
0.5
0.0
-0.5

Spatial autoregressive models

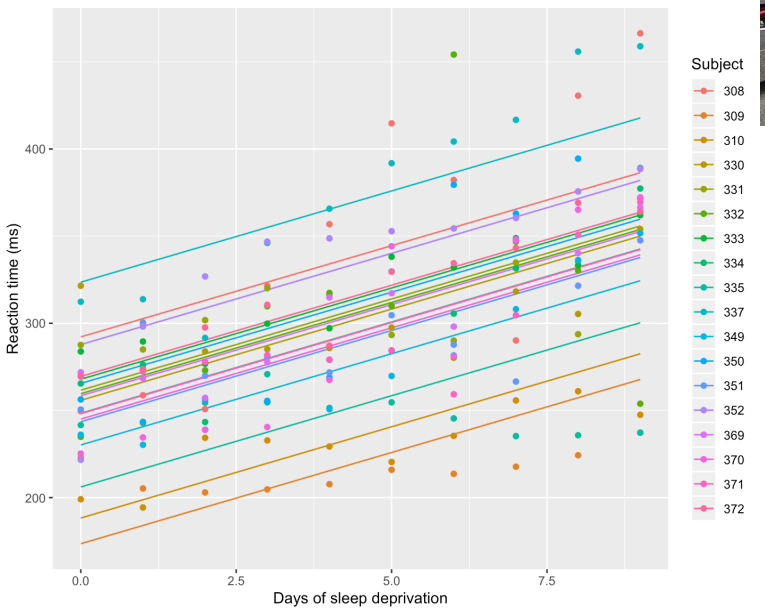
- Instead of assuming independence, account for spatial autocorrelation by modeling a covariance matrix of the residuals as a function of the locations where the response variable was collected.
  - Covariance decays exponentially with distance.
  - Keeps inference on regression parameters from being invalidated by residual autocorrelation
- AR models: a network of connections between samples is specified.
- This spatial dependence/correlation matrix is included in a hierarchical model
- Most common spatial autoregressive models:
  - Conditional autoregressive (CAR)
  - Simultaneous autoregressive (SAR)



$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

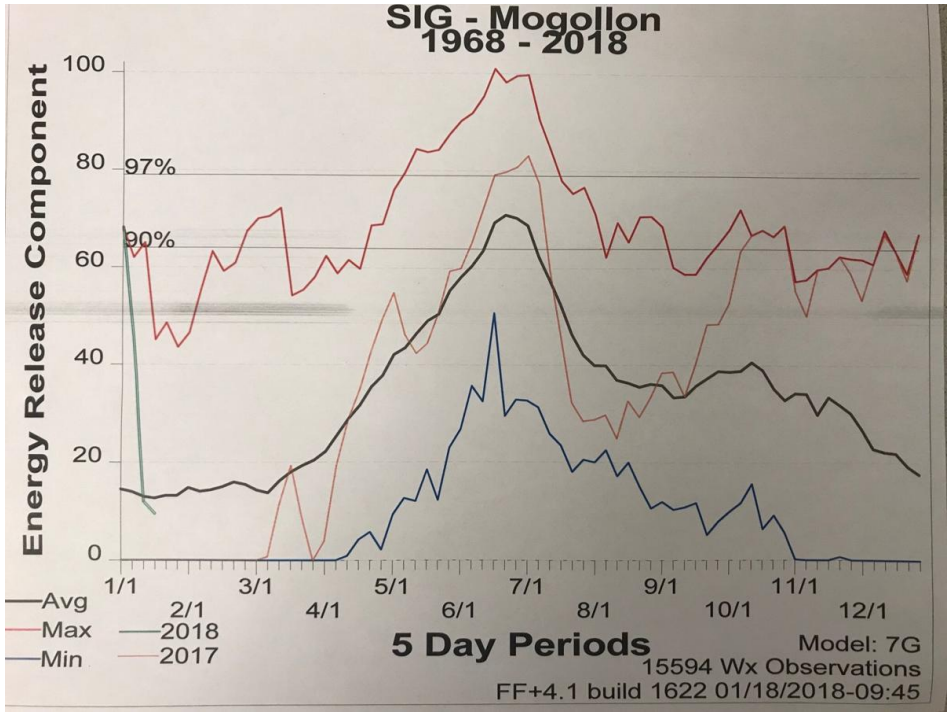
Temporal autocorrelation in regression

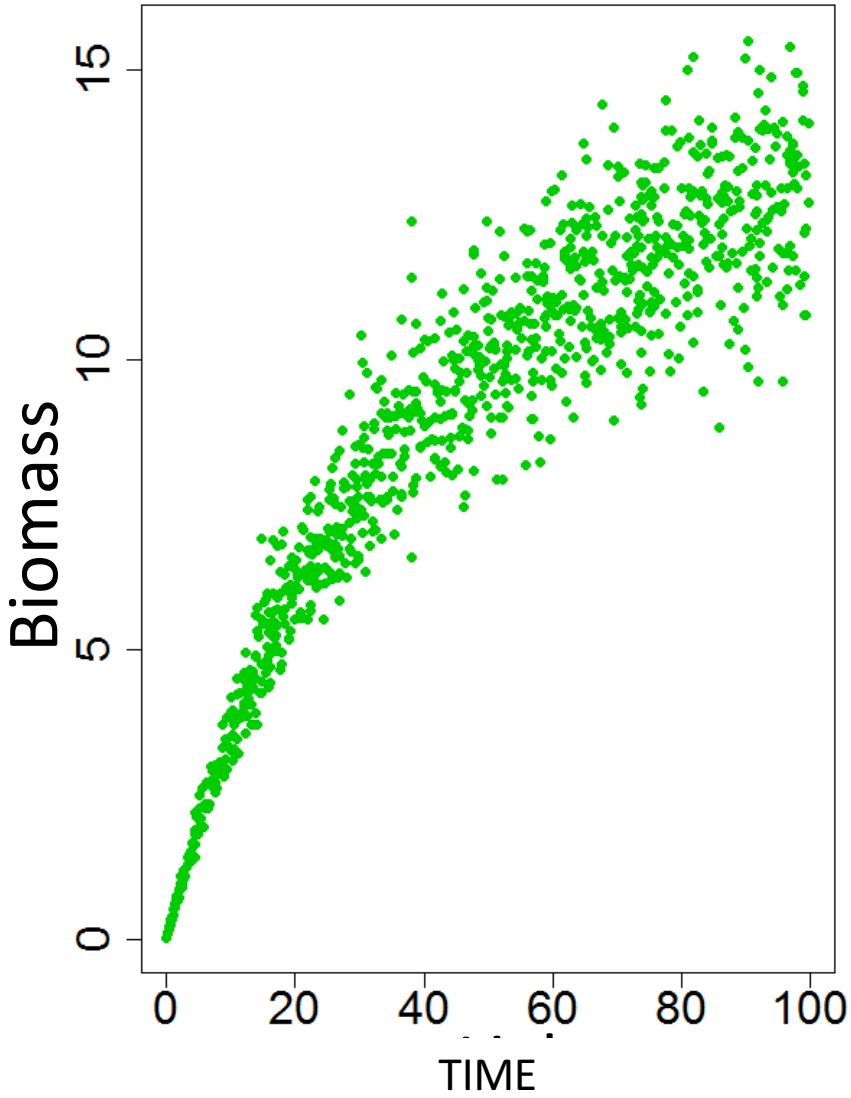
- Repeated measures from the same subject.
- Can use autoregressive terms to account for this.
- Can use random effects to account for this.
  - E.g. Truck drivers & Sleep deprivation treatment
  - Each observation nested within its subject

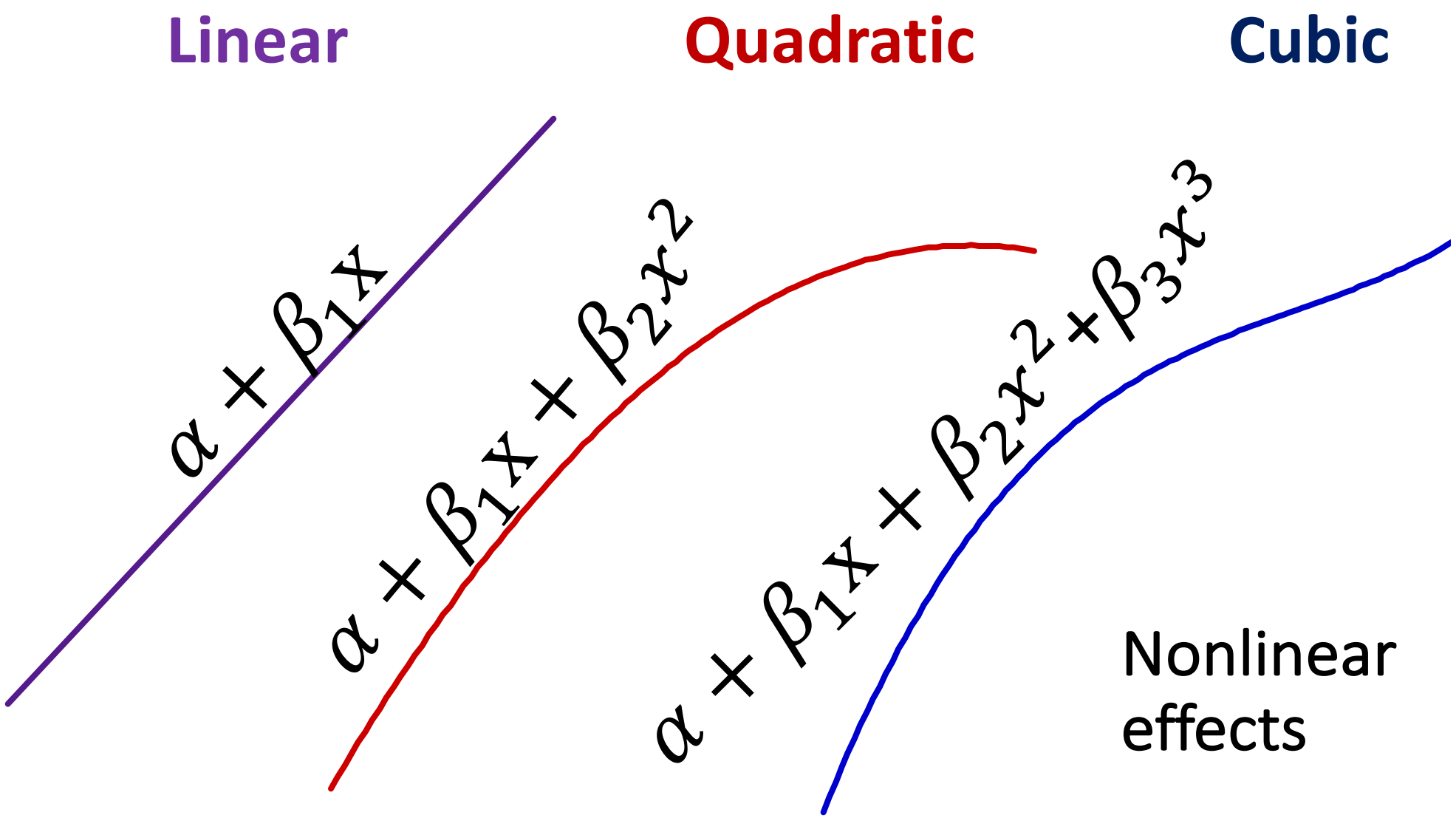


Non-linear responses

Time (and many other variables) may not have a linear effect





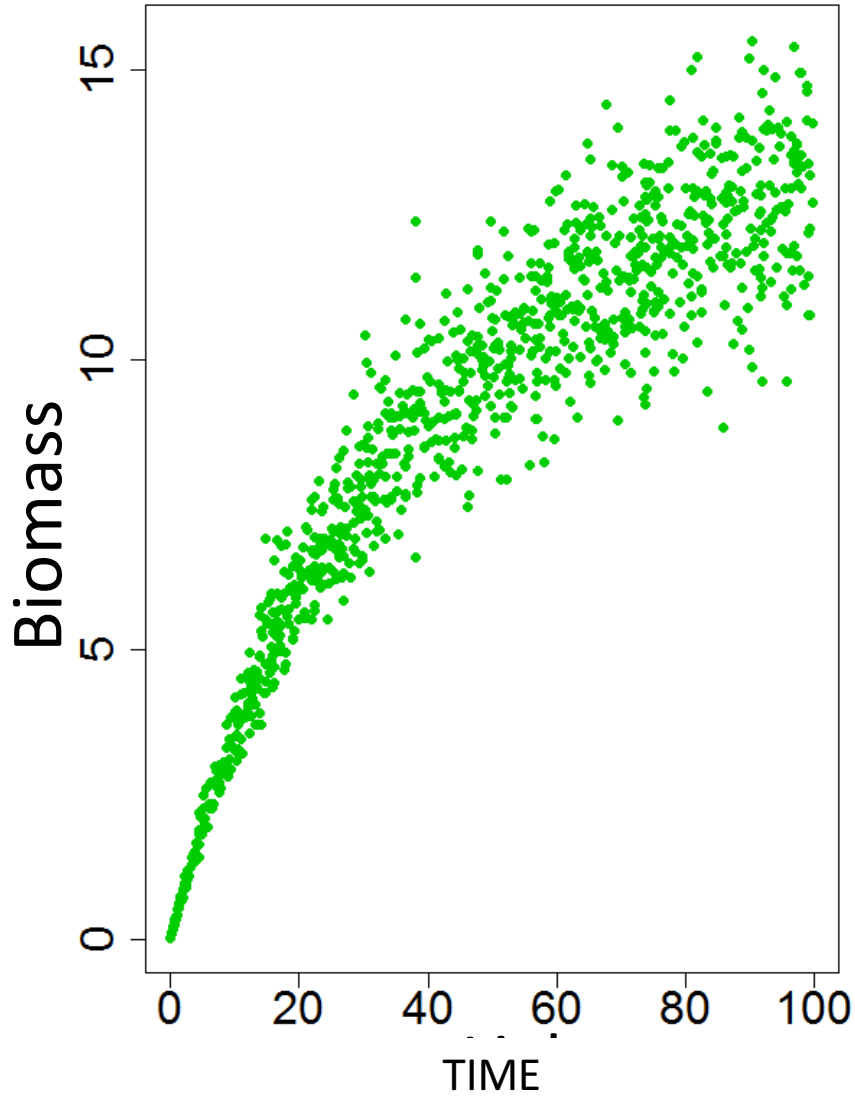

Linear
Quadratic
Cubic
$\alpha + \beta_1 x$
$\alpha + \beta_1 x + \beta_2 x^2$
$\alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
Nonlinear effects

Non-linear responses

Time does not always have a linear effect

In base R:

```
Time2 <- Time^2
glm(biomass~Time+Time2)
```

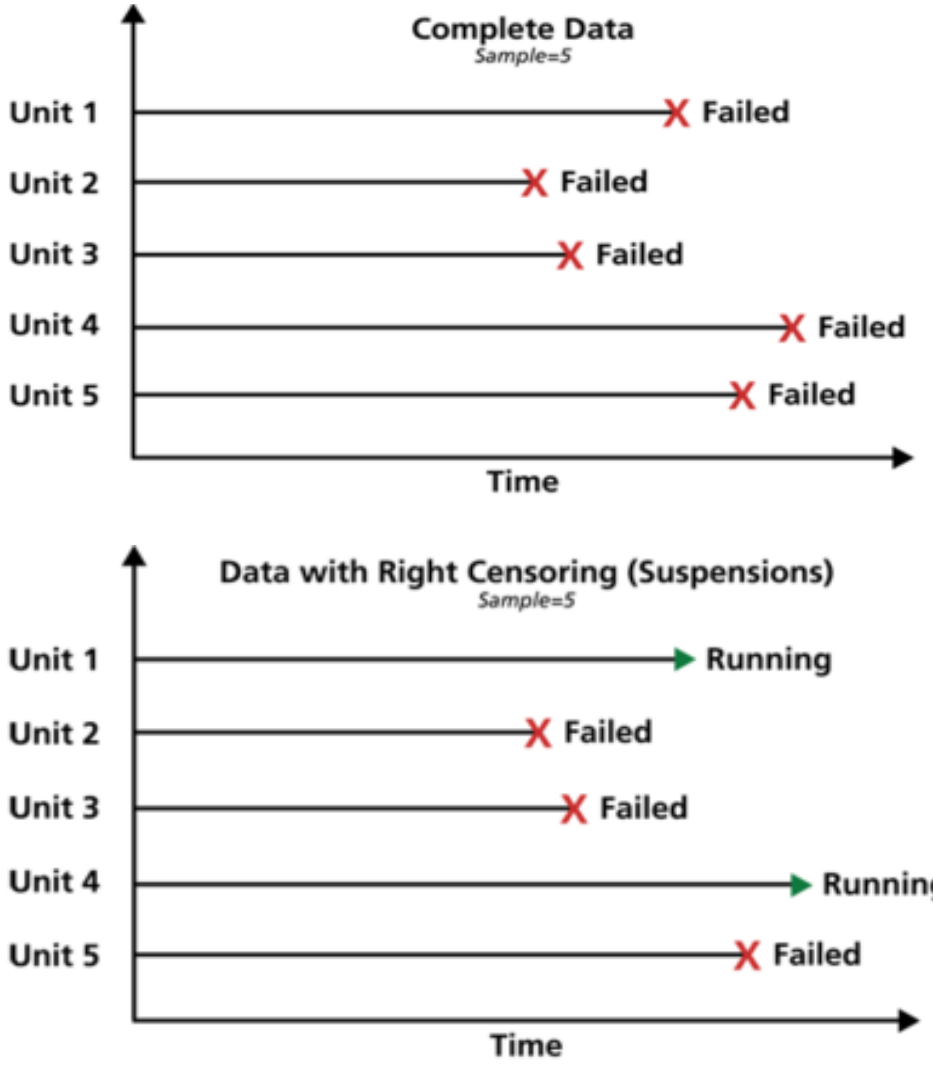


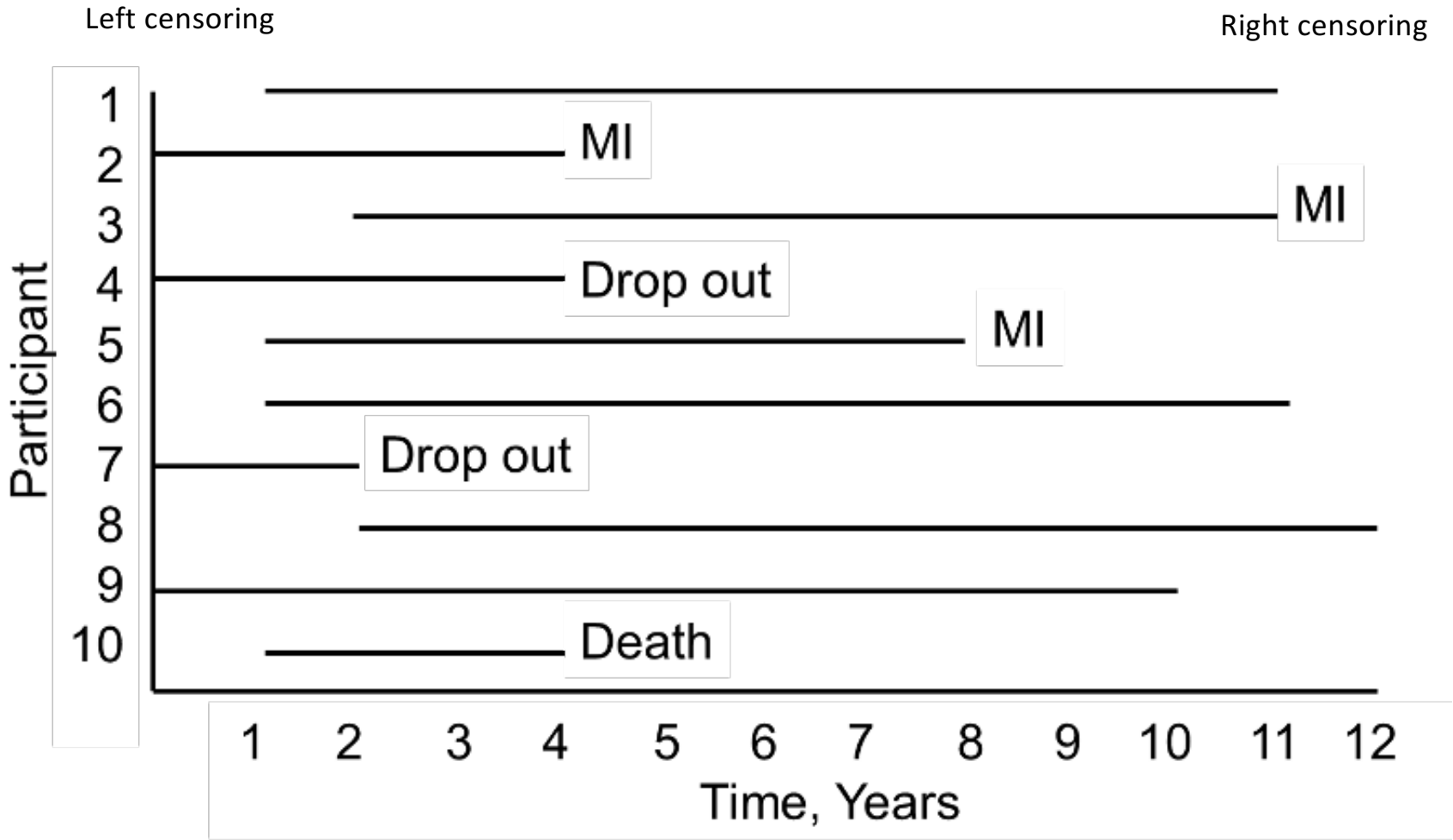
Survival analysis or "Time-to-Event" Models
(or "time to failure")

- Time until an individual dies (or is cured!)
- Time until a restored site reaches a threshold of native plant cover
- Time until a stream runs dry during the summer.
- Time until a dog gets adopted from the animal shelter.
  - Etc…
- OUTCOME = a duration
- Data structure:
  - Start time (of monitoring)
  - End time
  - If the event occurred
  - How long the individual was monitored
  - Data will often come with start and end dates rather than pre-calculated survival times

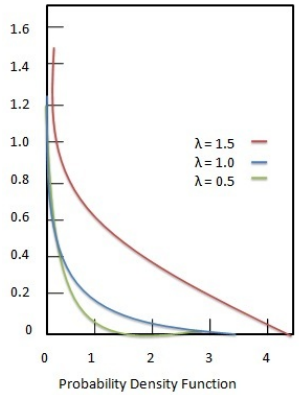
Censored data

- If there was complete information about each of our subjects, we could just use std. regression
- Event of interest does not always occur in the window of observation
  - Or, no follow-up (i.e. study funding ran out, human subject withdrew, animal not remeasured, etc).
- Can't just throw these out! It is a missing value. But we have some information that lets us narrow down its value.
- MODEL: Probability of any observed duration that end in the event as well as the probability that we would wait the observed duration without seeing the event




Left censoring
Right censoring
Participant
1
2
3
4
5
6
7
8
9
10
MI
MI
Drop out
MI
Drop out
Death
1
2
3
4
5
6
7
8
9
10
11
12
Time, Years

- Exponential distrib = special case of the gamma in which events have constant rate



If we witnessed the adoption:

$$D_i \sim \text{Exponential}(\lambda_i)$$

But what about the adoptions we did not witness??

Cumulative probability distribution

= Proportion of events that occurred by a certain time point

1- Cumulative probability distribution

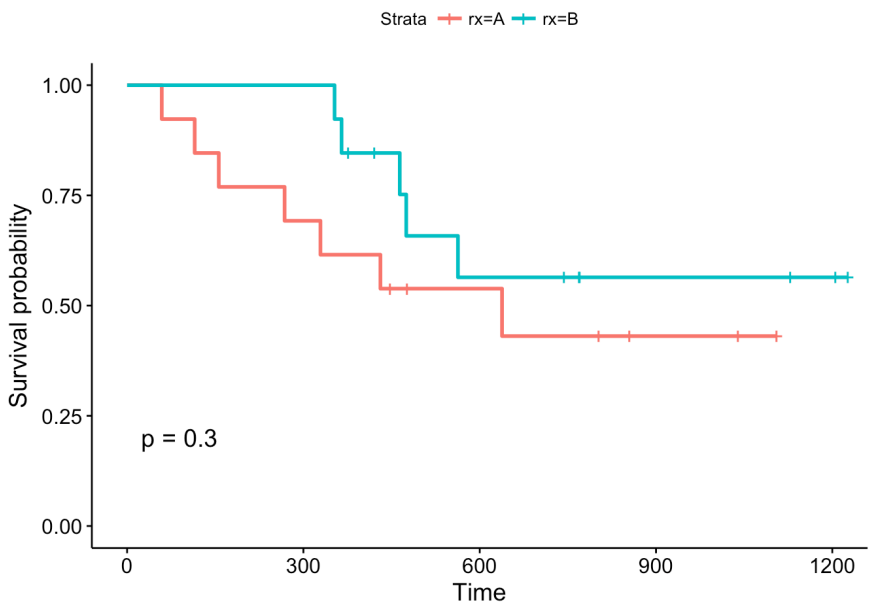
= Proportion of events that did not occur by a certain time point (AKA Survival)

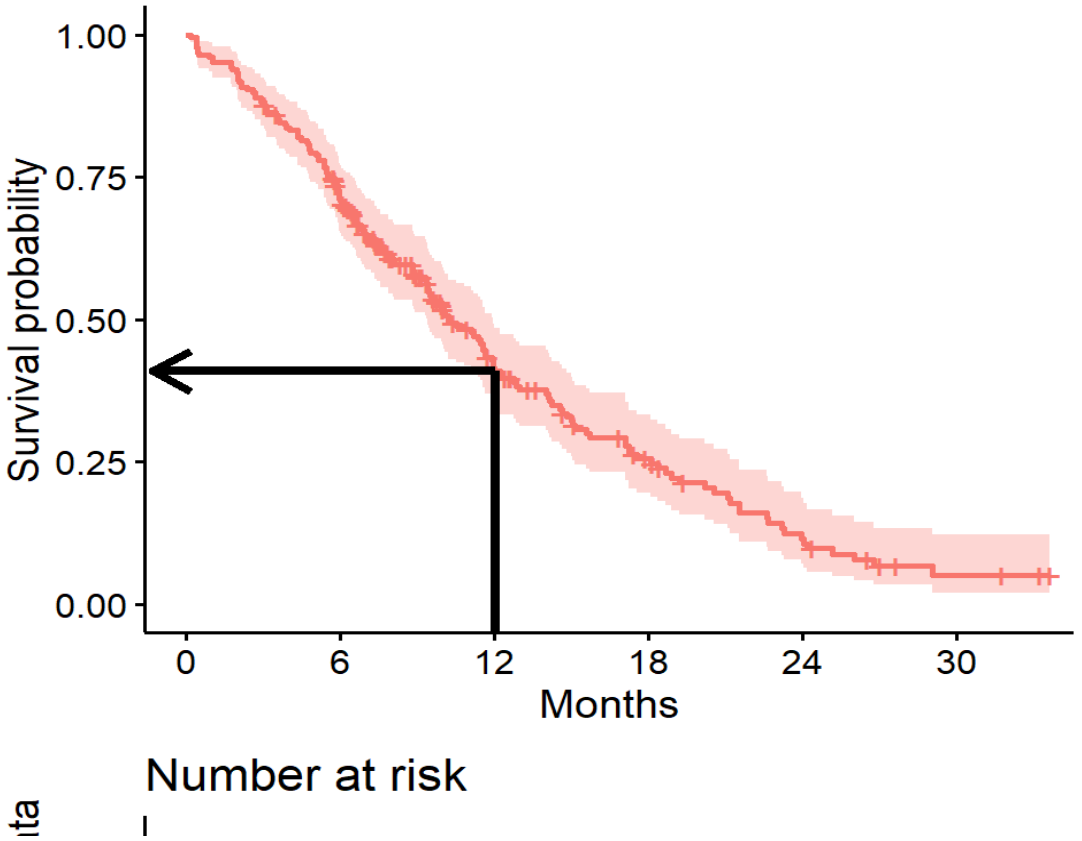
$$D_i|A_i = 1 \sim \text{Exponential}(\lambda_i)$$

$$D_i|A_i = 0 \sim \text{Exponential-CCDF}(\lambda_i)$$

$$\lambda_i = 1/\mu_i$$

$$\log \mu_i = a + b*X$$


Strata rx=A rx=B
1.00
0.75
0.50
0.25
0.00
Survival probability
p = 0.3
0
300
600
900
1200
Time


1.00
0.75
0.50
0.25
0.00
Survival probability
0
6
12
18
24
30
Months
Number at risk

Bayesian Survival Analysis Using the rstanarm R Package

Samuel L. Brilleman
Monash University, Melbourne, Australia

Eren M. Elci
Bayer AG, Berlin, Germany

Jacqueline Buros Novik
Generable Inc., New York, USA

Rory Wolfe
Monash University, Melbourne, Australia

Abstract

Survival data is encountered in a range of disciplines, most notably health and medical research. Although Bayesian approaches to the analysis of survival data can provide a number of benefits, they are less widely used than classical (e.g. likelihood-based) approaches. This may be in part due to a relative absence of user-friendly implementations of Bayesian survival models. In this article we describe how the **rstanarm** R package can be used to fit a wide range of Bayesian survival models. The **rstanarm** package facilitates Bayesian regression modelling by providing a user-friendly interface (users specify their model using customary R formula syntax and data frames) and using the Stan soft-

Exercise

- Create a survival analysis assessing the “time-to-adoption” for pets in an animal shelter, assessing differences between juvenile and adult animals.

CONGRATULATIONS!

- **For Thursday!**
  - Review game for bonus points
  - Wrap up and reflection from the course
  - Course evaluations