Introduction to Bayesian Statistics for Demographic Science

Day 2: Generalised linear models (GLMs) & hierarchical models

Outline

- General linear model (refresher)
- Generalised linear models (GLMs)
- Mark-recapture
- Hierarchical models
- State-space

Components of a model

1. Data model(s)

$$y_i \sim Normal(\mu_i, \sigma)$$

2. Process model(s)

$$\mu_i = \alpha + \beta x_i$$

3. Parameter model(s)

$$\alpha \sim Normal(0, 1)$$
$$\beta \sim Normal(0, 1)$$
$$\sigma \sim HalfCauchy(0, 1)$$

Diagnostics

- MCMC Diagnostics
  - Trace plots
  - Rhat

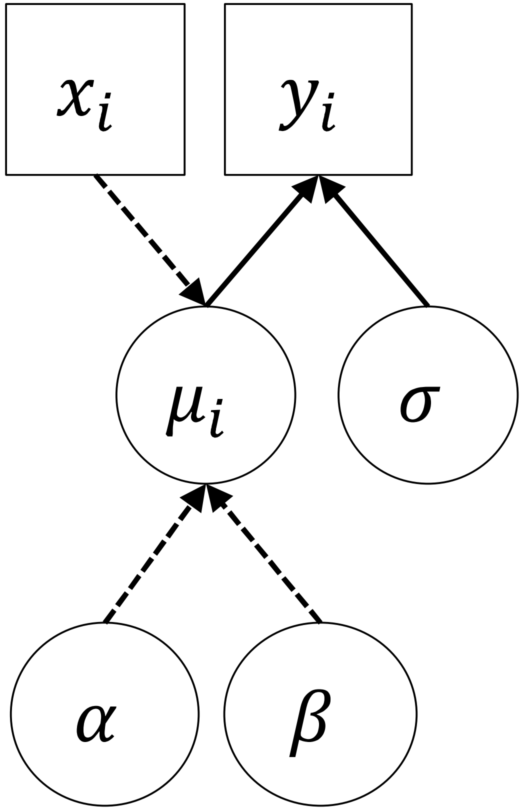
- Model fit
  - Out-of-sample cross validation
  - Posterior predictive checks
  - (your normal way of assessing fit for any other statistical model)

Linear regression

General linear model

General linear model

Refresher



$$y_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

$$\alpha \sim Normal(0,1)$$
$$\beta \sim Normal(0,1)$$
$$\sigma \sim HalfCauchy(0,1)$$

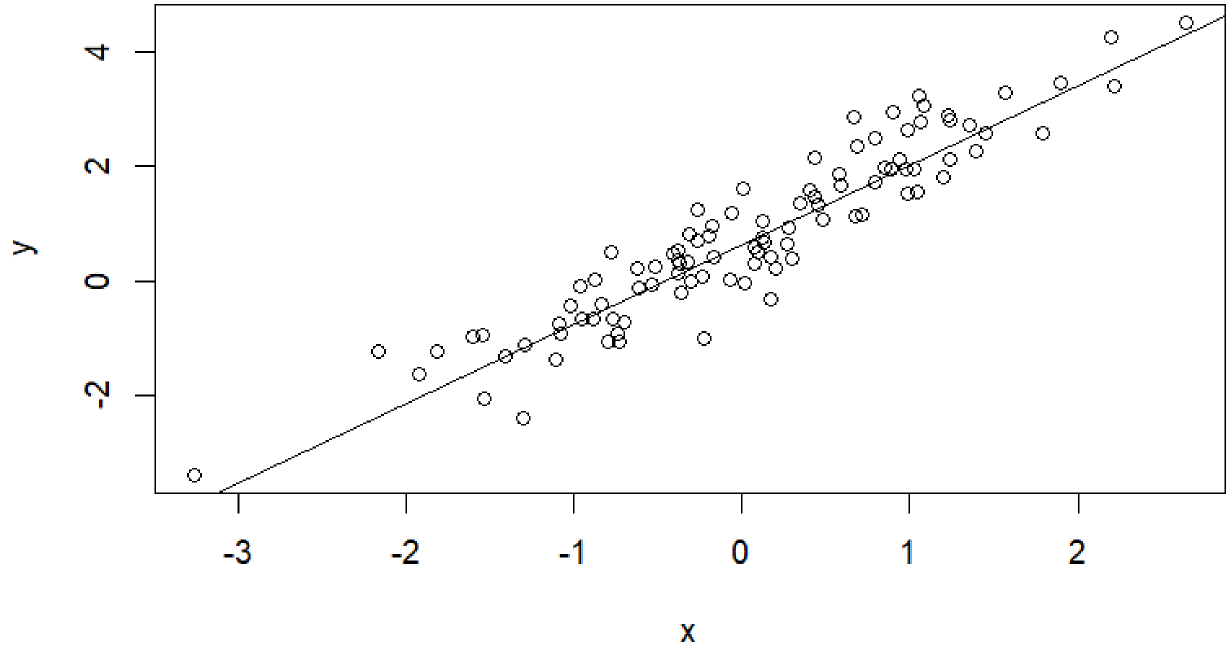
General linear model

Refresher

What are examples of quantities that y_i could represent?
Examples that it cannot be?

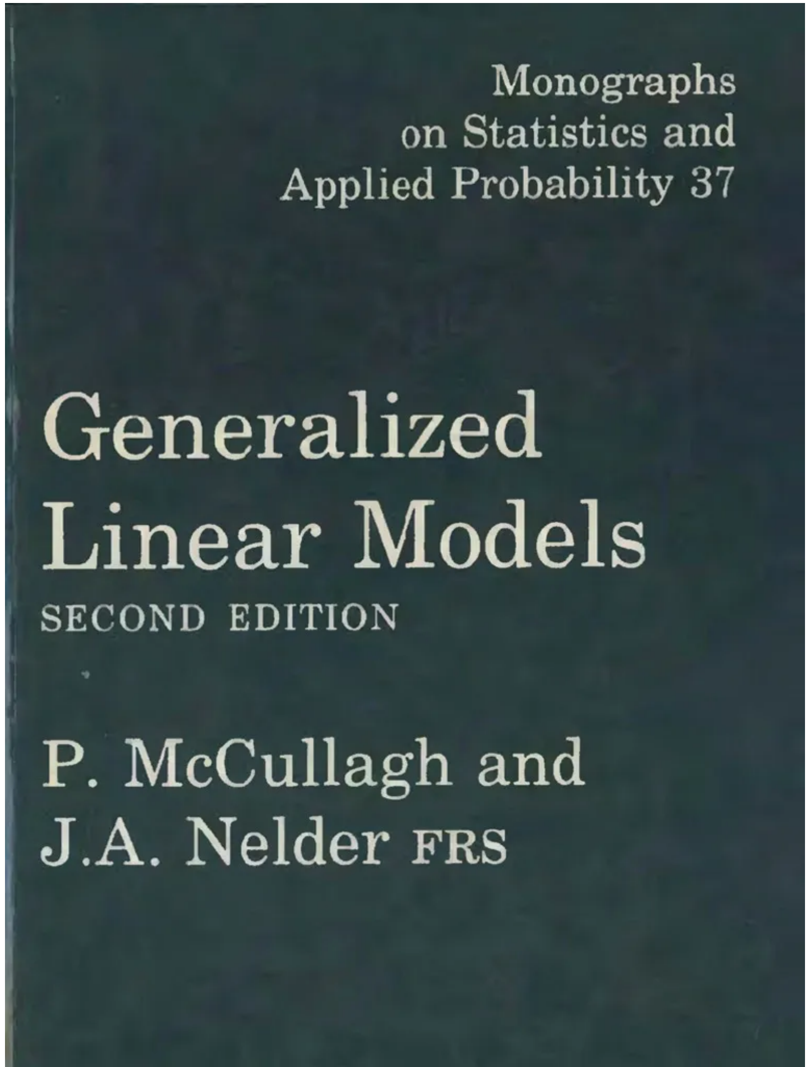
$$y_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$



$$\alpha \sim Normal(0,1)$$
$$\beta \sim Normal(0,1)$$
$$\sigma \sim HalfCauchy(0,1)$$

Generalised linear models



Monographs
on Statistics and
Applied Probability 37

Generalized
Linear Models
SECOND EDITION

P. McCullagh and
J.A. Nelder FRS

Generalised linear models

- Generalisation of a linear model to accommodate additional data types
- Three components:
    - Probability distribution
    - A linear predictor
    - Link function

Can you sketch a model?

We have **counts of people** y_i from a sample of locations i across a country.

We want to infer the relationship of population size with the **unemployment rate** $x_{1,i}$, **cost of living** $x_{2,i}$, and **number of pubs** $x_{3,i}$.

Generalised linear model

Poisson regression for count data

$$y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i}$$

$$\alpha \sim Normal(0, 15)$$
$$\beta \sim Normal(0, 10)$$

Can we build a model for these data?

We have **counts of active Facebook users** y_i from one hundred countries i. We also know the **total population** N_i of each country.

We want to study how country characteristics influence the probability that a person uses Facebook.

We have data on **gross domestic product** $x_{1,i}$, **broadband internet speeds** $x_{2,i}$, and **rates of attainment for secondary education** $x_{3,i}$.

Generalised linear model

Binomial regression for count data

$$y_i \sim Binomial(N_i, \theta_i)$$

$$logit(\theta_i) = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i}$$

$$\alpha \sim Normal(0, 1e3)$$
$$\beta \sim Normal(0, 1e3)$$

How about this one?

We have GitHub data from individual **Russian programmers** i **indicating if they have migrated (or not)** to a new country since 2022.

We want to study how individual characteristics influence their probability of migration.

From their profile, we have their **number of followers** $x_{1,i}$, **number of contributions** in the past year $x_{2,i}$, and **number of popular code repositories** $x_{3,i}$.

Generalised linear model

Logistic regression for binary outcomes

$$y_i \sim Bernoulli(\theta)$$

$$logit(\theta) = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i}$$

$$\alpha \sim Normal(0, 1e3)$$
$$\beta \sim Normal(0, 1e3)$$

Some other useful models beyond GLMs…

What do you think about this one?

We have data on the **proportion women** among employees of the top tech companies i globally.

We want to study how these ratios are influenced by company characteristics like **female CEO** $x_{1,i}$, **gender pay gap** $x_{2,i}$, and **days paid maternity leave** $x_{3,i}$.

Beta regression

$$y_i \sim Beta(\mu\tau, \tau - \mu_i\tau)$$

$$logit(\mu_i) = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i}$$

$$\alpha \sim Normal(0, 1e3)$$
$$\beta \sim Normal(0, 1e3)$$
$$\tau \sim Uniform(0, 1e3)$$

Last one…

We have data on **total fertility rates for every major city** i in the world.

We want to study how these rates are influenced by **population density** $x_{1,i}$, **hospitals per capita** $x_{2,i}$, and **4G coverage** $x_{3,i}$.

Log-linear regression

$$y_i \sim LogNormal(\mu_i, \sigma)$$

$$\mu_i = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i}$$

$$\alpha \sim Normal(0, 1e3)$$
$$\beta \sim Normal(0, 1e3)$$
$$\sigma \sim HalfNormal(0, 1)$$

What if we also wanted to include a **country effect**?
How would you do that?

Hierarchical Models

Hierarchical models

- Multiple levels (i.e. sub-models) within the same model
- The outcome of one level is an input to another level

Why do we want to do this?

- Sharing information among groups to increase statistical power
- Processes occurring at multiple scales (i.e. spatial, temporal, etc)
- Partition uncertainty into process error and observation error
- Different deterministic functions (i.e. regressions) for different levels of the model

Random effects

- Effects (e.g. intercept, slope) are dependent on another factor
- Usually a grouping factor (g)
- The effects are not independent among groups
- Prior model and hyper-priors

$$y_i \sim Poisson(\lambda_i)$$

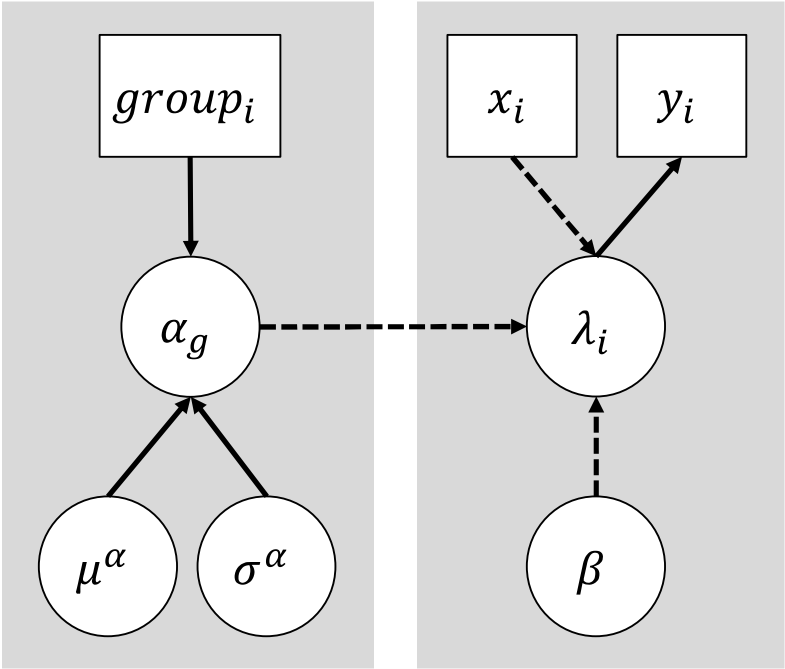
$$log(\lambda_i) = \alpha_g + \beta x_i$$

$$\alpha_g \sim Normal(\mu^\alpha, \sigma^\alpha)$$

$$\mu^\alpha \sim Normal(0, 1e3)$$
$$\sigma^\alpha \sim Cauchy(0, 1e3)$$

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1(3):515-533.

Random effects



$$y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \alpha_g + \beta x_i$$

$$\alpha_g \sim Normal(\mu^\alpha, \sigma^\alpha)$$

$$\mu^\alpha \sim Normal(0, 1e3)$$
$$\sigma^\alpha \sim Cauchy(0, 1e3)$$

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1(3):515-533.

Mark-recapture

- A method originally designed for estimating unobservable animal populations
- Useful in demography for estimating populations that are impossible to census
- **Lincoln-Peterson mark recapture** in the oldest and simplest approach (assuming a closed population)
- **Cormack-Jolly-Seber method** is an open-population method in which the population can change due to mortality.
- **“Multi-systems estimation”** is a based on these approaches and has been used to estimate casualties from armed conflicts (including unobserved casualties).

Manrique-Vallier et al. 2021. Capture-recapture for casualty estimation and beyond: Recent advances and research directions. Chapter in *Statistics in the Public Interest.* Pp. 15-31. Springer.

Stan User’s Guide. Mark-recapture models. https://mc-stan.org/docs/stan-users-guide/latent-discrete.html#mark-recapture-models

Mark-recapture model (Lincoln-Peterson)

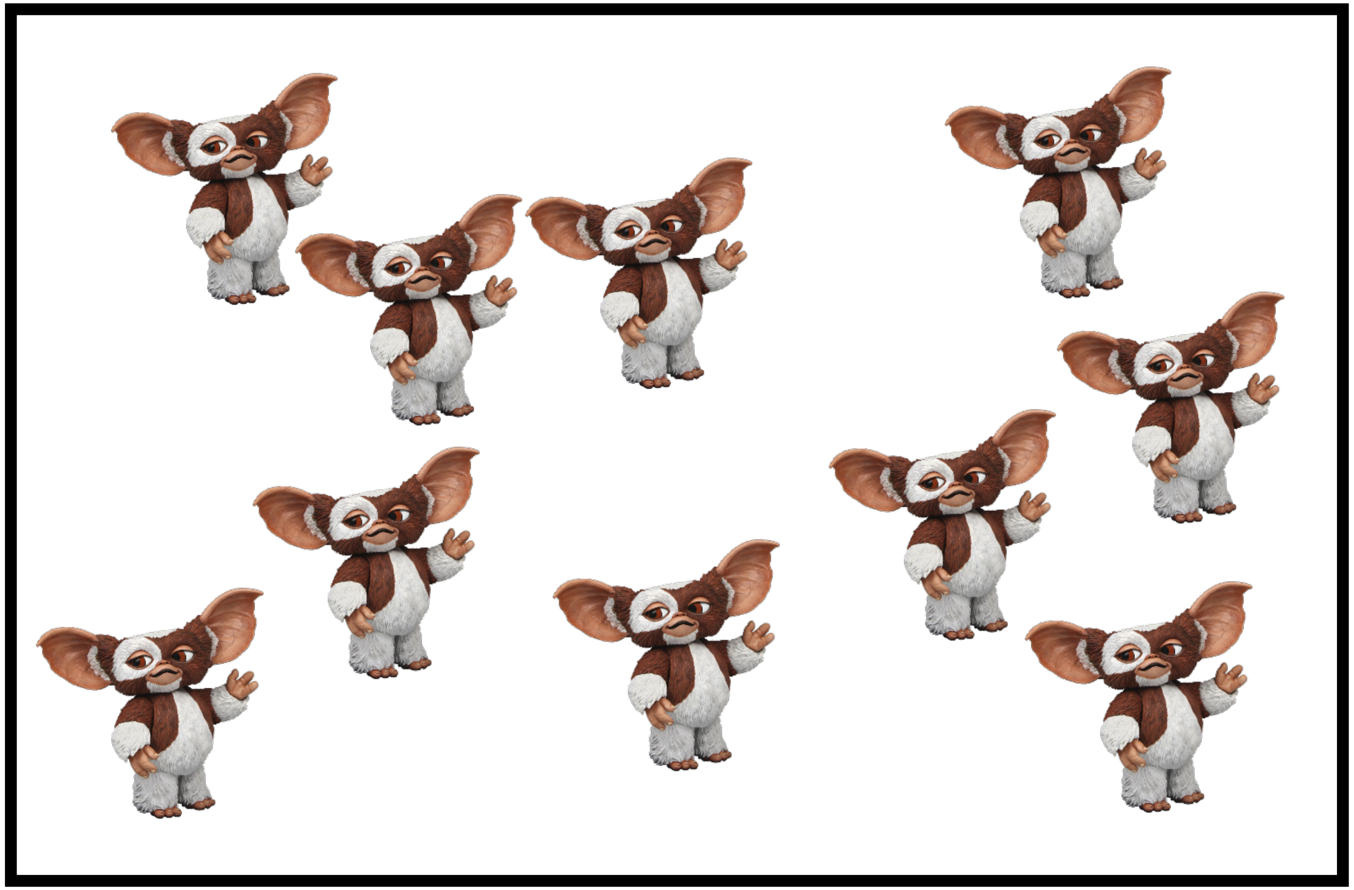
- Estimate an unobservable population size from mark-recapture data
- Conduct a survey randomly sampling M "marked" individuals from the population
- Conduct a second round of the survey randomly sampling C "captured" individuals
- Count the number of "recaptured" individuals R in the second round

$$R \sim Binomial\left(C, \frac{M}{N}\right)$$

$$N \sim Uniform(C - R + M, 1e6)$$

Mark-recapture model (Lincoln-Peterson)

survey = 1, mogwai = 10

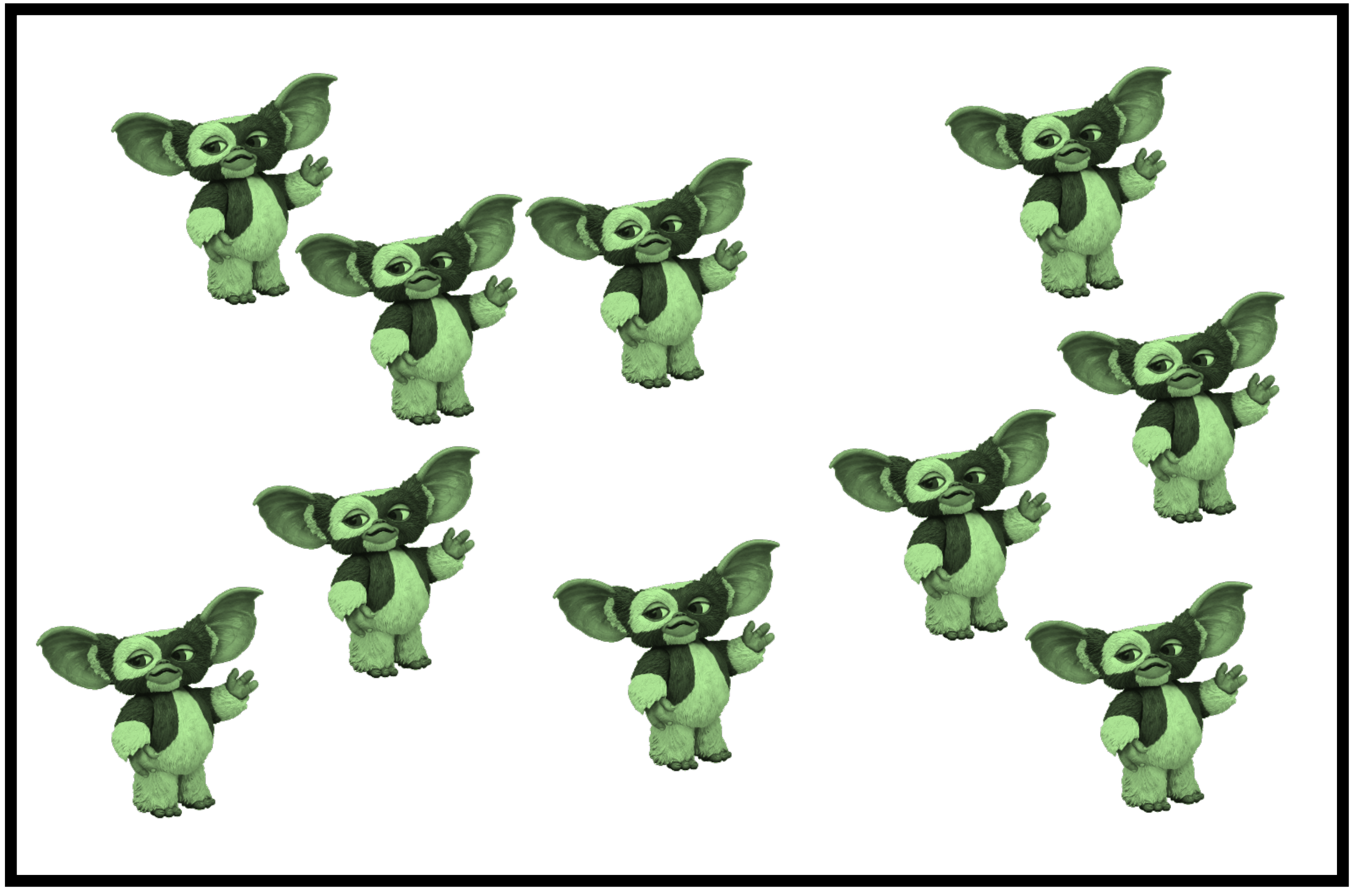


$$R \sim Binomial\left(C, \frac{M}{N}\right)$$

$$N \sim Uniform(C - R + M, 1e6)$$

Mark-recapture model (Lincoln-Peterson)

survey = 1, mogwai = 10



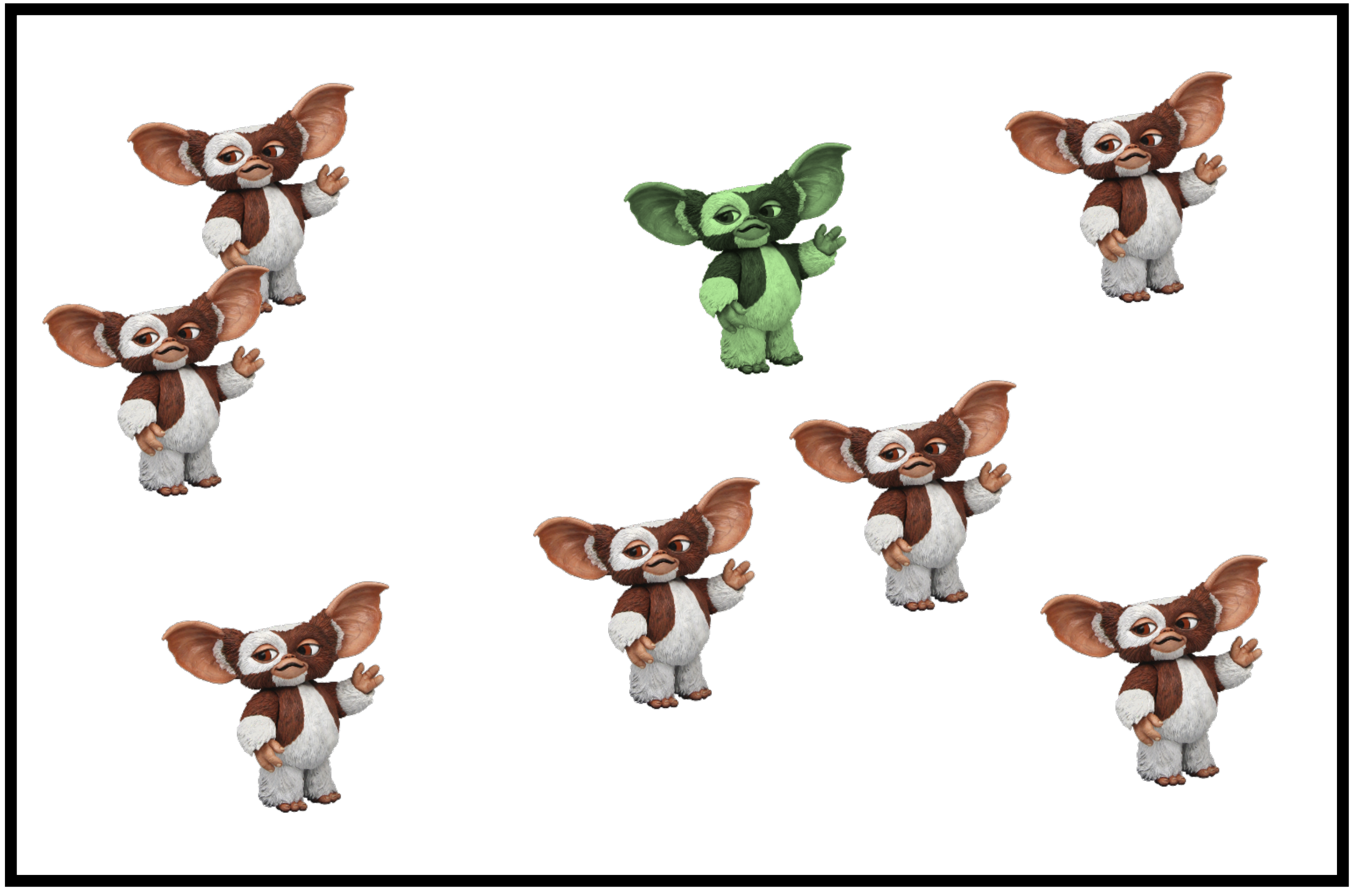
$$R \sim Binomial\left(C, \frac{M}{N}\right)$$

$$N \sim Uniform(C - R + M, 1e6)$$

Marked (M) = 10

Mark-recapture model (Lincoln-Peterson)

survey = 2, mogwai = 8



$$R \sim Binomial\left(C, \frac{M}{N}\right)$$

$$N \sim Uniform(C - R + M, 1e6)$$

Marked (M) = 10

Captured (C) = 8

Recaptured (R) = 1

Mark-recapture model (Lincoln-Peterson)

What is the big assumption in this model?

$$R \sim Binomial\left(C, \frac{M}{N}\right)$$

The detection probability in round 1 is equal to the probability of round 2 captures being recaptures.

$$N \sim Uniform(C - R + M, 1e6)$$

$$\frac{M}{N} = \frac{R}{C}$$

Sometimes we can’t “mark” individuals

A “mark” could simply be recording a person’s name and birthdate, but sometimes this is not possible.

N-mixture models of population size

- Repeated j counts $y_{i,j}$ of individuals at each location i.
- The repeated counts allow us to estimate detection probability $\rho_{i,j}$

$$y_{i,j} \sim Binomial\left(N_i, \rho_{i,j}\right)$$

$$logit\left(\rho_{i,j}\right) = \alpha + \beta x_{i,j}$$

- Population size N_i determined by characteristics of each location $x_{k,i}$.
- Assumes a closed population across j surveys

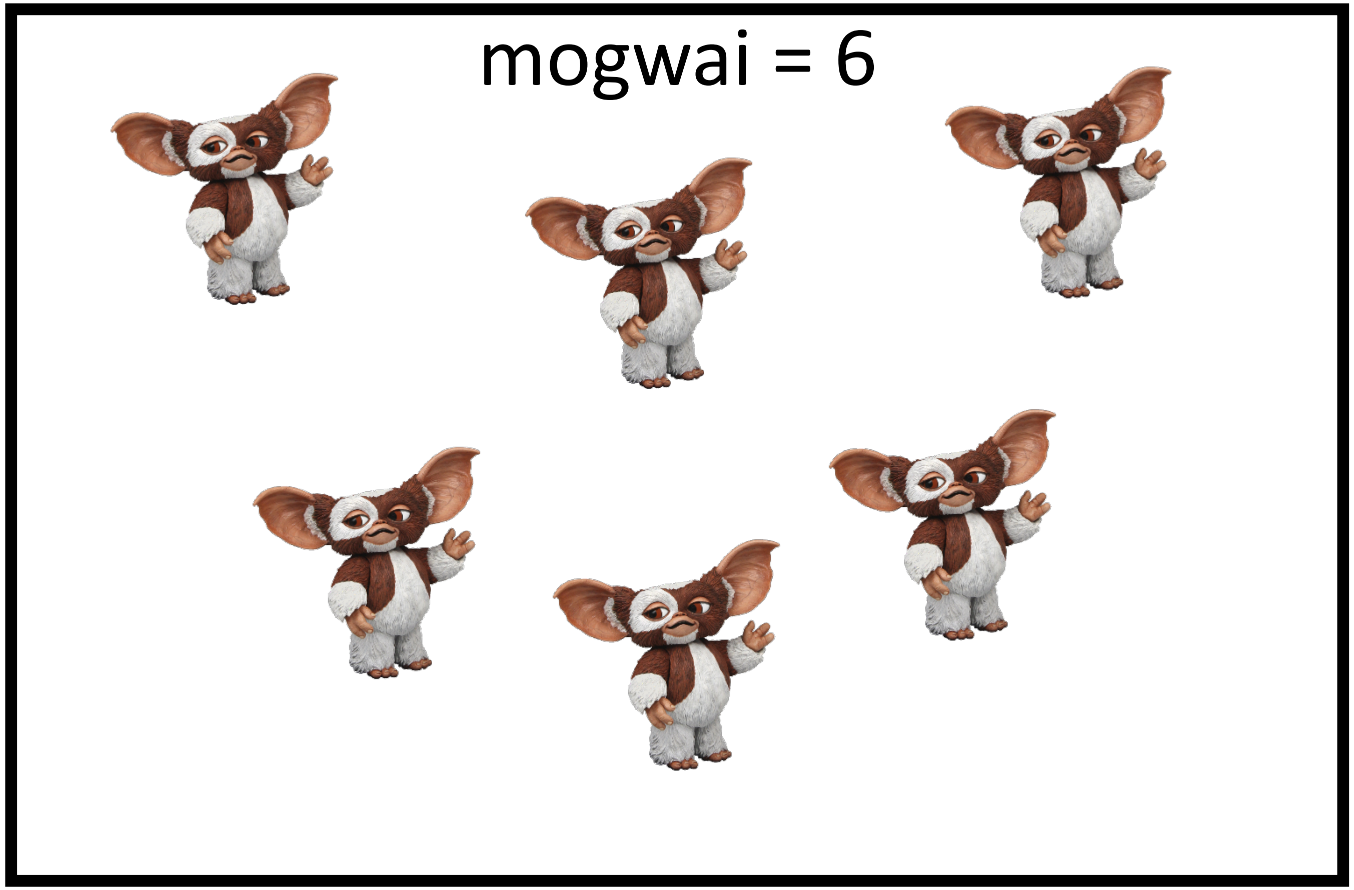
$$N_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \alpha + \beta x_i$$

Royle A. 2004. N-mixture models for estimating population size from spatially replicated counts. *Biometrics* 60(1):108-115.

N-mixture models of population size

location (i) = 1, survey (j) = 1



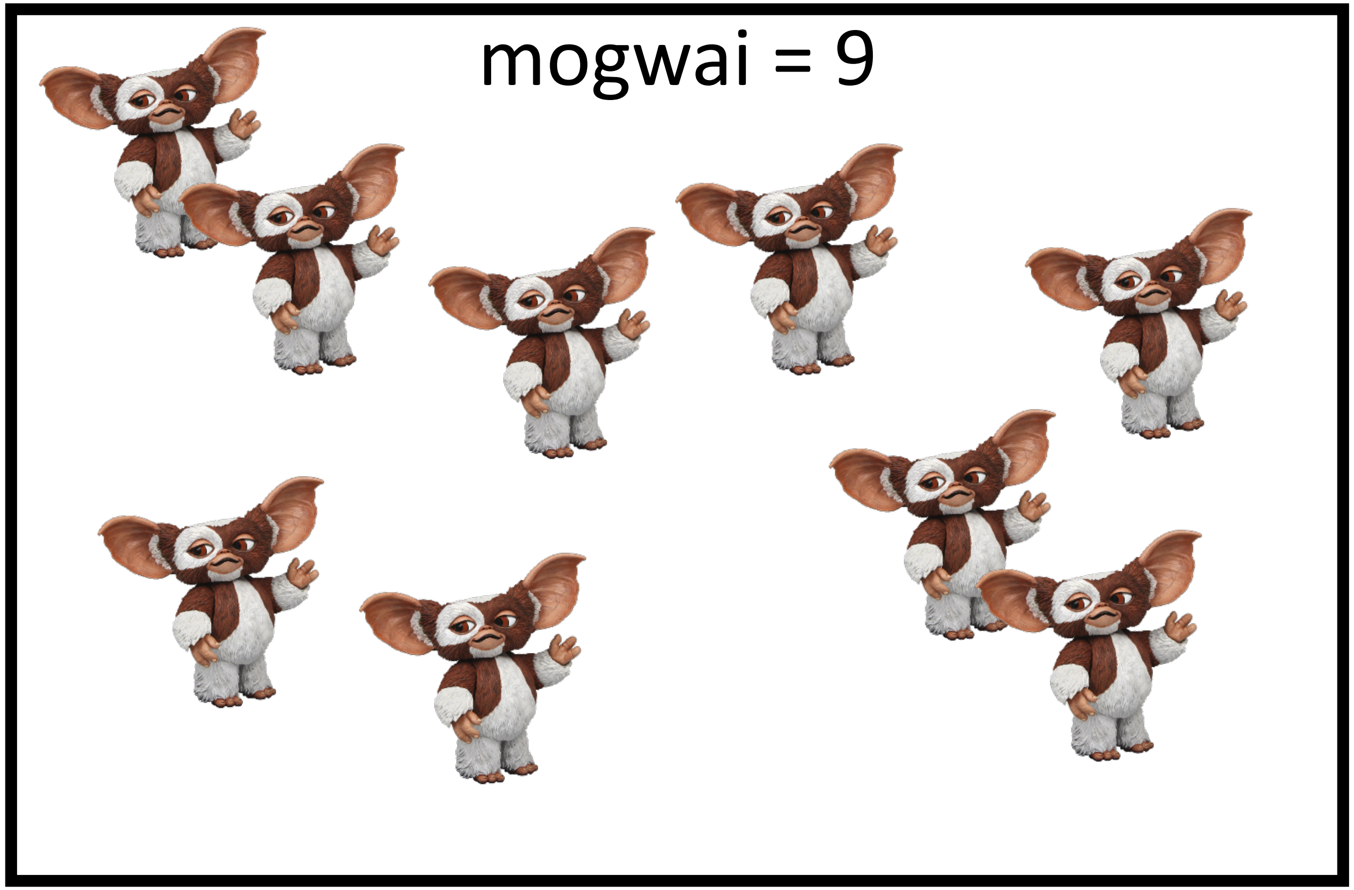
mogwai = 6

$$y_{1,1} = 6$$

Royle A. 2004. N-mixture models for estimating population size from spatially replicated counts. *Biometrics* 60(1):108-115.

N-mixture models of population size

location (i) = 1, survey (j) = 2



mogwai = 9

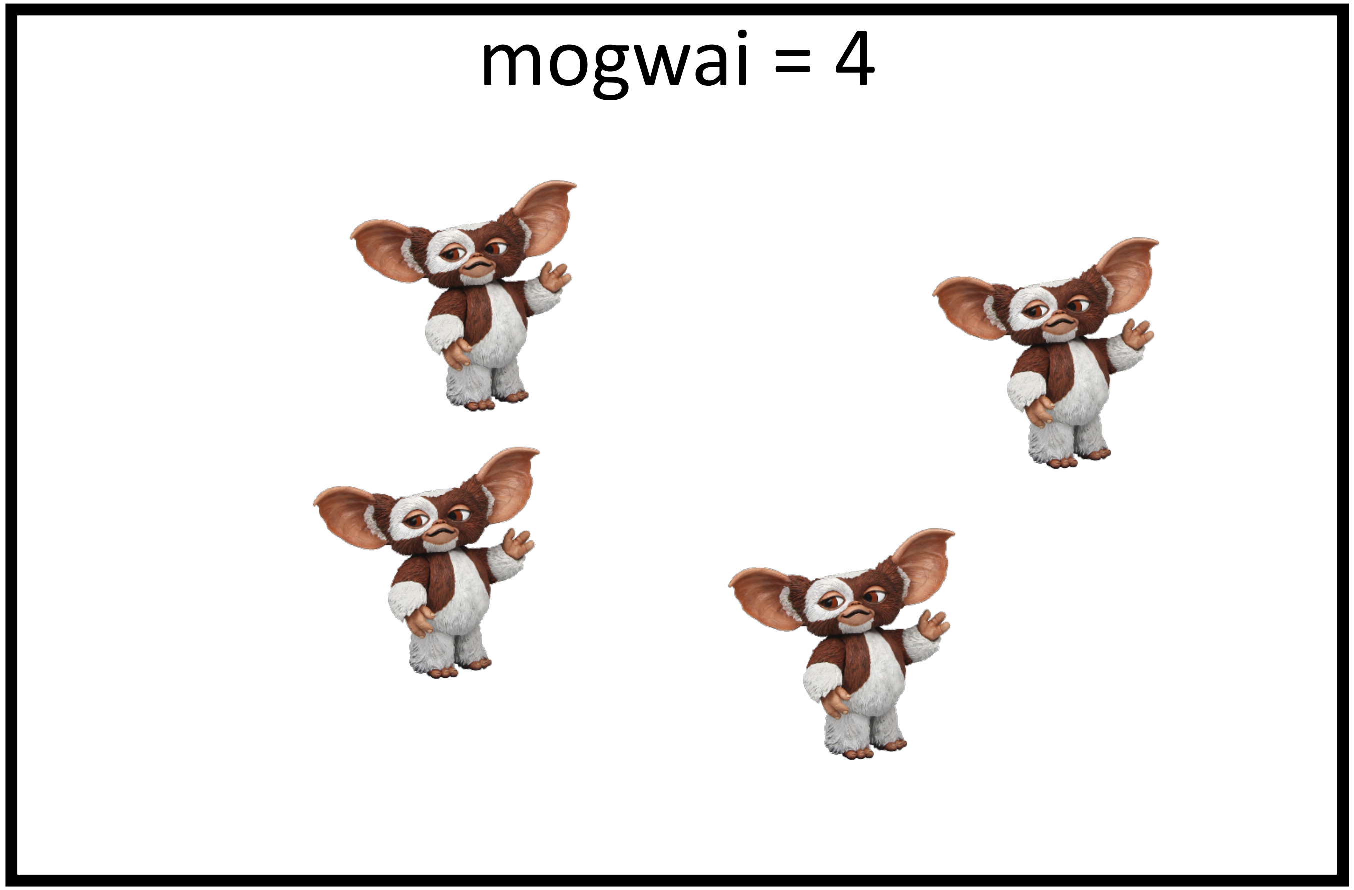
$$y_{1,1} = 6$$

$$y_{1,2} = 9$$

Royle A. 2004. N-mixture models for estimating population size from spatially replicated counts. *Biometrics* 60(1):108-115.

N-mixture models of population size

location (i) = 1, survey (j) = 3



mogwai = 4

$$y_{1,1} = 6$$

$$y_{1,2} = 9$$

$$y_{1,3} = 4$$

Royle A. 2004. N-mixture models for estimating population size from spatially replicated counts. *Biometrics* 60(1):108-115.

N-mixture models of population size

Observation model:

$$y_{i,j} \sim Binomial\left(N_i, \rho_{i,j}\right)$$

$$logit\left(\rho_{i,j}\right) = \alpha + \beta x_{i,j}$$

Process model:

$$N_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i}$$

Royle A. 2004. N-mixture models for estimating population size from spatially replicated counts. *Biometrics* 60(1):108-115.

State-space models

- Unobserved process in which the state z_t of the system at time t depends on the state in the previous time step with stochastic variation ϵ

$$P(z_t \mid \theta z_{t-1}, \epsilon)$$

- Observed data y_t that are dependent on the process state z_t with observation error σ

$$P(y_t \mid \beta z_t, \sigma)$$

- Make inferences about a process that you cannot observe directly
- Estimate process error separately from observation error

State-space models

Hidden Markov model is a specific case of this

Process model: $P(z_t \mid \theta z_{t-1}, \epsilon)$

Observation model: $P(y_t \mid \beta z_t, \sigma)$

Dail-Madsen open population models

- Apparent survival $S_{i,t}$ is the number of individuals surviving and not emigrating
- Recruitment $G_{i,t}$ is the number of new individuals entering the population

$$S_{i,t} \sim Binomial(N_{i,t-1}, \omega)$$

$$G_{i,t} \sim Poisson(\lambda N_{i,t-1})$$

$$N_{i,t} = S_{i,t} + G_{i,t}$$

- N-mixture observation model

$$y_{i,j,t} \sim Binomial(N_{i,t}, \rho_j)$$

$$logit(\rho_j) \sim Normal(\mu^{\rho}, \sigma^{\rho})$$

Dail & Madsen. 2011. Estimating abundance from repeated counts of an open metapopulation. *Biometrics* 67:577-587.
Hostetler & Chandler 2015. Improved state-space models for inference about spatial and temporal variation in abundance. *Ecology* 96(6)

Mark-recapture state-space model

- A population process
- Why a log-normal? Hmm….
- Priors not shown

$$N_t \sim LogNormal(\log(N_{t-1}\lambda_t), \sigma)$$

$$\log(\lambda_t) = \alpha + \sum_{k=1}^{K} \beta_k x_{k,t}$$

- Mark-recapture observations
- A prior is only needed to initialise N in the first time step

$$R_t \sim Binomial\left(C_t, \frac{M_t}{N_t}\right)$$

$$N_{t=1} \sim Uniform(C - R + M, 1e6)$$

Stan and discrete latent parameters

- Stan does not allow parameters to be integers. Only data can be integers.
- This is infinitely annoying, but the Stan developers are firm in their crusade against discrete latent parameters in the noble pursuit of accurately estimating the tails of distributions and increasing MCMC efficiency.
- If you have a model with a latent discrete parameter (e.g. unobserved population size), then you have three options:
  1. Re-parameterise your model to "marginalise out" the integer parameter.
  2. Use a continuous distribution for your integer parameter (e.g. LogNormal).
  3. Use JAGS instead of Stan.

Stan Users Guide. Latent discrete parameters. https://mc-stan.org/docs/stan-users-guide/latent-discrete.html#latent-discrete.chapter
See Stan Forums for discussion on marginalising out integer parameters for various types of specific models.

Recap

- Ordinary linear models
- Generalised linear models
  - Logistic regression
  - Binomial regression
  - Poisson regression
- Log-linear models
- Beta regression
- Mark-recapture

Hierarchical Models

- Random effects
- N-mixture models
- State-space
  - Dail-Madsen N-mixture
  - Mark-recapture state-space

Components of a model

1. Data model (likelihood) $y_i \sim Normal(\mu_i, \sigma)$

2. Deterministic process $\mu_i = \alpha + \beta x_i$

3. Parameter model (*e.g.* priors)

$$
\begin{aligned}
&\alpha \sim Normal(0,1)\\
&\beta \sim Normal(0,1)\\
&\sigma \sim HalfCauchy(0,1)
\end{aligned}
$$

Tie it all together…

Ordinary linear regression

$$y_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

Priors not shown

Tie it all together…

Generalised linear model

$$y_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \alpha + \beta x_i$$

Priors not shown

Tie it all together…

Generalised linear model

$$y_i \sim Binomial(N_i, \rho_i)$$

$$\text{logit}(\rho_i) = \alpha + \beta x_i$$

Priors not shown

Tie it all together…

Random effects (hierarchical model)

$$y_i \sim Binomial(N_i, \rho_i)$$

$$\mathrm{logit}(\rho_i) = \alpha_g + \beta x_i$$

$$a_g \sim Normal(\mu, \sigma)$$

Priors not shown

Tie it all together…

Hierarchical models (e.g. N-mixture model)

Latent process model:

$$N_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \alpha + \beta x_i$$

Observation model:

$$y_{i,j} \sim Binomial(N_i, \rho_{i,j})$$

$$\text{logit}(\rho_{i,j}) = \alpha + \beta x_{i,j}$$

Priors not shown

Tie it all together…

State space model

$$N_{i,t} \sim Poisson(N_{i,t-1}\lambda_{i,t})$$

$$\log(\lambda_{i,t}) = \alpha + \beta x_{i,t}$$

$$y_{i,j,t} \sim Binomial(N_{i,t}, \rho_{i,j,t})$$

$$\text{logit}(\rho_{i,j,t}) = \alpha + \beta x_{i,j,t}$$

Priors not shown

Tie it all together…

State space model

$$P(z_t \mid \lambda z_{t-1})$$

$$P(y_t \mid \rho z_t)$$

Priors not shown