Generalized linear models

Liliana M. Dávalos
Ecology & Evolution
10 July 2013

Stony Brook University | The State University of New York

Today's lecture

- Model checking
  - Prediction
  - Robustness
- Hypothesis testing
  - Bayes Factors
  - BIC
- Non-Gaussian responses
  - Count variables
  - Binary variables

Model checks

- Want to determine how **sensitive** the model is to decisions we made in building it
- Also need to see if it captures salient aspects of the data or has been tainted by the data (outliers)
  - This is **robustness**
- In the end, we want to verify that we have **model adequacy** or that we have properly characterized the uncertainty
- Note the separate parts here (local) sensitivity, (global) robustness, and model adequacy
- This is distinct from hypothesis testing (more on that later)

From: Brandt 2011

Components to be checked

- Prior and hyperpriors
- Likelihood

From: Brandt 2011

Sensitivity analysis

- Informal process of checking how the posterior responds to assumptions, e.g.,
  - Which variables belong in the model
    - Add or take away variables
  - Choice of the form and parameters in the prior
  - Look at critical model diagnostics
- This is an informal analysis of the effects of local changes on the posterior

From: Brandt 2011

Robustness

- This is a total, posterior insensitivity to assumptions. It is a systematic process looking at how the posterior responds to **misspecification** of the prior, likelihood and data outliers
  - Global robustness looks at how changes in the prior affect inferences from the posterior.
  - Local robustness looks at the volatility of results around key values (usually with differential calculus)
  - Both are often specific to a model or application

From: Brandt 2011

Approach

- Common approach to sensitivity and robustness is looking at the posterior predictive distribution of the model
  - This involves simulating data from the fitted model for comparison
- Typically we would want to graphically compare the data to the replicated data from the posterior
- Often this will include summary statistics

From: Brandt 2011

How is this done?

- Compute posteriors with different **priors** and compare the results
- Compare posteriors with the same functional prior, but different **hyperparameters**
  - i.e., coded differently
- Check that the model responds as expected to different priors
- Look at mixtures or perturbed priors
- Check for outliers
- Compute and compare posterior predictive distributions to the data

From: Brandt 2011

Posterior Predictive Distribution

- Basic idea is from Rubin (1984): if your model fits well then simulated data from the model should
  - be similar to the original data
  - summarize the main features of the original data
  - be insensitive and robust
- Note that this combines sensitivity and robustness checking and is easy!

From: Brandt 2011

Posterior Prediction

- **Prior predictive distribution** is the **pdf** of a new value before seeing the data:

$$p(x_{new}) = \int_{\Theta} p(x_{new}, \theta) d\theta = \int_{\Theta} p(x_{new} \mid \theta) p(\theta) d\theta$$

- So we can condition this on the data to get the **posterior predictive distribution**:

$$p(x_{new} \mid x) = \int_{\Theta} p(x_{new}, \theta \mid x) d\theta$$

$$= \int_{\Theta} \frac{p(x_{new}, \theta \mid x)}{p(\theta \mid x)} p(\theta \mid x) d\theta$$

$$= \int_{\Theta} p(x_{new} \mid \theta, x) p(\theta \mid x) d\theta$$

From: Brandt 2011

Posterior Prediction

- **Posterior predictive distribution**:

$$p(x_{new} \mid x) = \int_{\Theta} p(x_{new} \mid \theta, x)\, p(\theta \mid x)\, d\theta$$

PDF new observations

Likelihood of x

Uncertainty in □

- So the posterior predictive distribution is the product of
  - the new observations **pdf**
  - the likelihood
  - integration over the uncertainty in □
- depends only on the observed data!

From: Brandt 2011

How to do posterior predictive checks

1. Fit your Bayesian model
2. Use the posterior parameter estimates to simulate new dataset values
3. Compare the replicated / simulated data to the (in or out of) samples inferential quantities to see how well they compare
4. Often graphical and summary statistics comparisons will be useful for this

From: Brandt 2011

Posterior prediction in practice

- Suppose you need to predict an outcome y for a set of predictors X:
  - Estimate the parameters in $f(y \mid X, \beta)$.
  - Fix X
  - Simulate β^i for $i \in 1,\ldots,m$.
  - Generate samples from $f(y^i \mid X, \beta^i)$.
  - Compare the distributions of y and the summary of y^i over the m values
- We will do an example of prediction this afternoon

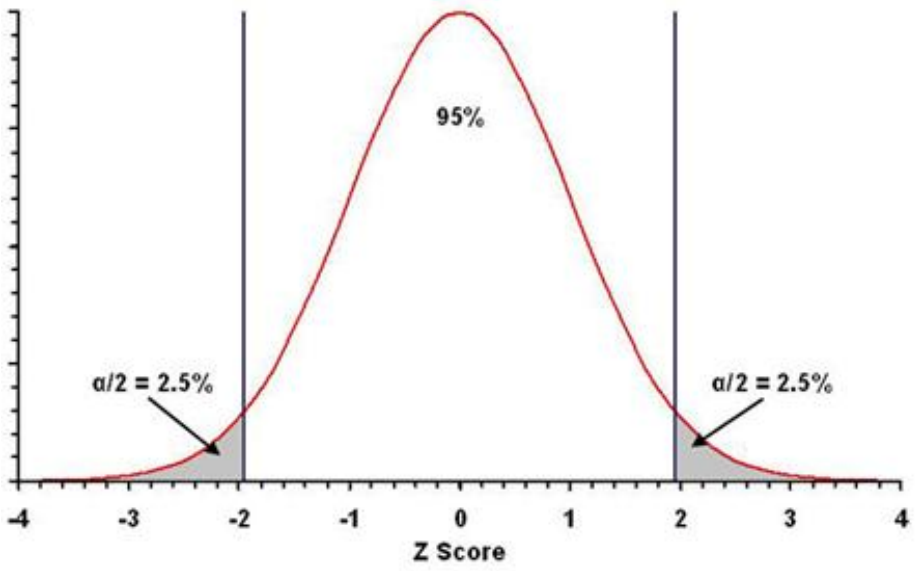
From: Brandt 2011

Today's lecture

- Model checking
  - Prediction
  - Robustness
- Hypothesis testing
  - Bayes Factors
  - BIC
- Non-Gaussian responses
  - Count variables
  - Binary variables

Brainstorm

- What are some hypotheses you would like to test as part of your analyses?
- Can you formulate them as Bayesian models?
- Have you tested them using null hypothesis testing?



What do we do with Bayesian results?

- Answer questions about hypotheses
- Weigh the evidence for model M_i
- Determine sensitivity of inferences to priors

From: Brandt 2011

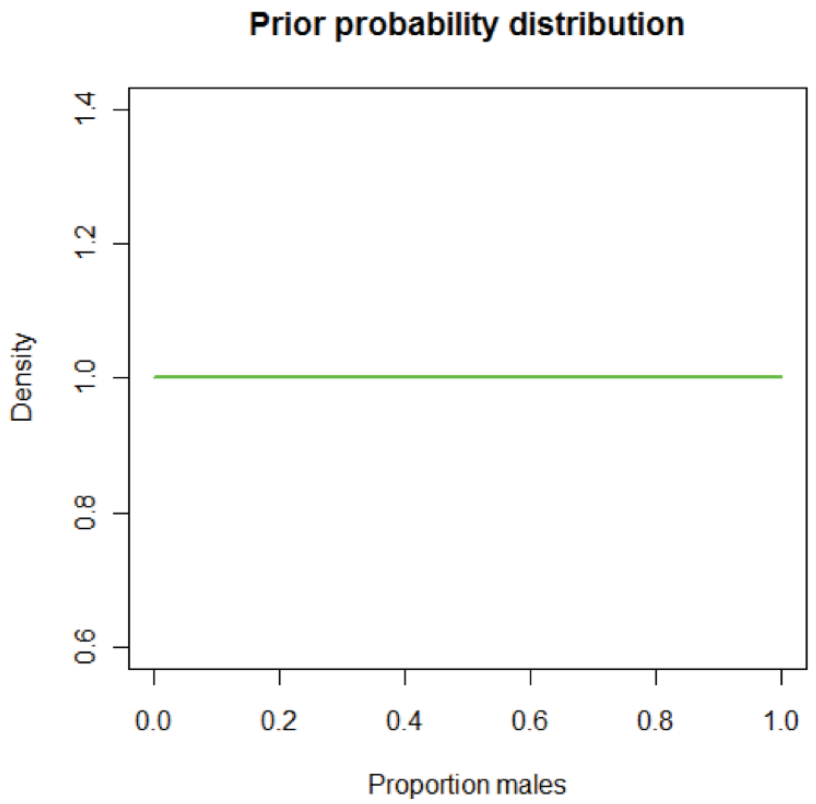
Example Bayesian hypothesis

- Study of the sex ratio of the communal--□ving bee, (Paxton and Tengo, 1996, *J. Insect. Behav.)*
- What is the proportion of males in the reproductive adults emerging from colonies?

From: Schluter

Bayesian estimation of a proportion

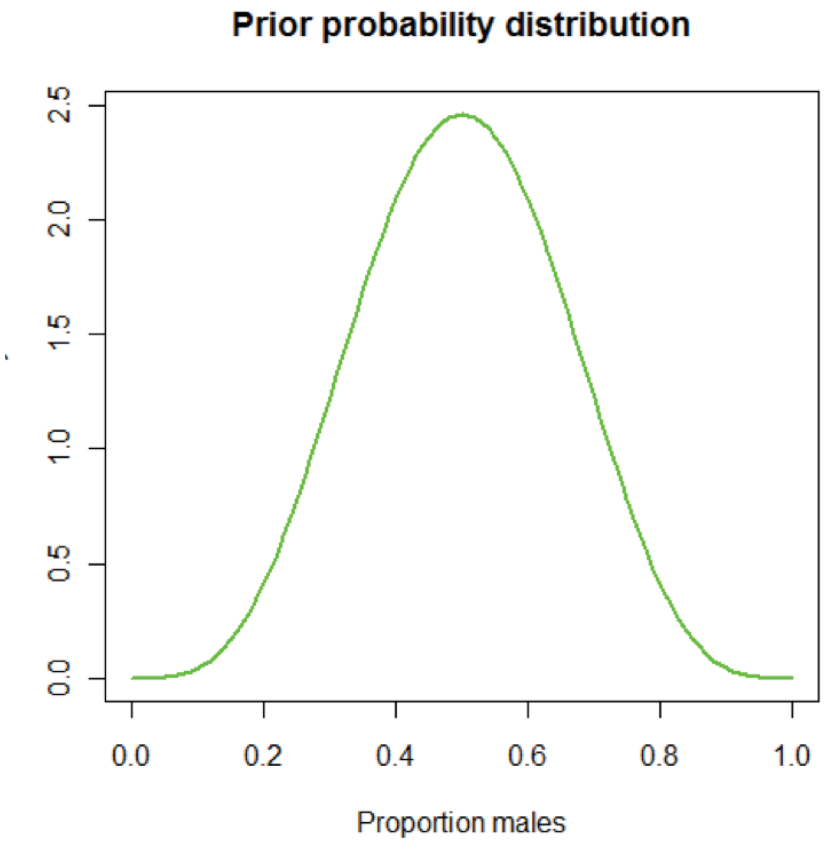
- To begin, we need to come up with a prior probability distribution for the proportion
- Case 1: an "noninformative" prior: expression of total ignorance



From: Schluter

Case 2

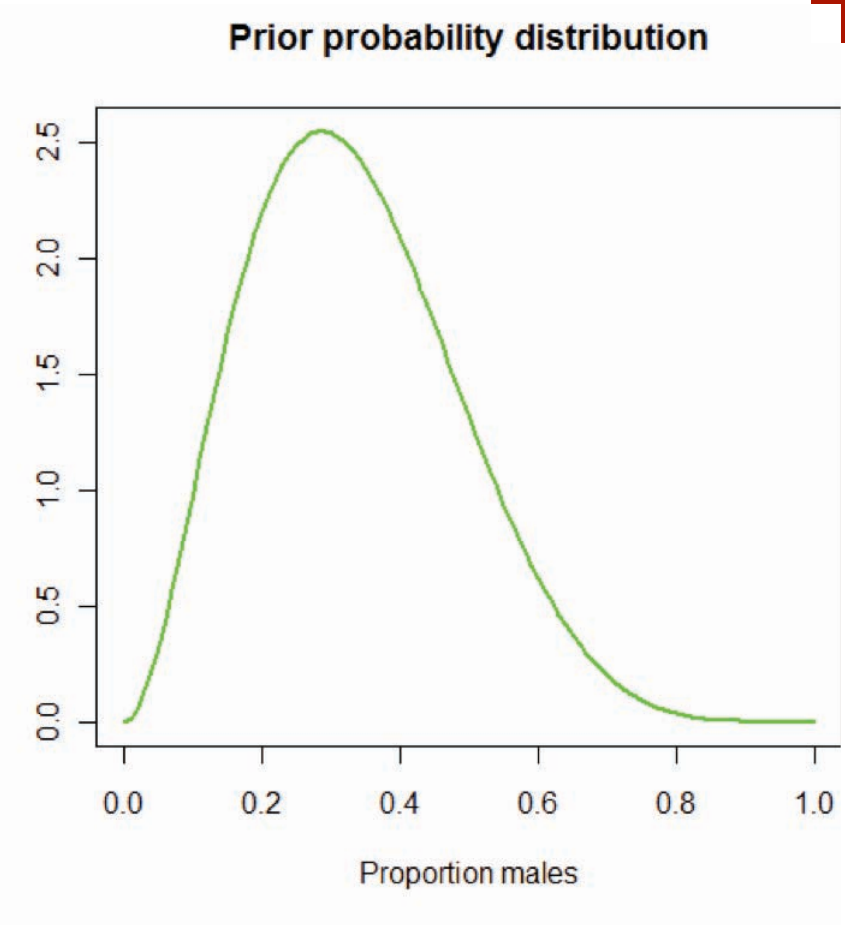
- Most species have a sex ratio close to 50:50, and this is predicted by simple sex-ratio theory
- The following prior attempts to capture this previous information (this is really what priors are for)



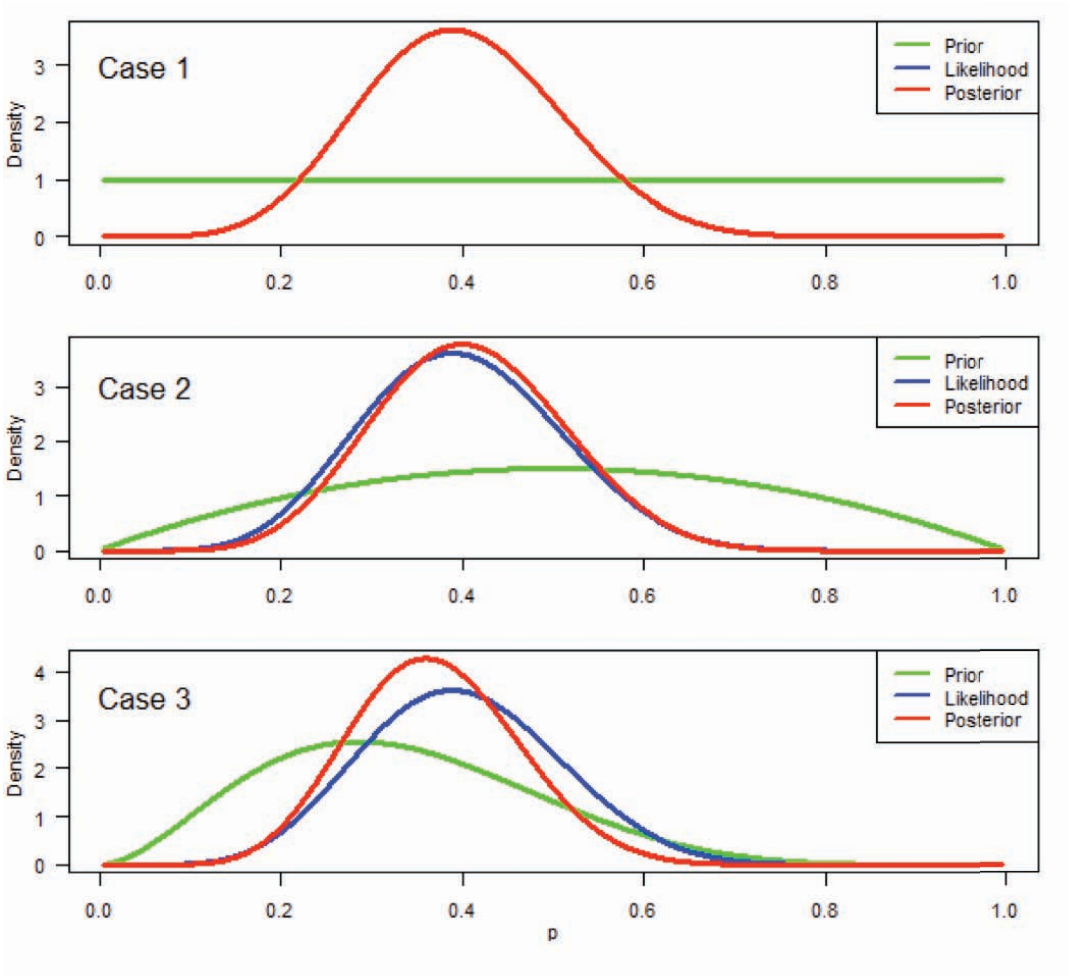
From: Schluter

Case 3

- Then again, female-biased sex ratios do exist in nature, more than male-biased sex ratios, especially in bees and other hymenoptera
- The following prior attempts to capture this previous information



Data: From day 148 at nest S31: 7 males, 11 females, $p.hat_{MLE} = 0.39$



Bayesian estimation of a proportion: results

- The estimate having maximum posterior probability depends on the prior probability distribution for the estimate
- Potential source of controversy: The prior is partly subjective. Different researchers may use different priors, hence obtain different estimates with the same data
- To resolve this we might all agree to use "noninformative" priors. But this prevents us from incorporating prior information, which is regarded as one of the strengths of the Bayesian approach
- Conflict can be resolved if we base the prior on a survey of preexisting evidence (lot of work). Would you agree with this?

Describing distributions

- If probability is subjective, then we want to describe posterior **distribution**, not **single** points under them
- Point hypothesis makes little sense outside a strict belief in long-run frequencies
- Need to rethink some of how we use hypothesis testing

From: Brandt 2011

Frequentist hypothesis testing

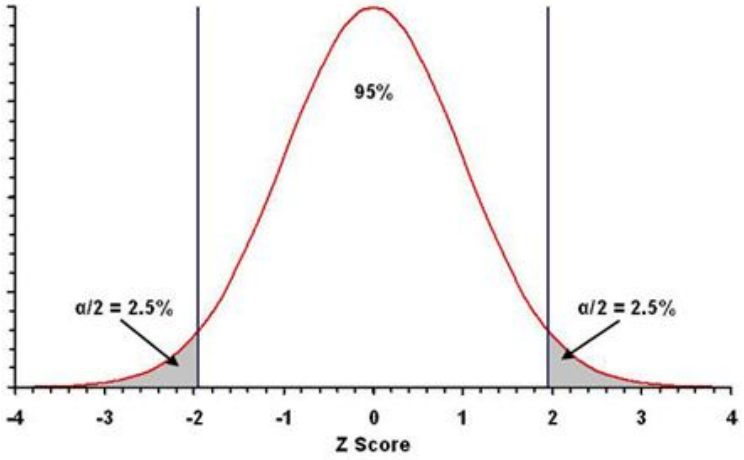
- Suppose we have $f(x \mid \theta)$ based on $X_1, X_2, \ldots X_n$. Then
  - One sided test:
    - $H_0: \theta \leq 0, H_1 > 0$
    - Two sided test:
    - $H_0: \theta = 0, H_1 \neq 0$
- with a *p*–value for a test statistic T of

$$p(x) = p(T(x) \geq T \mid \theta = 0) = \int_T^{\infty} f(t \mid \theta = 0) dt$$

From: Brandt 2011

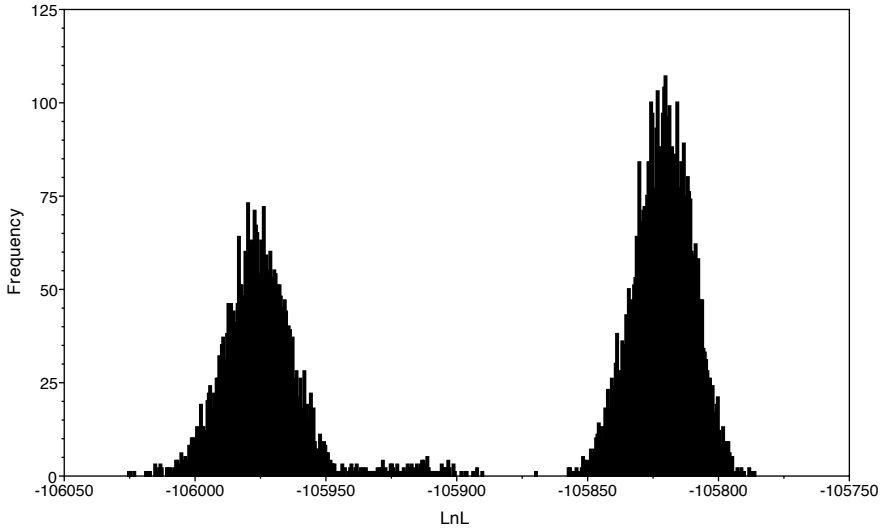
Issues with null hypothesis testing

- How do we choose the size of the test, α?
- Does not really measure Type I error
- Translate complex processes into cartoon null hypotheses
  - E.g., assume coefficients take value 0 or 1



Bayesian methods directly estimate parameters

- The posterior probability of a hypothesis is its probability given the data and the prior
  - Can depend strongly on the prior
- E.g., the probability that the value in the middle is the true value given the data, the model, and the prior is close to 0



Model selection

- Model selection: the problem of deciding the best candidate model fitted to data
- Requires a criterion to compare models, and a strategy for finding the best model
- One Bayesian approach uses Bayes factors
- Another option: BIC (Bayesian Information Criterion) as the criterion to select best model

Bayes factors

- Dominate hypothesis evaluation in Bayesian analysis
- Can be applied to non-nested hypotheses
  - Nested = one hypothesis simpler version of the other
- Allow for multiple comparisons without adjustments for multiple tests
- Compares prior odds to the posterior odds of a hypothesis being true

From: Brandt 2011

Definition

- Suppose we have two models that have different parameters:
$$M_1 : f_1(x \mid \theta_1) \quad M2 : f_2(x \mid \theta_2).$$
- These can be nested, or non-nested
- Assume we have a prior for each parameter vector, $p(\theta_i)$
- We can then find the prior probabilities for the two models, $p(M_1)$ and $p(M_2)$

From: Brandt 2011

To compare the evidence for one model against another

- for model 1 over model 2 is

Likelihood of data conditional on M_1!!!

$$B(x) = \frac{Pr(x \mid M_1)}{Pr(x \mid M_2)}$$

- No inherent scale to the ratio.
- Often computed in logarithms for numerical stability
- Can be sensitive to the specification of priors (which is good)!

From: Brandt 2011

Jeffreys Bayes factor scale

- $B(x) \geq 1$ model 1 supported
- $1 > B(x) \geq 10^{-1/2}$ minimal evidence against model 1
- $10^{-1/2} > B(x) \geq 10^{-1}$ substantial evidence against model 1
- $10^{-1} > B(x) \geq 10^{-2}$ strong evidence against model 1
- $10^{-2} > B(x)$ decisive evidence against model 1

From: Brandt 2011

Issues with Bayes factors

- Need to be able to compute the integration constant:

$$\int_{\theta_1} f_1(x \mid \theta_1) p(\theta_1) d\theta_1$$

  - which can be non-trivial
- Does not work for improper priors: the above integral is undefined
- Need to assume that both models can be wrong (Gelman and Rubin 1995)
- Can use a Bayesian information criteria (BIC)
  - This is independent of priors
  - This is a rough approximation to *log(B(x))*

From: Brandt 2011

Bayes information criterion BIC

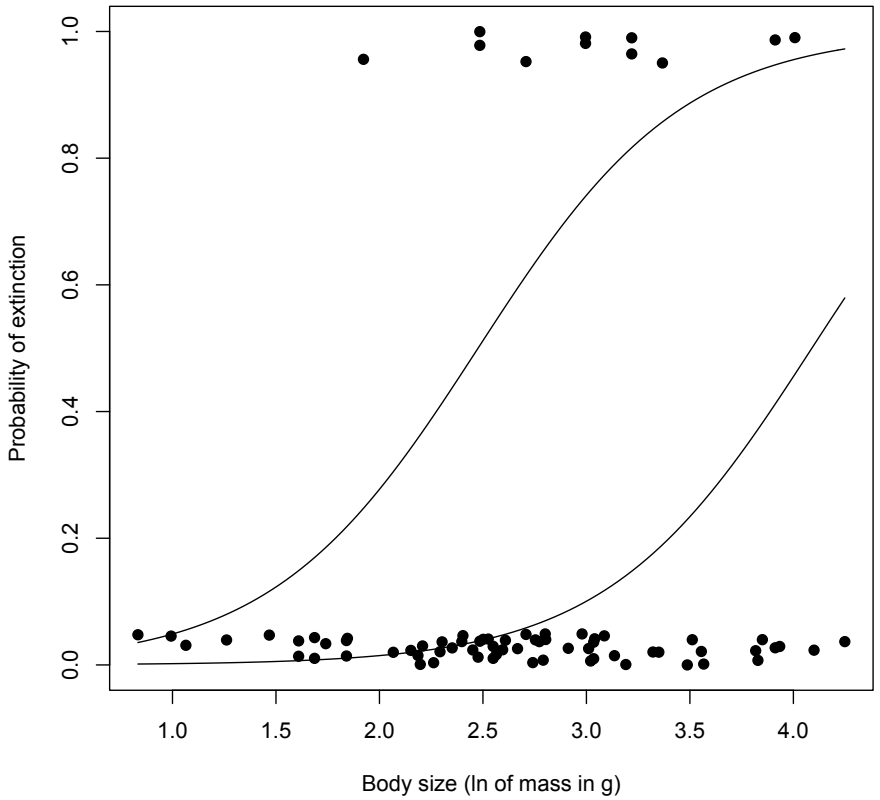
- Derived from a wholly different theory, but yields a formula similar to that of AIC (Akaike information criterion)
- It assumes that the "true model" is one of the models included among the candidates
- The approach has a tendency to pick a simpler model than that from AIC ("penalty" is more severe)
  - AIC = $-2 \ln L(\textit{fitted}\ \text{model} \mid \text{data}) + k \log(n)$
  - BIC = $-2 \ln L(\textit{fitted}\ \text{model} \mid \text{data}) + 2k$
    - k is the number of parameters estimated in the model (including intercept and σ^2), n is the sample size

Today's lecture

- Model checking
  - Prediction
  - Robustness
- Hypothesis testing
  - Bayes Factors
  - BIC
- Non-Gaussian responses
  - Count variables
  - Binary variables

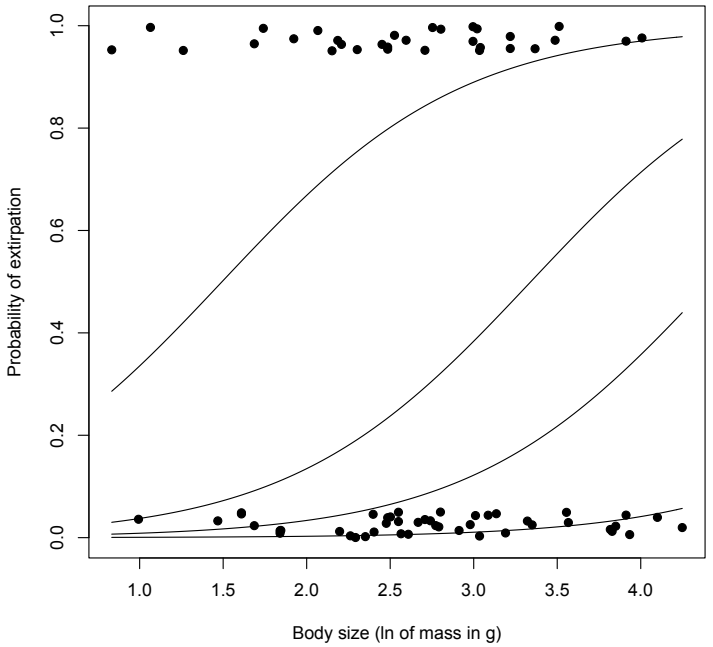
What if your response isn't normal?

- Often we want to model variables that are not normal or even continuous
- This is particularly important in the spatial context
  - E.g., presence or absence of a particular habitat type
  - Number of individuals of rare species detected



Think, pair, share

- Talk to your neighbor about hypotheses you discussed earlier
- How many involve continuous responses?
- How many involve presence/absence or count responses?
- What are they?



Spatial generalized linear models

- Often data sets preclude Gaussian modeling: $Y(\mathbf{s})$ may not even be continuous
- Example: $Y(\mathbf{s})$ is a **binary** or **count** variable
  - species presence or absence at location **s**
  - species abundance from count at location **s** continuous forest
  - variable is high or low at location **s**
- Replace Gaussian likelihood by exponential family member

From: Banerjee & Finley 2013

How to model

- **First stage**: $Y(\mathbf{s}_i)$ are conditionally independent given β and $w(\mathbf{s}_i)$, so $f(y(\mathrm{s}_i) \mid \beta, w(\mathrm{s}_i), \gamma)$ equals

$$h(y(\mathrm{s}_i), \gamma) \exp(\gamma[y(\mathrm{s}_i)\eta(\mathrm{s}_i) - \psi(\eta(\mathrm{s}_i))])$$

- Where $g(E(Y(\mathrm{s}_i))) = \eta(\mathrm{s}_i) = \mathrm{x}^T(\mathrm{s}_i)\beta + w(\mathrm{s}_i)$ is the canonical link function and γ is a dispersion parameter

- **Second stage**: Model $w(\mathbf{s})$ as a Gaussian process:

$$\mathrm{w} \sim N(0, \sigma^2 R(\phi))$$

- **Third stage**: Priors and hyperpriors

- No process for $Y(\mathbf{s})$, only a valid joint distribution

- Not sensible to add a pure error term $\varepsilon(\mathbf{s})$

From: Banerjee & Finley 2013

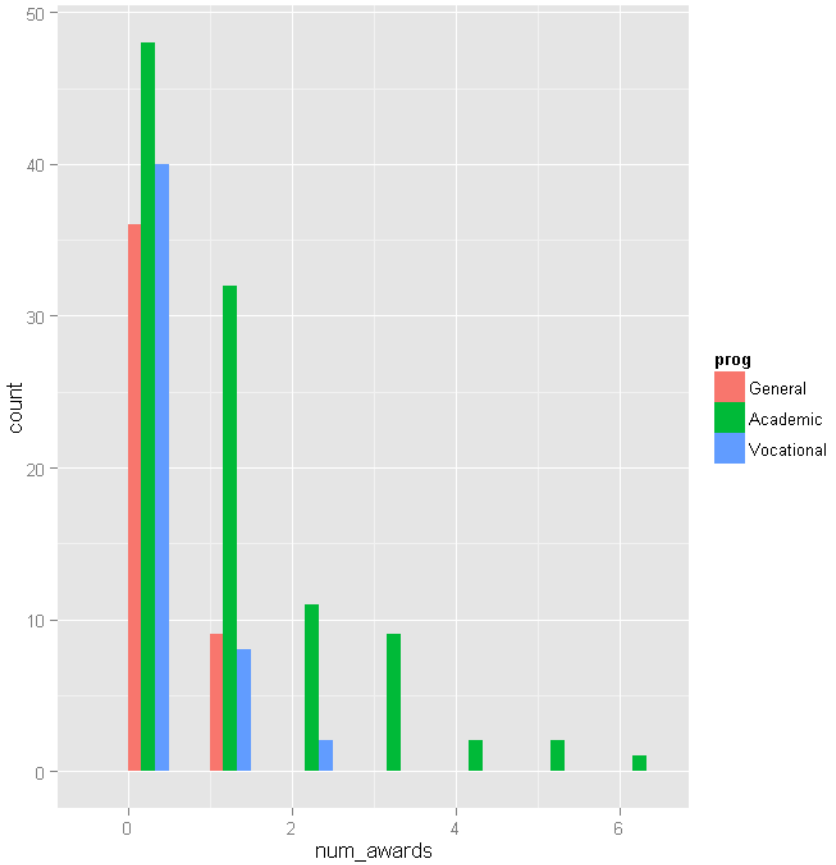
How does this work

- We are modeling with **spatial random effects**
- Introducing these in the transformed mean encourages means of spatial variables at proximate locations to be close to each other
- Marginal spatial dependence is induced between, e.g., Y(**s**) and Y (**s**'), but observed Y(**s**) and Y (**s**') need not be close to each other
- **Second stage** spatial modeling is attractive for spatial explanation in the **mean**
- **First stage** spatial modeling more appropriate to encourage proximate observations to be close

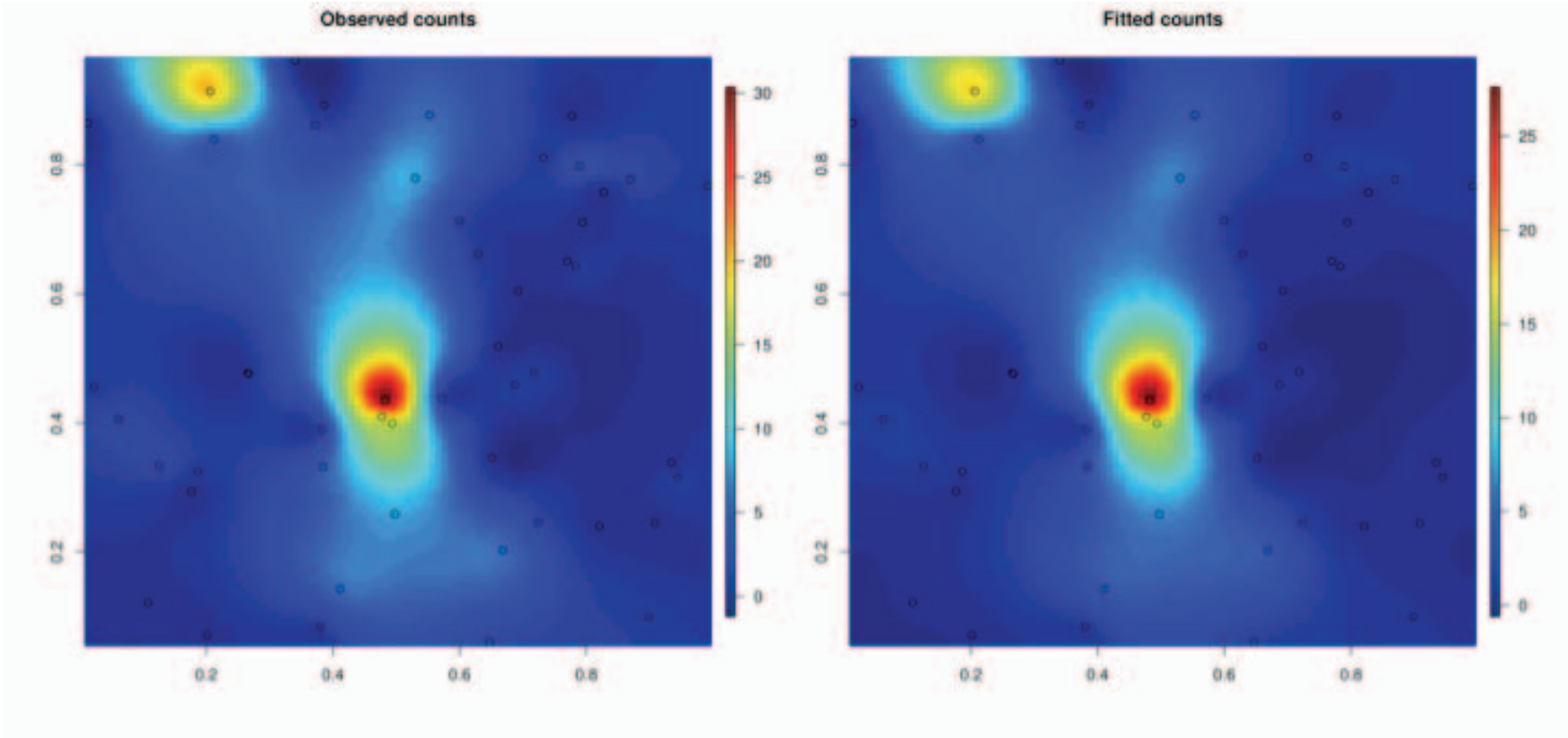
From: Banerjee & Finley 2013

Counts: Poisson regression

- Used to model dependent count variables
- Examples:
  - The number of persons killed by mule or horse kicks in the Prussian army per year
  - The number of people in line in front of you at the grocery store
  - The number of awards earned by students at one high school



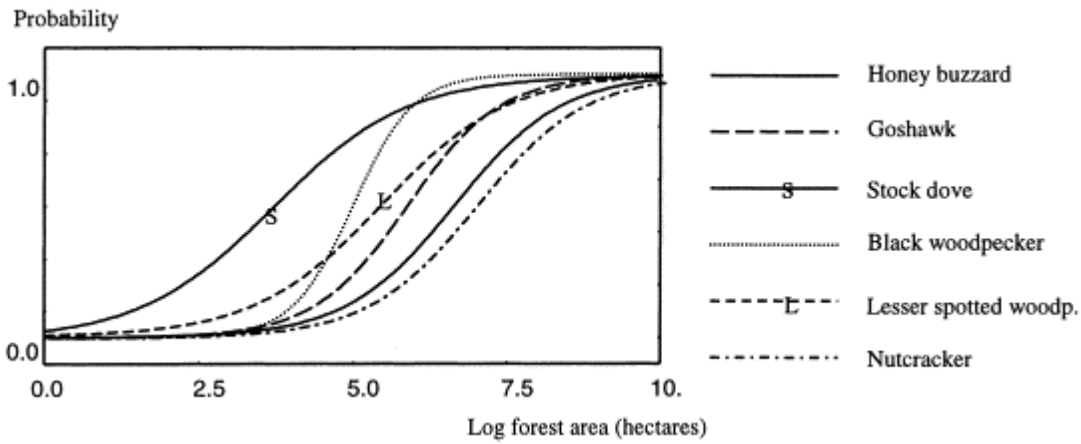
An example using spBayes (spGLM)



From: Banerjee & Finley 2013

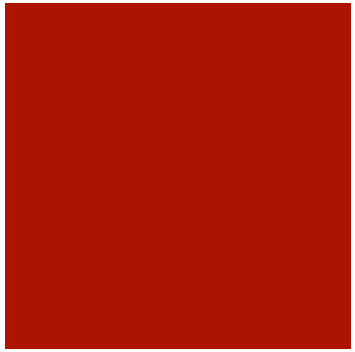
Binary/binomial: logistic regression

- Used to model dependent binary variables
- Examples:
  - Threatened status in bats
  - Having or not having a disease
  - Presence/absence of forest, a species, or soil type



Mortberg et al. 2000

Most landscapes will be modeled this way



- Objective is to make pixel-level prediction of forest/non-forest across the domain
- Data: Observations are from 500 georeferenced USDA Forest Service Forest Inventory and Analysis (FIA) inventory plots within a 32 km radius circle in MN, USA
- The response $Y(\mathbf{s})$ is a binary variable, with
  - $Y(\mathbf{s}) = 0$ if inventory plot is forested
  - $Y(\mathbf{s}) = 1$ if inventory plot is not forested
- Observed covariates include the coinciding pixel values for 3 dates of 30 × 30 m resolution Landsat imagery

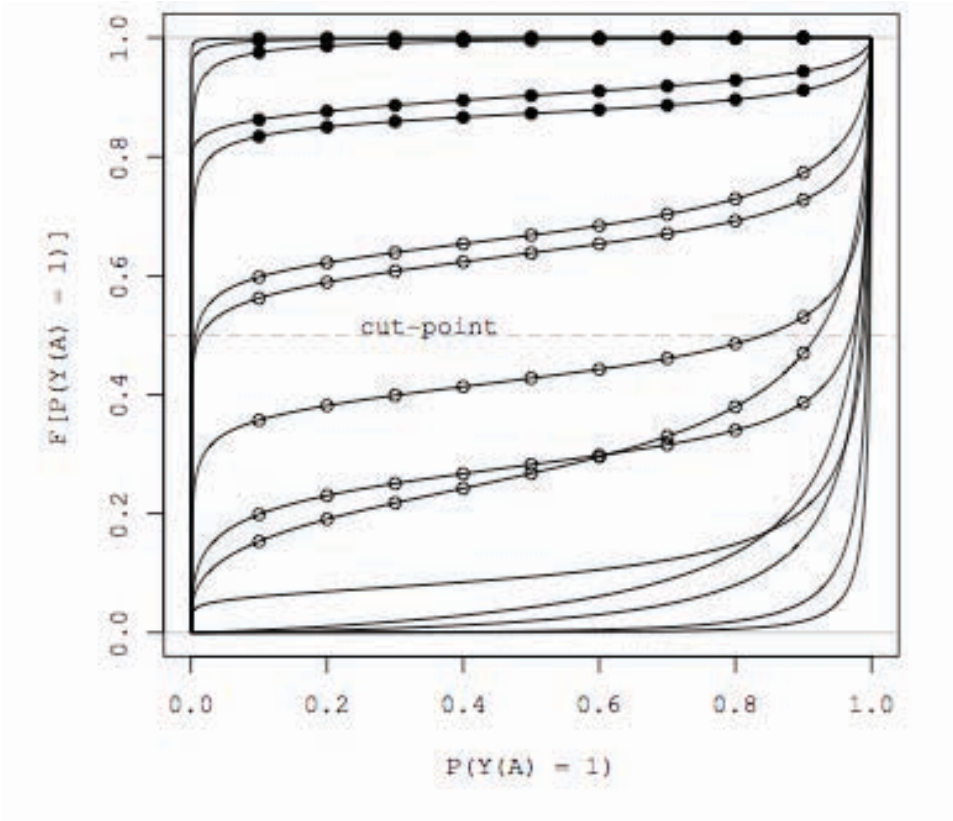
From: Banerjee & Finley 2013

Binary spatial regression: forest/non-forest

- We fit a generalized linear model where
  - $Y(\mathbf{s}_i) \sim Bernoulli(p(\mathbf{s}_i))$, $logit(p(\mathbf{s}_i)) = \mathbf{x}^T(\mathbf{s}_i)\beta + w(\mathbf{s}_i)$
  - Assume vague *flat* for β, a Uniform(3/32, 3/0.5) prior for ϕ, and an inverse-Gamma(2, ·) prior for σ^2

From: Banerjee & Finley 2013

Classification of 15 20 × 20 pixel areas (based on visual inspection of imagery)



From: Banerjee & Finley 2013

Evaluating a spatial logistic regression

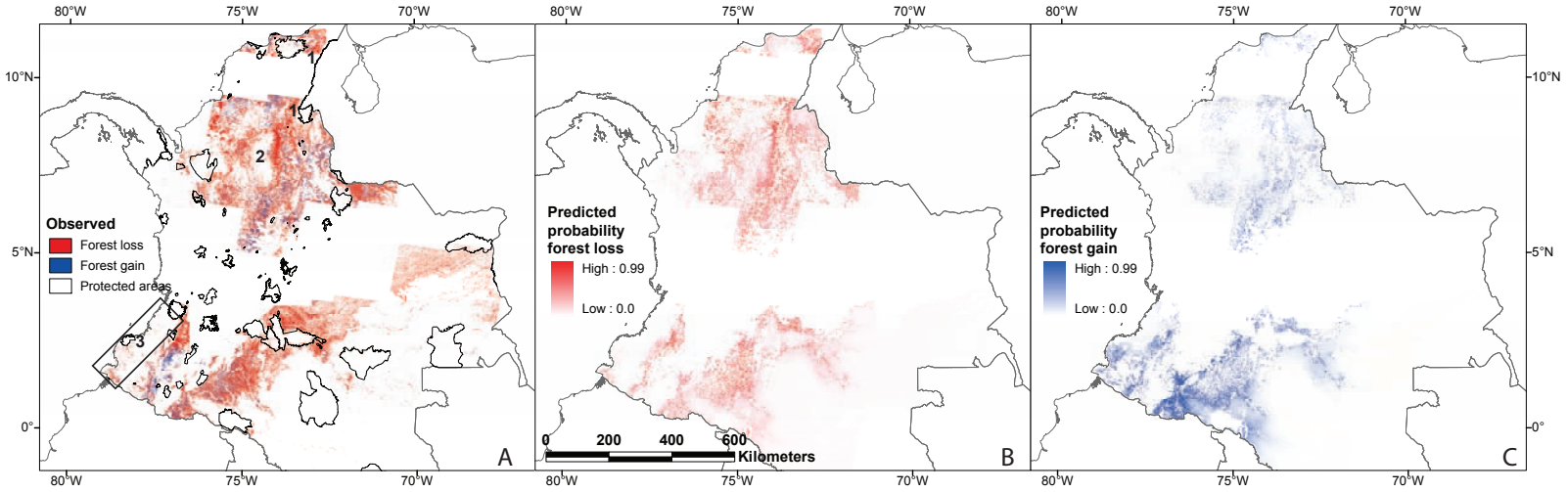
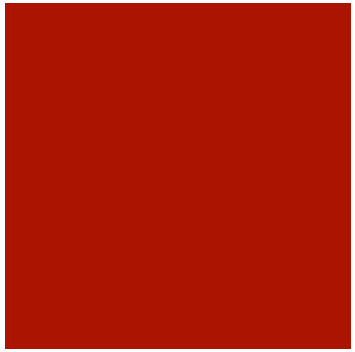


Figure 2



- THE END