Introduction to Generalized Linear Models

FW8051 Statistics for Ecologists

Department of Fisheries, Wildlife and Conservation Biology

Learning Objectives

- Understand the role of random variables and common statistical distributions in formulating modern statistical regression models
- Be able to fit appropriate models to count data and binary data (yes/no, presence/absence) in both R and JAGS
- Be able to evaluate model goodness-of-fit
- Be able to describe a variety of statistical models and their assumptions using equations and text and match parameters in these equations to estimates in computer output.

Outline

- Introduction to generalized linear models (today)
- Models for count data (Poisson and Negative Binomial regression)
- Models for Binary data (logistic regression)
- Models for data with lots of zeros

Linear Regression

Often written in terms of "signal + error":

$$y_i = \underbrace{\beta_0 + x_i\beta_1}_{\text{Signal}} + \underbrace{\epsilon_i}_{\text{error}}, \text{ with}$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Possible because the Normal distribution has separate parameters that describe:

- mean: $E[Y_i|X_i] = \mu_i = \beta_0 + x_i\beta_1$
- variance: $Var[Y_i|X_i] = \sigma^2$

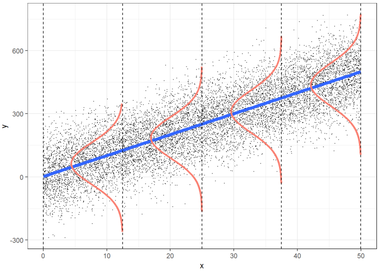
Remember: for Poisson, Binomial distributions, the variance is a function of the mean.

Linear Regression

$$Y_i|X_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 X_{1,i} + \ldots \beta_p X_{p,i}$$

This description highlights:

1. The distribution of Y_i depends on a set of predictor variables X_i
2. The distribution of the response variable, conditional on predictor variables is Normal
3. The mean of the Normal distribution depends on predictor variables (X_1 through X_p) and regression coefficients (the β_1 through β_p)
4. The variance is constant and given by σ^2.



From Broadening Your Statistical Horizons:
https://bookdown.org/roback/bookdown-bysh/

Linear Regression = General Linear Model

Sometimes referred to as: General Linear Model

- t-test (categorical predictor with 2 categories)
- ANOVA (categorical predictor with $>$ 2 categories)
- ANCOVA (continuous and categorical predictor, no interaction so common slope)
- Continuous and categorical variables, with possible interactions

Generalized Linear Models

Generalized linear models further unifies several different regression models:

- General linear model
- Logistic regression
- Poisson regression
- ...

Rather elegant general theory developed for exponential family of distributions

Generalized Linear Models (glm)

Systematic component: $g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$

Some transformation of the the mean, $g(\mu_i)$, results in a linear model.

- $g()$ is called the link function
- $\eta_i = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$ is called the linear predictor.
- $\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_1 + \ldots \beta_p x_p)$

Random component: $Y_i|X_i \sim f(y_i|x_i), i = 1, \ldots, n$

- $f(y_i|x_i)$ is in the exponential family (includes normal, Poisson, binomial, gamma, inverse Gaussian)
- $f(y_i|x_i)$ describes unmodeled variation about $\mu_i = E[Y_i|X_i]$

Generalized Linear Models (glm)

Linear Regression:

- $f(y_i|x_i) = N(\mu_i, \sigma^2)$
- $E[Y_i|X_i] = \mu_i = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$
- $g(\mu_i) = \eta_i = \mu_i$, the identity link
- $\mu_i = g^{-1}(\eta_i) = \eta_i = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$

Poisson regression:

- $f(y_i|x_i) \sim Poisson(\lambda_i)$
- $E[Y_i|X_i] = \mu_i = \lambda_i$
- $g(\mu_i) = \eta_i = log(\lambda_i) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$
- $\mu_i = g^{-1}(\eta_i) = exp(\eta_i) = exp(\beta_0 + \beta_1 x_1 + \ldots \beta_p x_p)$

Other GLMs

Logistic regression:

- $f(y_i|x_i) \sim \text{Bernoulli}(p_i)$
- $E[Y_i|X_i] = p_i$
- $g(\mu_i) = \eta_i = logit(p_i) = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$
- $\mu_i = g^{-1}(\eta_i) = \frac{\exp^{\eta_i}}{1+\exp^{\eta_i}} = \frac{\exp^{\beta_0+\beta_1 x_1+\ldots\beta_p x_p}}{1+\exp^{\beta_0+\beta_1 x_1+\ldots\beta_p x_p}}$

Link functions and sample space

Link functions allow the "structural component" $(\beta_0 + \beta_1 x_1 + \ldots \beta_p x_p)$ to live on $(-\infty, \infty)$ while keeping the μ_i consistent with the range of the response variable.

Poisson (counts) = $0, 1, 2, \ldots, \infty$

- $g(\mu_i) = \eta_i = log(\lambda_i) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$, range = $(-\infty, \infty)$
- $\mu_i = exp(\beta_0 + \beta_1 x_1 + \ldots \beta_p x_p)$, range = $[0, \infty]$

Logistic regression:

- $g(\mu_i) = \eta_i = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$, range = $(-\infty, \infty)$
- $\mu_i = g^{-1}(\eta_i) = \frac{\exp^{\eta_i}}{1+\exp^{\eta_i}} = \frac{\exp^{\beta_0+\beta_1 x_1+\ldots\beta_p x_p}}{1+\exp^{\beta_0+\beta_1 x_1+\ldots\beta_p x_p}}$, range = $(0, 1)$

Probit regression

Probit regression model:

- $f(y_i|x_i) \sim \text{Bernoulli}(p_i)$
- $E[Y_i|X_i] = p_i$
- $g(\mu_i) = \eta_i = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$
- $\mu_i = g^{-1}(\eta_i) = \Phi(\beta_0 + \beta_1 x_1 + \ldots \beta_p x_p)$

where Φ is the cumulative probability distribution function, $F(X) = P(X \leq x)$, for a standard normal distribution (goes between 0 and 1).

Probit regression model:

- $g(\mu_i) = \Phi^{-1}(p_i)$ (think `qnorm`), range = $(-\infty, \infty)$
- $\mu_i = g^{-1}(\eta_i) = \Phi(\beta_0 + \beta_1 x_1 + \ldots \beta_p x_p)$, (think `pnorm`), range = $(0, 1)$

By Gary Noon - Flickr, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=4077294

- X_i = amount of grassland cover
- $Y_i|X_i \sim Poisson(\lambda_i)$
- $log(\lambda_i) = \beta_0 + \beta_1 X_i$

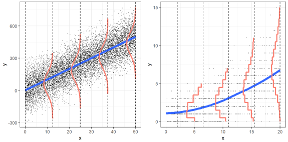
Because the mean of the Poisson distribution is λ:

- $E[Y_i|X_i] = \lambda_i = \exp(\beta_0 + \beta_1 X_i) = exp(\beta_0)exp(\beta_1 X_i)$
- The mean number of pheasants increases by a factor of $exp(\beta_1)$ as we increase X_i by 1 unit.

Assumptions

1. **Poisson Response**: The response variable is a count per unit of time or space, described by a Poisson distribution.
2. **Independence**: The observations must be independent of one another.
3. **Mean=Variance**: By definition, the mean of a Poisson random variable must be equal to its variance.
4. **Linearity**: The log of the mean rate, $\log(\lambda)$, must be a linear function of x.

Visually



- small values of λ are associated with skewed distributions
- as λ increases, the variance increases and the response looks more Normal
- $log(E[Y_i|X_i]) = \beta_0 + \beta_1 X_i$ so $E[Y_i|X_i] = exp(\beta_0 + \beta_1 X_i)$

From Broadening Your Statistical Horizons: https://bookdown.org/roback/bookdown-bysh/

Next Steps

- Understand how Maximum Likelihood is used to fit modern statistical regression models (`glm`)
- Be able to fit regression models appropriate for count data in R and JAGS
  - Poisson regression models
  - Quasi-Poisson (R only)
  - Negative Binomial regression
- Interpret estimated coefficients and describe their uncertainty using confidence and credible intervals
- Use simple tools to assess model fit
  - Residuals (deviance and Pearson)
  - Goodness-of-fit tests
- Use deviances and AIC to compare models.
- Use an offset to model rates and densities, accounting for variable survey effort
- Be able to describe statistical models and their assumptions using equations and text and match parameters in these equations to estimates in computer output.