Dichotomous Predicted Variable

• This section focuses on dichotomous predicted variables: such as whether a basketball player will get a hit or if a bird will be located in a spatial grid.

• Traditionally, these types of methods are generally implemented with logistic regression.

• The model can be written as:

\[ y \sim \text{Bernoulli}(\mu) \]

\[ \mu = \logistic(\beta_0 + \beta_1 x_1 + \beta_2 x_2), \]

where the logistic function is \( \logistic(x) = 1/(1 + \exp(-x)) \).

• Q: What are the three components of a GLM and what are they in this specific setting?
  1. Sampling Model:
  2. Linear Combination of Predictors:
  3. (inverse) Link function:

• To fit this in a Bayesian framework, we need to specify priors. What parameters require priors and what distributions would be reasonable?
Lab Exercise

We will revisit the Swiss birds dataset for this lab to construct a logistic regression model for presence of the Willow Tit.

```r
swiss.birds <- read.csv('http://www.math.montana.edu/ahoegh/teaching/stat491/data/willowtit2013.csv')
kable(head(swiss.birds))
```

<table>
<thead>
<tr>
<th>siteID</th>
<th>elev</th>
<th>rlength</th>
<th>forest</th>
<th>birds</th>
<th>searchDuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q001</td>
<td>450</td>
<td>6.4</td>
<td>3</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td>Q002</td>
<td>450</td>
<td>5.5</td>
<td>21</td>
<td>0</td>
<td>190</td>
</tr>
<tr>
<td>Q003</td>
<td>1050</td>
<td>4.3</td>
<td>32</td>
<td>1</td>
<td>150</td>
</tr>
<tr>
<td>Q004</td>
<td>950</td>
<td>4.5</td>
<td>9</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>Q005</td>
<td>1150</td>
<td>5.4</td>
<td>35</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Q006</td>
<td>550</td>
<td>3.6</td>
<td>2</td>
<td>0</td>
<td>115</td>
</tr>
</tbody>
</table>

This dataset contains 242 sites and 6 variables:
- siteID, a unique identifier for the site, some were not sampled during this period
- elev, mean elevation of the quadrant in meters
- rlength, the length of the route walked by the birdwatcher, in kilometers
- forest, percent forest cover
- birds, binary variable for whether a bird is observed, 1 = yes
- searchDuration, time birdwatcher spent searching the site, in minutes

1. (5 points) Model Specification

Clearly write out the model, using proper notation for the variables in bird dataset. You don’t need use all variables, but you should state which are included.

2. (5 points) Priors

Describe and justify the necessary priors for this model.

3. (5 points) Fit MCMC

Fit the JAGS code for this model. You will have to put this together following the specification in the previous examples, but the following statement can be used for the sampling model portion.

```r
model {
  for (i in 1:Ntotal) {
    y[i] ~ dbern(mu[i])
    mu[i] <- ilogit(beta0 + sum( beta[1:Nx] * x[i,1:Nx] ))
  }
  # priors inserted here
}
```

4. (5 points) Summarize inferences from model

Talk about the model and discuss which and how predictor variables influence the observation of a bird.
Interpretation of Logistic Regression Coefficients

Recall this model can be written as:

- When $x_i$ increases or decreases by 1 unit,

- The logit function for $\mu$ can be expressed as $\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

- In this setting, $\mu$ is the probability of $y = 1$, so

- This ratio:

Suppose the logistic regression for the Swiss birds had the following coefficients: $\beta_0 = -6$, $\beta_{\text{elev}} = .002$, and $\beta_{\text{forest}} = .06$. Compute the probability for observing a bird in a quadrant with: elevation = 1500 meters and forest cover of 60%.

- How do we interpret the meaning for the coefficient $\beta_{\text{forest}}$?

- The log-odds are different than a probability as in this case (conditional on the elevation) the probability increases to: $\frac{1}{1 + \exp(-(-6 + .002 \times 1500 + .06 \times 70))}$ =

- Note that an unit increase in the log-odds does not have a unit increase in the probability.
Count Predicted Variable

- Now reconsider the willow tit dataset and consider modeling not just the presence / absence of birds, but directly modeling the number of birds observed in each spatial region.

```r
data(swiss.birds)
kable(head(swiss.birds))
```

<table>
<thead>
<tr>
<th>siteID</th>
<th>elev</th>
<th>rlength</th>
<th>forest</th>
<th>bird.count</th>
<th>searchDuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q001</td>
<td>450</td>
<td>6.4</td>
<td>3</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td>Q002</td>
<td>450</td>
<td>5.5</td>
<td>21</td>
<td>0</td>
<td>190</td>
</tr>
<tr>
<td>Q003</td>
<td>1050</td>
<td>4.3</td>
<td>32</td>
<td>3</td>
<td>150</td>
</tr>
<tr>
<td>Q004</td>
<td>950</td>
<td>4.5</td>
<td>9</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>Q005</td>
<td>1150</td>
<td>5.4</td>
<td>35</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Q006</td>
<td>550</td>
<td>3.6</td>
<td>2</td>
<td>0</td>
<td>115</td>
</tr>
</tbody>
</table>

```r
ggplot(aes(bird.count), data=swiss.birds) + geom_bar() + ylab('Frequency') + xlab('Number of Birds Observed') + ggtitle('Bird Counts of Willow Tit')
```
1. Sampling Model: In general the Poisson model will be used as a sampling model for count data: \( y \mid \mu \).

- the mean and variance of the Poisson

- if this is not a reasonable assumption, the negative binomial distribution can be used

2. Linear Combination of Predictors:

3. Link Function: The support for count data is values greater than or equal to zero.
Homework Assignment

We will revisit the Swiss birds dataset for this lab to construct a Poisson regression model for abundance of the Willow Tit.

```r
# swiss.birds <- read.csv('http://math.montana.edu/ahoegh/teaching/stat491/data/willowtit2013_count.csv')
kable(head(swiss.birds))
```

<table>
<thead>
<tr>
<th>siteID</th>
<th>elev</th>
<th>rlength</th>
<th>forest</th>
<th>bird.count</th>
<th>searchDuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q001</td>
<td>450</td>
<td>6.4</td>
<td>3</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td>Q002</td>
<td>450</td>
<td>5.5</td>
<td>21</td>
<td>0</td>
<td>190</td>
</tr>
<tr>
<td>Q003</td>
<td>1050</td>
<td>4.3</td>
<td>32</td>
<td>3</td>
<td>150</td>
</tr>
<tr>
<td>Q004</td>
<td>950</td>
<td>4.5</td>
<td>9</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>Q005</td>
<td>1150</td>
<td>5.4</td>
<td>35</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Q006</td>
<td>550</td>
<td>3.6</td>
<td>2</td>
<td>0</td>
<td>115</td>
</tr>
</tbody>
</table>

This dataset contains 242 sites and 6 variables. The only difference from the lab is:
- bird.count, count variable for abundance of birds observed, rather than a binary outcome

1. (5 points) Model Specification

Clearly write out the model, using proper notation for the variables in bird dataset. You don’t need use all variables, but you should state which are included.

2. (5 points) Priors

Describe and justify the necessary priors for this model.

3. (5 points) Fit MCMC

Fit the JAGS code for this model. You will have to put this together following the specification in the previous examples, but the following statement can be used for the sampling model portion.

```jags
model {
  for (i in 1:Ntotal) {
    y[i] ~ dpois(mu[i])
    mu[i] <- exp(beta0 + sum( beta[1:Nx] * x[i,1:Nx] ))
  }
  # priors inserted here
}
```

4. (5 points) Summarize inferences from model

Talk about the model and discuss which and how predictor variables influence the observation of a bird. Furthermore, discuss the differences between this model and the logistic regression setting.