MTH225 Fall2016 Final Problem 2

Logistic regression is a special case of a class of techniques known as *generalized* linear models. Another member of this class is Poisson regression.

Poisson regression may be an appropriate choice when the data represents counts. In this problem, we have data from two sites on counts of gypsy moth eggmasses in randomly selected plots. One of the sites (site 2) was sprayed with Bt (bacillus thuringiensis) the previous year, and we are interested in determining whether there is a detectable difference in the number of eggmasses in the year following spraying. In addition, it is thought that soil moisture, as a proxy for relative wetness or dryness in that location, may influence eggmass production, so a measure of relative soil moisture is included as a covariate.

The data is in `MTH225_Spring2017_Final_Problem2.csv`. The variables are:

- `count` Count of eggmasses for this plot.
- `moisture` Relative soil moisture for this plot.
- `site` Which site the plot is in (1 or 2).

Use STAN to model the data as a sample from a Poisson distribution, with parameter λ given by the linear model

$$\lambda = \alpha_i + \beta \cdot x$$

An ordinary least squares (OLS) model can be written in terms of the conditional expectation of y given x as:

$$E(y|x) = X\beta$$

The Poisson regression model has the form

$$\ln(E(y|x)) = X\beta$$

STAN provides a convenience function called `poisson_log` that automatically applies the natural log function to $X\beta$, so the likelihood part of the model can be coded as:

The Stan model file `Poisson_regression1.stan` can be used as a template.

```
y ~ poisson_log(alpha+beta*x);
```

Where y represents the observed counts, α_i represents the mean for site i and x is the soil moisture.

Use the posterior draw to compare the counts in the sprayed (site 2) and unsprayed (site 1) sites.

- 2 points: Write R code to read the data and convert it to an R data frame.
- 1 point: Write the data block of a STAN model file that extracts the data from the R workspace.
- 1 point: Write the parameters block of a STAN model file that declares the paramter(s) of your model.
- 2 points: Write the model block of a STAN model file that specifies the priors and likelihood for your model.
- 1 point: Write R code to apply the `extract` function to the data structure output from the `stan` function.
- 1 point: Use the `extract()` function of the RSTAN package to obtain the values for the parameters from the posterior draw.
- 1 point: Compute 95% confidence intervals for the difference between the alpha parameters for sites 1 and 2, adjusted for soil moisture.
- 1 point: Using the posterior draw for α and β with the mean $\overline{x}$ of the moisture values, compute 95% credible intervals for the counts at site 1 and site2. For each of the 4,000 pairs of α and β values in the posterior draw, you can calculate the count when the moisture is $\overline{x}$ for site i as:

  $$n_{ij} = \exp(\alpha_{ij} + \overline{x} * \beta_j) \quad i = 1, 2 \quad j = 1, 2, \ldots, 4000$$

  To get the credible intervals, calculate the quantiles of the n_{1j} and n_{2j} arrays.

(10 points possible)