In-class exercise: Logistic regression

Names: (signatures only please, printed names will not be counted)

1.) 4.)

2.) 5.)

3.) 6.)

Overview

In this exercise we suppose we model the number of days with measurable rainfall in a month at the Amherst Massachusetts weather station using monthly records from 1900 through 1992 in `logistic1.csv`

We model the number of days with measurable rainfall in a month using a bionomial distribution with the number of trials n equal to the number of days in the month, and probability of success p that depends on time through a regression model.

$$y \sim \text{binomial(n,p)}$$

The probability of measurable rainfall is represented using a *logit* transform, which greatly improves the numerical stability of the regression model. The logit transform is:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

This function maps values in the interval $(0, 1)$ to the interval $(-\infty, \infty)$, and its inverse

$$\text{logit}^{-1}(z) = \frac{1}{1+e^{-z}}$$

maps $(-\infty, \infty)$ into $(0, 1)$. Our regression model is written in terms of z,

$$z_i = \beta_0 + \beta_1 x_i$$

STAN has a built-in likelihood function called `binomial_logit` that automatically does the transform from z to p for the binomial likelihood. The line that specifies this in the model statement would look like this:

```
y[i] ~ binomial_logit(n[i],beta0+beta1*x[i]);
```

It is generally a good idea to center the values of the independent variable x so that they has a mean of zero, and scale them so that the range is fairly compact.

A reasonable choice in this case is, for year y and month m,

$$x = \frac{12(y - 1947) + m}{279}$$

For this data, this formula produces values between -2 and $+2$. The x column contains values computed from the year and month column using this formula.

Instructions

As usual, start by bringing your copy of the `MTH225_Fall2016` archive up to date.

Open a command prompt or terminal window, and use the `cd` command to change to the `MTH225_Fall2016` subdirectory. Then type the command:

```
git pull origin master
```

The pull operation should download the following files:

- The R-knitr code: `MTH225-11_logistic_regression.Rnw`
- The data in Rdata format: `logistic1.csv`
- The STAN model file: `logistic1.stan`

Questions

Use the *Compile PDF* button to run the model, and use the output to answer the following questions:

1) What is the point estimate and 95% confidence interval for the slope parameter?

2) What is the point estimate and 95% confidence interval for the intercept for the month of April?

3) What is the point estimate and 95% confidence interval for the probability of rainfall on a day in January, 1900? In January 1947? In January 1990?

4) What is the median of, and 95% confidence interval for, the number of rainy days in January, 1900? In January, 1990?

5) What probability does the model give that the number of rainy days in January, 1990 is greater than or equal to the number of rainy days in 1900?