37457

Advanced Bayesian Methods

Data Pre-Processing for Bayesian Analyses

A New Species of Snake!

A zoologist discovers a new species of snake in a remote part of the Amazon. The snakes are very long and the zoologist measures 100 of them.

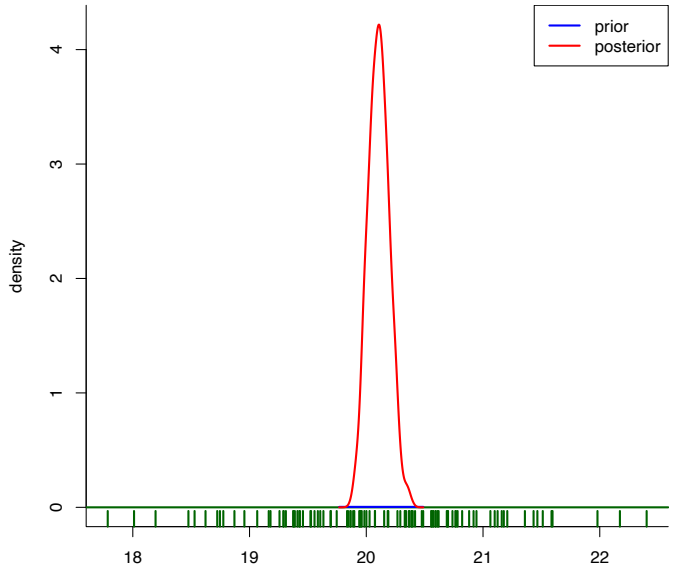
The data are sent back to her lab and a statistician fits the following Bayesian model in Stan:

$$x_i|\mu,\sigma^2 \overset{\text{ind.}}{\sim} N(\mu,\sigma^2)$$

$$\mu \sim N(0,100^2), \quad \sigma \sim \text{Half-Cauchy}(0,100^2)$$

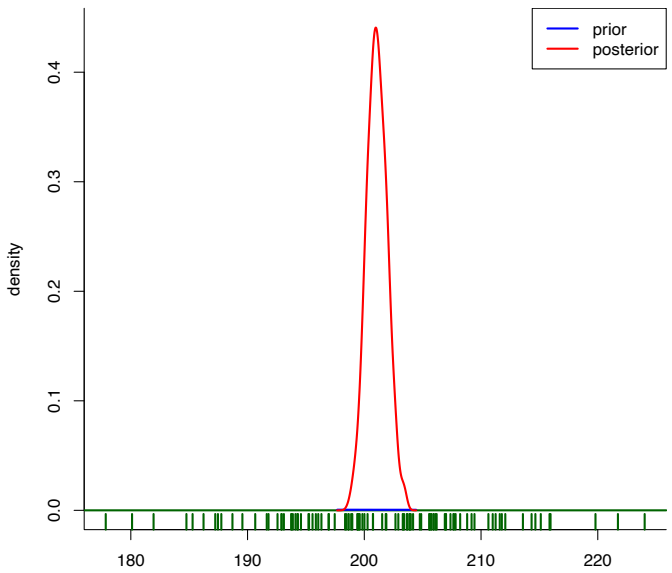
where $x_1,\ldots,x_{100}$ are the snake lengths in metres. The statistician wants to make Bayesian inference about:

$\mu =$ mean length of snake species.



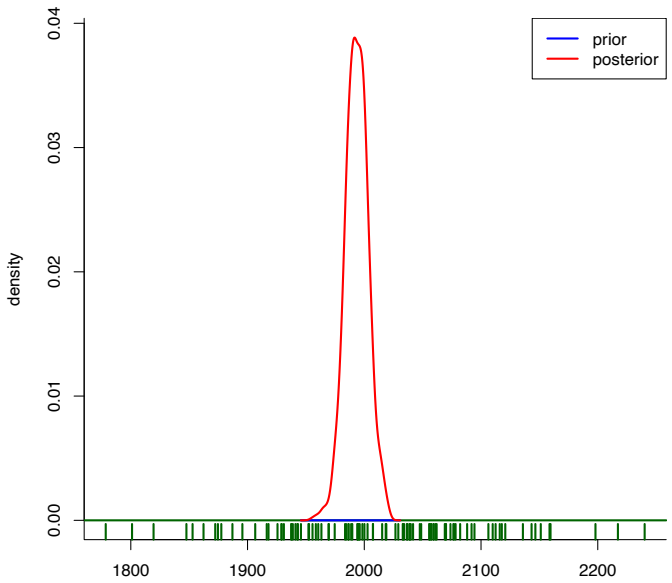
INSTEAD, MEASURE LENGTH

IN DECIMETRES



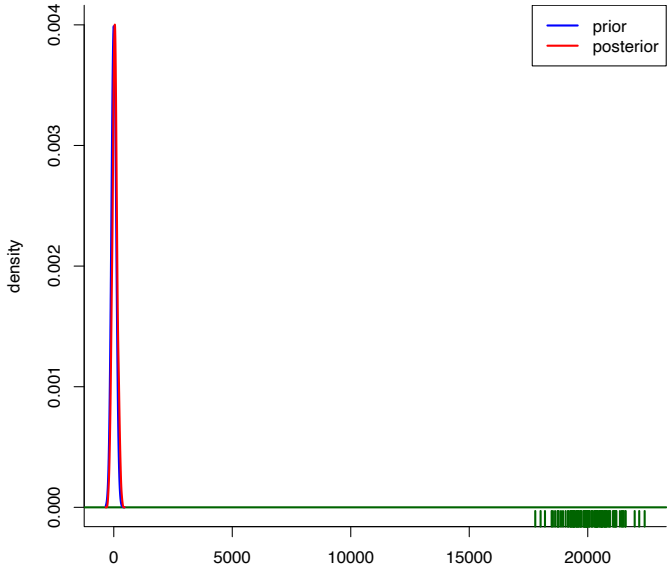
INSTEAD, MEASURE LENGTH

IN CENTIMETRES



INSTEAD, MEASURE LENGTH

IN MILLIMETRES



Summary of Bayesian Analyses With Different Length Units

length units	classical sample mean	Bayes estimate of μ
metres	20.11	20.11
decimetres	201.1	201.1
centimetres	2011	1993
millimetres	20110	44.7

CONCLUSION:

Bayesian inference with a fixed prior specification (such as $\mu \sim N(0, 100^2)$) depends on the units of measurement – sometimes crucially.

For big spreadsheets of data with dozens of columns this is a concern, since there can be data of many variable types on different scales and with different units.

Remedy

If

$$x_1^{\text{orig}}, \ldots, x_{100}^{\text{orig}}$$

are the original snake length data then first standardise to

$$x_i \equiv \frac{x_i^{\text{orig}} - \text{sample mean}}{\text{sample standard deviation}}.$$

Then alright to use:

$$x_i | \mu \overset{\text{ind.}}{\sim} N(\mu, \sigma^2)$$

$$\mu \sim N(0, 100^2), \quad \sigma \sim \text{Half-Cauchy}(0, 100^2)$$

Conversion Back to Original Units

If we use a Bayesian inference engine to do regression analysis for a model such as:

$$\text{price}_i | \beta_0, \beta_1, \sigma^2 \overset{\text{ind.}}{\sim} N\Big(\beta_0 + \beta_1 \, \text{age}_i, \sigma^2\Big)$$

for data on used cars then

$$\beta_1 = \text{depreciation rate}.$$

If standardised data used then best to transform β_1 to meaningful units (e.g. dollars per year).

See Exercise 1 of Assignment 6 (to be handed out next week).

♦ ♦ ♦