Using Stan

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

Normal models

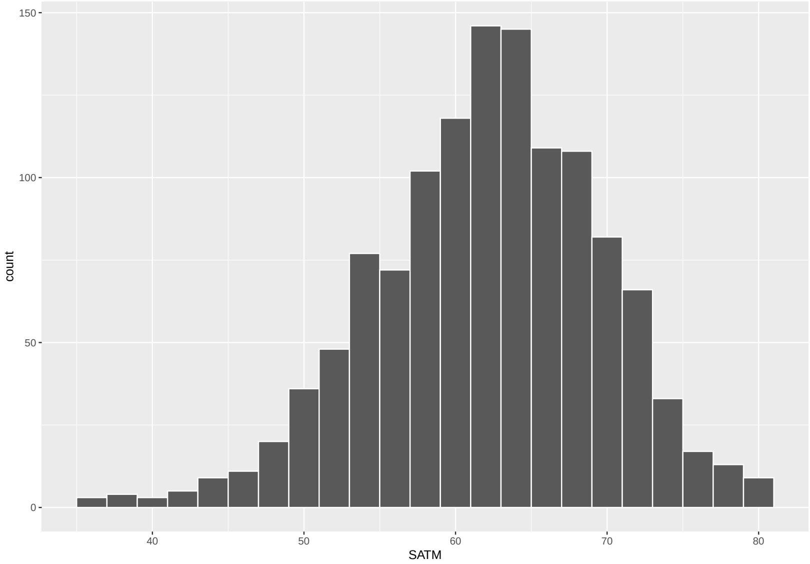

Figure 1: Histogram of mathematical SAT scores in a sample of student in a US university.

Normal models

- Despite is lack of symmetry, a simple and almost default model of this data would be as follows.

$$y_i \sim N(\mu, \sigma^2), \quad \text{for } i \in 1 \ldots n,$$

  where y_i is the maths SAT score of student i and where there are n students in total.
- Obviously, we have two unknowns, μ and σ, and so in a Bayesian model, we first put priors over these two variables.
- Common choices for a prior on the μ parameter of the normal distribution is another normal distribution.
- For the prior over σ, Gelman et al generally recommends heavy tailed distributions over the positive real values such as a half-Cauchy or half-t distribution.

Normal models

- Following these suggestions, our Bayesian model becomes, for example:

$$y_i \sim N(\mu, \sigma^2), \quad \text{for } i \in 1 \dots n,$$
$$\mu \sim N(\nu, \tau^2), \quad \sigma \sim \text{Student}_+(\kappa, \phi, \omega),$$

where Student_+ is the upper half of the (nonstandard) Student t-distribution centered at ϕ, with scale parameter ω, and with degrees of freedom κ. For this choice of prior, we therefore have in total 5 hyper-parameters ν, τ, ϕ, ω and κ.

Using Stan

- A Stan program implementing this model is in the file `normal.stan`.
- We can run this program with `rstan::stan` as follows.

```
y <- read_csv('data/MathPlacement.csv') %>%
  select(SATM) %>%
  na.omit() %>%
  pull(SATM)

N <- length(y)

math_data <- list(y = y,
                  N = N,
                  nu = 50,
                  tau = 25, phi = 0, omega = 10, kappa = 5)
M_math <- stan('normal.stan', data = math_data)
```

Stan Summary

- We can view the summary of the results with `stan_summary`.

```
summary(M_dice,
        pars = pars = c('mu', 'sigma'),
        probs = c(0.025, 0.975))$summary
```

Regression models

- Normal linear regression models are extensions of the normal distribution based model just described.

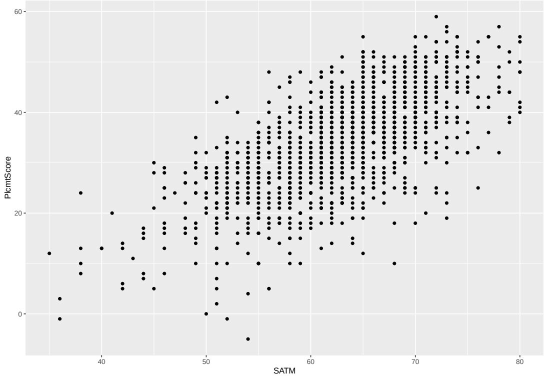

Figure 2: A scatterplot of scores on a mathematics placement exam against maths SAT scores.

Regression models

- Denoting the `PlcmtScore` by y and `SATM` by x, the model can be written as follows.

$$\text{for } i \in 1 \ldots n \quad y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \beta_1 x_i.$$

- There are now three parameters in the model: β_0, β_1, σ.
- We will place normal priors on β_0 and β_1, and half t-distribution on σ.
- As such the full Bayesian model is as follows.

$$y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \beta_1 x_i,$$
$$\beta_0 \sim N(\nu_0, \tau_0^2), \quad \beta_1 \sim N(\nu_1, \tau_1^2), \quad \sigma \sim \text{Student}_+(\kappa, \phi, \omega)$$

Regression models

- The Stan code for this model is in `normallinear.stan`.
- For this example, we will choose the hyperparameters to lead to effectively uninformative priors on β_0 and β_1.
- Specifically, the normal distributions will be centered on zero, i.e. $\nu_0 = \nu_1 = 0$, and will be sufficiently wide, i.e., $\tau_0 = \tau_1 = 50$, so as to be effectively uniform over all practically possible values for β_0 and β_1.
- For the prior on σ, as above, we will use the upper half of Student's t-distribution centered at 0 and with a relatively low degrees of freedom and with a scale ω equal to the MAD of the outcome variable y.

Regression models

- If we place the x and y data vectors and the values of the hyperparameters in the list math_data_2, we can call the Stan program as using rstan::stan as we did above.

```
x <- pull(math_df_2, SATM)
y <- pull(math_df_2, PlcmtScore)

math_data_2 <- list(
  x = x,
  y = y,
  N = length(x),
  tau = 50, omega = mad(y), kappa = 3
)

M_math_2 <- stan('normlinear.stan', data = math_data_2)

stan_summary(M_math_2, pars = c('beta_0', 'beta_1', 'sigma'))
```