# STA365 Homework 4

*Daniel Simpson*

*2020-03-19*

Please submit your answers as a **single pdf document** via Quercus. The documents should be prepared in an RMarkdown document. It will be **Due Tuesday 31 March at 12pm**. Late submissions will be heavily penalized.

This homework counts towards 5% of the final grade for the course.

## Multilevel Regression and Poststratification

A very scientific (and definitely very real) survey was conducted to find out whether a person prefered cats or dogs. The data is here:

```
survey <- readr::read_csv(file = "survey.csv" )
```

```
## Parsed with column specification:
## cols(
##   cat_pref = col_double(),
##   male = col_double(),
##   age = col_double(),
##   eth = col_double(),
##   income = col_double(),
##   state = col_double(),
##   id = col_double()
## )
```

```
head(survey)
```

```
## # A tibble: 6 x 7
##   cat_pref  male   age   eth income state    id
##      <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1        1     0     7     3      1    13     1
## 2        1     0     7     2      1    37     2
## 3        1     1     5     3      2    45     3
## 4        1     1     7     1      1     1     4
## 5        0     1     5     1      3    12     5
## 6        1     0     6     3      3    14     6
```

The first column `cat_pref` records a `1` if the respondent prefered cats over dogs. The remaining columns encode if a person is male or not, their age group (1-7), race/ethnicity (1-3), income group (1-3), and state (1-50). The `id` column is not useful for us.

The full population information is contained in the following poststratification matrix.

```
poststrat <- readr::read_csv(file = "poststrat.csv" )
```

```
## Parsed with column specification:
## cols(
##   male = col_double(),
##   eth = col_double(),
##   age = col_double(),
##   income = col_double(),
##   state = col_double(),
##   N = col_double()
## )
```

```
head(poststrat)
```

```
## # A tibble: 6 x 6
##    male   eth   age income state      N
##   <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl>
## 1     0     1     1      1     1  94877
## 2     0     1     1      1     2 156645
## 3     0     1     1      1     3 137803
## 4     0     1     1      1     4 141987
## 5     0     1     1      1     5 121577
## 6     0     1     1      1     6  93574
```

## Questions

1. Write a stan program that fits a multilevel logistic regression to the survey data. It should tread `male` as a fixed covariate (no random effect) and `age`, `ethnicity`, `income`, and `state` as varying intercepts. In mathematical notation, this is the model

$$y_i \mid p_i \sim \text{Binomial}(1, p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \beta \text{male} + \alpha^{(\text{age})}_{\text{age}(i)} + \alpha^{(\text{eth})}_{\text{eth}(i)} + \alpha^{(\text{inc})}_{\text{inc}(i)} + \alpha^{(\text{state})}_{\text{state}(i)}$$

   where the notation age($i$) is the age group that observation $i$ is in.

2. Use the poststratification matrix and the samples from the posterior (which you can extract using the `extract` function) to get the posterior of the total proportion of cat lovers. Plot the posterior.