STAT 491: Midterm Exam
Due: March 10 at 11:59 PM
Name:

Please turn in the exam to D2L and include the R Markdown code, SAS code *and either* a Word or PDF file with output. While the exam is open book, meaning you are free to use any resources from class, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including class members.** The instructor will answer questions related to the data, expectations, and understanding of the exam, but will not fix or troubleshoot broken code.

1. (36 points OKCupid Data Analysis)

This question will focus on the OKCupid dataset, which can be accessed using http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/OKCupid_profiles_clean.csv. More details on a larger version of this dataset are available at https://github.com/rudeboybert/JSE_OkCupid/blob/master/okcupid_codebook.txt.

The goal will be to fit a model for height for the respondents with sex tagged as 'm' or 'f', note these value are self reported and are recorded in inches.

a. (4 points)

Describe the dataset, what do the columns and variables mean?

```
okc <- read.csv('http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/OKCupid_profiles_clean.cs
str(okc)
```

```
## 'data.frame':    22123 obs. of  10 variables:
##  $ age      : int  22 35 31 37 28 30 29 31 33 27 ...
##  $ body_type: chr  "a little extra" "average" "average" "athletic" ...
##  $ diet     : chr  "strictly anything" "mostly other" "mostly anything" "mostly anything" ...
##  $ drinks   : chr  "socially" "often" "socially" "not at all" ...
##  $ drugs    : chr  "never" "sometimes" "never" "never" ...
##  $ ethnicity: chr  "asian, white" "white" "white" "white" ...
##  $ height   : int  75 70 65 65 72 66 62 71 72 67 ...
##  $ job      : chr  "transportation" "hospitality / travel" "artistic / musical / writer" "student" .
##  $ sex      : chr  "m" "m" "f" "m" ...
##  $ smokes   : chr  "sometimes" "no" "no" "no" ...
```

The dataset contains information scraped from publicly available profiles on OkCupid.com. Each row represents a single user and the following columns are categories about the user and their lifestyle:

- age: age in years
- body_type: a set of categories for self-reported body type
- diet: a set of categories to describe eating habits
- drinks: a set of categories about alcohol habits
- drugs: a set of categories about drug habits
- ethnicity: a set of categories for ethnicity, respondents may choose more than one ethnicity
- height: self-reported height in inches
- job: a set of categories for employment
- sex: male or female

- smokes: a set of categories about smoking habits

b. (4 points)

Specify a prior distribution for height for both genders. We are going to fit these models independently, so there should be two different priors here. For full credit, defend your prior choices - that is why are they reasonable?

c. (4 points)

Select and justify a sampling model for this problem. You can use the same sampling model for each gender.

d. (4 points)

Use JAGS to fit the two posterior distributions. Include your code and for full credit, detail what each line (or chunk) of code is doing and defend your choice for options in the functions.

e. (4 points)

Use the `coda` plotting functionality (or equivalent) to summarize your posterior samples. Describe what the figures mean.

f. (4 points)

Print and interpret the posterior HDIs for the mean height for male and female users.

g. (4 points)

Summarize the posterior for heights of female and male OKCupid users. This should be a written summary without statistical lingo.

h. (4 points)

Use the posterior samples to compute the probability that the mean height for men is taller than 68 inches. Do the same for female heights.

i. (4 points)

Using what you have learned from this dataset, compute the probability that a randomly selected female will be taller than a randomly selected male.

2. (29 points - Basketball Model)

Suppose you have been hired as a statistical consultant by a NBA expansion team that will be moving to Big Sky, MT. Your goal is to help identify basketball players to bring to Montana.

Specifically, the team has one more spot to fill and is looking to add a free throw shooting specialist. Here are two options:

- Bugs Bunny. Bugs Bunny is a shooting guard that was 2 for 7 on free throws last year.
- Tasmanian Devil. Tasmanian Devil is a center that was 25 for 40 on free throws last year.

a.

Use uniform priors, Beta(1,1), to model the free throw shooting for Bugs Bunny and Tasmanian Devil.

a1 (6 points)

Plot and summarize the posterior distribution for each player.

a2 (4 points)

Using the posteriors above, compute the probability that Bugs Bunny has a higher shooting percentage (θ parameter in the binary setting).

a3 (2 points)

If both players had 25 free throws, who do you think will make more free throws? Why?

b.

Now assume that we know that shooting guards, like Bugs Bunny, tend to make about 80 percent of free throws and centers, like Tasmanian Devel, tend to make about 60 percent of free throws. This information can be imparted by more informative priors: Use Beta(40,10) for Bugs Bunny and Beta(30,20) for Tasmanian Devil as the prior distributions.

b1 (6 points)

Plot and summarize the posterior distribution for each player.

b2 (4 points)

Using the posteriors above, compute the probability that Bugs Bunny has a higher shooting percentage (θ parameter in the binary setting).

b3 (2 points)

If both players had 25 free throws, who do you think will make more free throws? Why?

c. (5 points)

Reflect on the results from question a and question b.