STAT 532: Final Exam

For the exam, you may use any course materials or information you find on the internet; however, use citations as appropriate. You may not discuss questions or work together with classmates. You are welcome to contact the instructor with any questions. For complete (and partial credit) please show all work and turn in a reproducible document (PDF or DOC) as well as your source code (.RMD).

This exam will be focused on a data set with housing prices in Washington, D.C. The complete dataset is available on Kaggle.com, but we will be using a smaller, filtered dataset which is presented below.

```
DC <- read_csv('http://math.montana.edu/ahoegh/teaching/stat532/data/DC.csv')
```

Data Dictionary

- BATHRM: number of bathrooms
- HF_BTHRM: number of half bathrooms (no shower or bathtub)
- AC: Does the house have air conditioning
- BEDRM: Number of bedrooms
- STORIES: number of stories
- GBA: Gross Building Area (square feet)
- PRICE: sale price
- CNDTN: House condition: (Excellent/Very Good/Good/Average/Fair/Poor)
- LANDAREA: Land area of the property in square feet
- FULLADDRESS: address of property
- ASSESSMENT_NBHD: neighborhood of the property
- WARD: Ward in the city the property is located in
- QUADRANT: Quadrant in the city that the property is located in

Q1. Frequentist ANOVA (4 points)

Interpret the output from the one-way ANOVA model specified below.

```
quadrant_anova <- aov(PRICE ~ QUADRANT , data = DC)
summary(quadrant_anova)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## QUADRANT       3 1.391e+14 4.637e+13   185.1 <2e-16 ***
## Residuals   2299 5.760e+14 2.505e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(quadrant_anova, 'QUADRANT')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = PRICE ~ QUADRANT, data = DC)
##
## $QUADRANT
##             diff       lwr        upr     p adj
## NW-NE  444265.43  382005.1  506525.79 0.0000000
## SE-NE -123424.58 -193132.1  -53717.04 0.0000331
## SW-NE -216065.30 -493970.2   61839.63 0.1886585
## SE-NW -567690.00 -637022.3 -498357.67 0.0000000
```

```
## SW-NW -660330.73 -938141.8 -382519.67 0.0000000
## SW-SE  -92640.73 -372215.0  186933.54 0.8294890
```

Q2. 1-way ANOVA (22 points)

Notice that the ANOVA model can be parameterized similarly as a linear model where

$$y_{ij} = \mu_j + \epsilon_{ij}$$

where y_{ij} is the price of house i in group $j \in 1, 2, 3, 4$, μ_j is the mean for group j, and $\epsilon_{ij} \sim N(0, \sigma^2)$.

a. (4 points)

Specify priors for the parameters in the model above.

b. (6 points)

Using a Bayesian procedure, and your priors, find credible intervals for the parameters in the model. Please include your code for the procedure.

c. (4 points)

Now using your posterior results, find credible intervals for a contrast between the mean prices in the NW and SE quadrants. Please include your code for the procedure.

d. (4 points)

Assume a classmate has been offered a job in Washington, D.C. Describe your results from parts b and c.

e. (4 points)

How is the interval from part c similar to the one calculated in Q1 and how is it different? This should include a discussion that focuses on both the philosophical interpretation and the practical values.

Q3. Hierarchical Model (30 points)

a. (6 points)

Write out a hierarchical model to estimate the mean price for each neighborhood. Please include all prior specification as well.

b. (4 points)

Compare and contrast a 1-way ANOVA from Question 2 and the hierarchical model in Question 3a. In other words, how are the models similar and how are they different?

c. (6 points)

Fit the hierarchical model. Report posterior credible intervals for the overall population mean as well as the following neighborhoods (Anacostia, Capital Hill, and Georgetown) Please include your code for the procedure. Note you should include all neighborhoods in the analysis, but only report the credible intervals for the three listed above.

d. (4 points)

Again assume a classmate has been offered a job in Washington, D.C. Describe your results from part c.

e. (6 points)

Create a plot(s) of the posterior predictive distribution for the mean housing price in Anacostia, Capital Hill, and Georgetown. How do your results compare with data?

f. (4 points)

Assume your classmate wanted to know how the price per square foot (of Gross Building Area) differed across the three neighborhoods. Summarize how you would answer this question (but you do not need to implement it).

Q4. Multinomial Regression Model (10 points)

Another possible question of interest might be a model that attempts to predict which quadrant a house is located in using relevant covariates. This is an extension of a GLM used for binary regression (such as logistic regression) to account for more than 2 possible outcomes.

The sampling model can be written as

$$Y_i \sim Multinomial(n = 1, \tilde{\theta_i})$$

where $Y_i \in (1, 2, 3, 4)$ denoting the four possible quadrants for house i and $\tilde{\theta_i} = (\theta_{i1}, \theta_{i2}, \theta_{i3}, \theta_{i4})$ is a vector with four probabilities that correspond to the probability of each quadrant.

The linear combination of predictors can be mapped to the probabilities using the softmax function where

$$\begin{aligned}
\theta_{i1} &= Pr[Y_i = 1] = \frac{\exp(x_i^T \tilde{\beta}_1)}{\sum_{k=1}^{4} \exp(x_i^T \tilde{\beta}_k)} \\
\theta_{i2} &= Pr[Y_i = 2] = \frac{\exp(x_i^T \tilde{\beta}_2)}{\sum_{k=1}^{4} \exp(x_i^T \tilde{\beta}_k)} \\
\theta_{i3} &= Pr[Y_i = 3] = \frac{\exp(x_i^T \tilde{\beta}_3)}{\sum_{k=1}^{4} \exp(x_i^T \tilde{\beta}_k)} \\
\theta_{i4} &= Pr[Y_i = 4] = \frac{\exp(x_i^T \tilde{\beta}_4)}{\sum_{k=1}^{4} \exp(x_i^T \tilde{\beta}_k)}
\end{aligned}$$

where x_i^T is a vector of covariates (e.g. stories, price, GBA, ...) and $\tilde{\beta}_k$ is a vector of coefficients that correspond to the k^{th} class. Note that each class has a distinct vector of coefficients.

a. Priors (4 points)

List all of the parameters in the model that require priors and select and defend prior distributions for these parameters. Note that this should include values, not just the parameteric families.

b. Metropolis Algorithm (6 points)

In one or two sentences, describe the Metropolis Algorithm and discuss why it is necessary (in general). Then sketch out how a Metroplis algorithm could be used in this situation. Note you do not need to implement the procedure, just describe the necessary components.