Stat 532 **Name:** _____

Take home midterm

Due 10/17/16 at 10:00 AM

For the take home exam, you may use the textbook, any course materials provided on D2L, homeworks, and labs. You **may not** discuss questions or work together with classmates. You are welcome to contact the instructor with any questions related to better explantion or understanding of the questions themselves. Any relevant material from questions will be posted to D2L for the benefit of the entire class. For complete (and partial credit) please show all work, whether that be by hand or printed R code.

1. (10 points) Let $p(\theta|\sigma^2) \sim N(\mu_0, \sigma^2/\kappa_0)$ and $p(1/\sigma^2) \sim Gamma(\nu_0/2, \nu_0\sigma_0^2/2)$. If $p(y_1, \ldots, y_n|\theta, \sigma^2) \sim N(\theta, \sigma^2)$, show that the marginal posterior distribution $p(1/\sigma^2|y_1, \ldots, y_n) \sim Gamma(\nu_n/2, \nu_n\sigma_n^2/2)$, where $\nu_n = \nu_0 + n$ and $\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}\right]$ and $s^2 = \sum_{i=1}^n (y_i - \bar{y})/(n-1)$.

2. Capital Bikeshare is a bike rental company in the Washington D.C. area with almost 400 hubs and over 3500 bikes. Using actual data collected by the company, estimate θ the average number of bikes rented at a hub across the system. The data can be found on D2L in the `MidtermBikes.csv` file.

   (a) (4 points) As you will see the data consist of counts collected on June 23, 2016 for each station. What is a reasonable sampling model for the data that we have talked about in class?

   (b) (4 points) Choose and justify a prior distribution on θ, the average number of bikes rented per station.

   (c) (5 points) Import the data and find the posterior distribution of $p(\theta|y_1, \ldots, y_n)$. With a judicious selection of the sampling model and prior, this should be available in closed form.

(d) (5 points) Draw samples from the posterior predictive distribution $p(y^*|y_1, \ldots, y_n)$. Recall you do not need to know the form of the posterior predictive distribution to draw these samples. Plot histograms of the posterior distribution ($p(\theta|y_1, \ldots, y_n)$) and the posterior predictive distribution ($p(y^*|y_1, \ldots, y_n)$) side by side on the same scale. Comment on the differences between the distributions.

(e) (5 points) There are 8 slots at each of the 392 stations in the dataset. Compute a point estimate and a 95% credible interval for the average number of bikes rented per slot (note this is a function of θ).

(f) (2 points) Is your point estimate in part (e) a *Maximum A'Posteriori* (MAP) point estimate? Justify your answer.

(g) (2 points) Is your credible interval for part(e) a Highest Posterior Density (HDP) region? Justify your answer.

3. A simple change point model contains a break point where the mean of a process shifts. Consider data from housing sales in the Bozeman area from 2005-2013. We are interested in testing whether a change point occured between 2008 and 2009. Note Lehman Brothers filed for bankruptcy on Sept. 15, 2008. While this is a very rich dataset and you could likely formulate a time-series regression model, we will focus only on the average sales price per square foot (Closing_Price_per_sqft) and YearSold and assume the sales price comes from a normal distribution. The dataset can be found on D2L and is called `BozemanHousing.csv`. Note you will need to partition the dataset into two parts based on the YearSold variable. Here is one way to do this:

```
time1 <- subset(BozemanHousing, YearSold ==‘2005-2008’
```

   (a) (4 points) Set prior distributions for $\{\theta_1, \theta_2, \sigma_1^2, \sigma_2^2\}$, where $\{\theta_1, \sigma_1^2\}$ correspond to 2005 - 2008 and $\{\theta_2, \sigma_2^2\}$ correspond to 2009 - 2013. This question is intended to assess your MCMC skills so use the form $p(\theta, \sigma^2) = p(\sigma_2)p(\theta)$. Justify your choice. Note you may use the precision term $(1/\sigma^2)$ if you'd prefer.

   (b) (5 points) Implement a Gibbs sampler for $p(\theta_1, \sigma_1^2|y_1, \ldots, y_{n_1})$ and $p(\theta_2, \sigma_2^2|y_1, \ldots, y_{n_2})$. Comment on the convergence of your algorithm.

(c) (5 points) Display the marginal posterior distributions $p(\theta_1|y_1, \ldots, y_{n_1})$ and $p(\theta_2|y_1, \ldots, y_{n_2})$ on the same scale.

(d) (5 points) Compute the $Pr(|\theta_2 - \theta_1| > c)$, where $c = \$10/\text{ft}^2$ is a constant with a magnitude signifying that a change point has or has not occured.

(e) (4 points) Summarize and reflect on your findings in part (d).