Activity 7

Name here

This analysis will focus on small dataset containing information from Indeed.com, which can be accessed using http://www.math.montana.edu/ahoegh/teaching/stat491/data/bzn__jobs.csv.

```
bzn_jobs <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat491/data/bzn_jobs.csv')
```

```
## Rows: 30 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): normTitle
## dbl (3): jobAgeDays, estimatedSalary, localClicks
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

This dataset contains the following variables:

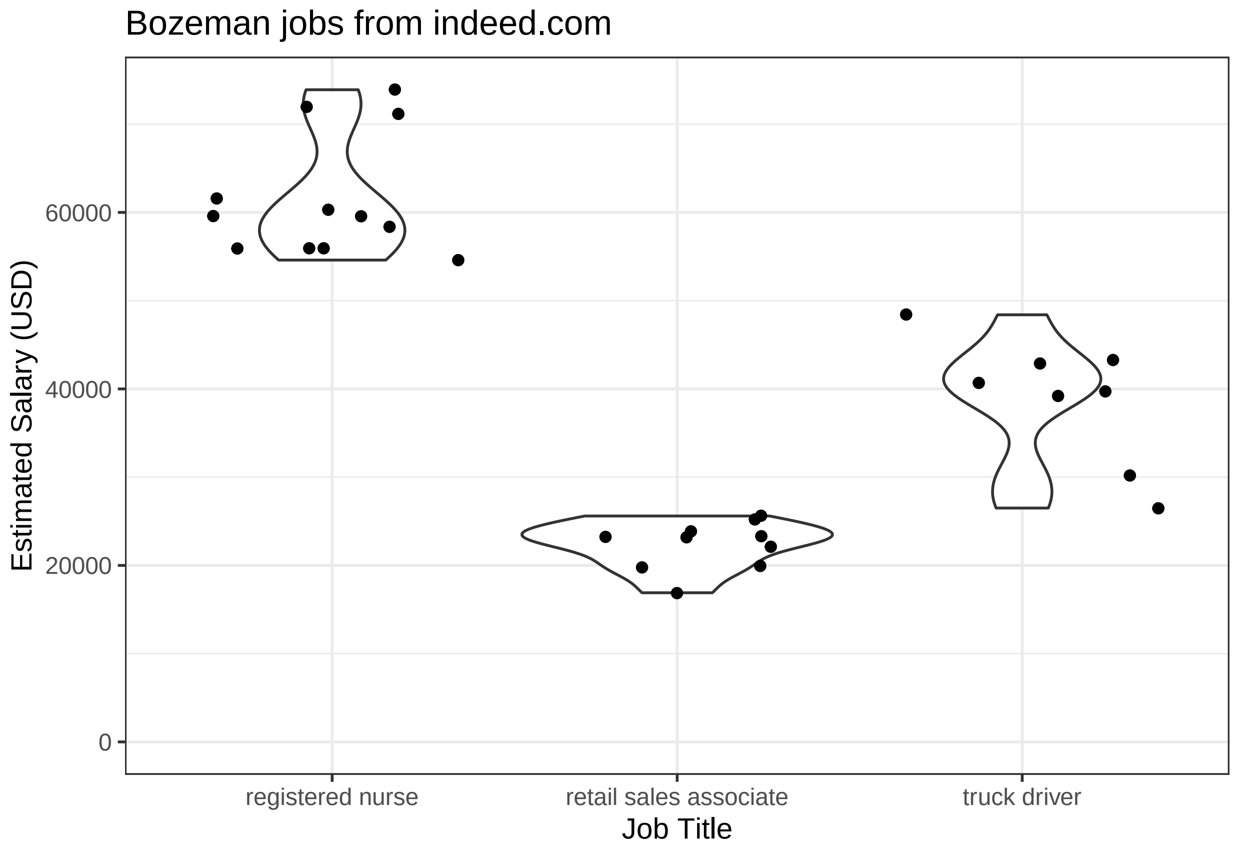
```
- jobAgeDays: number of days the job has been posted on Indeed.com
- normTitle: name of job position (registered nurse, sales associate, truck driver)
- estimatedSalary: estimated annual salary
- localClicks: number of people clicking on job posting
```

Bayesian ANOVA

For this question we will fit a regression analysis (ANOVA) to model estimated salary across three diffferent job types. variables as predictors.

a. Data Viz Create a figure of salary by `normTitle`. It is good practice to show all data points.

```
bzn_jobs %>% ggplot(aes(y = estimatedSalary, x = normTitle)) +
   theme_bw() + geom_violin() + geom_jitter() +
  xlab("Job Title") + ylab("Estimated Salary (USD)") + ylim(0, NA) +
  ggtitle('Bozeman jobs from indeed.com')
```



b. Interpret the following R output.

```
anova_fit <- lm(estimatedSalary ~ normTitle - 1, data = bzn_jobs)
summary(anova_fit)
```

```
##
## Call:
## lm(formula = estimatedSalary ~ normTitle - 1, data = bzn_jobs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12362.5  -3008.7    181.2   2626.9  12325.0
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## normTitleregistered nurse         61575       1706   36.10  < 2e-16 ***
## normTitleretail sales associate   22310       1869   11.94 2.78e-12 ***
## normTitletruck driver             38862       2089   18.60  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5909 on 27 degrees of freedom
## Multiple R-squared:  0.9852, Adjusted R-squared:  0.9835
## F-statistic: 597.2 on 3 and 27 DF,  p-value: < 2.2e-16
```

```
confint(anova_fit)
```

```
##                                     2.5 %   97.5 %
## normTitleregistered nurse       58075.08 65074.92
## normTitleretail sales associate 18476.03 26143.97
## normTitletruck driver           34575.99 43149.01
```

This is an ANOVA model, formulated through the reference case framework. We have estimated average salaries for each job title.

c. Select and Justify a sampling model for your response.

We will fit the following model:

$$salary = \beta_1 \times I_{job=rn} + \beta_2 \times I_{job=sales} + \beta_1 \times I_{job=driver} + \epsilon; \epsilon \sim N(0, \sigma^2)$$

While all salaries are positive, a normal distribution looks reasonable in this case as the values are all pretty far removed from 0. This model also inherently has an assumption about equal variance between the groups, which looks reasonable.

d. Explain the purpose of this model - you can assume you talking to a freshman in high school.

This model looks at estimated salary for a few different job titles. It will allow us to understand how estimated salary differs between these three job types.

e. State and Justify Priors Used for your Model

There will need to be three similar priors for β_1, β_2, and β_3. With little knowledge of salaries of these three job types, I'll use fairly uninformative priors. In particular, these will be Normal priors centered at \$50,000, with standard deviation of \$25,000. The net result is that most of my prior mass will be between 0 and 100k.

I'll use an informative inverse gamma prior on σ^2 with values .01 and .01.

e. Modify existing JAGS code to fit this model.

```
indicator_data <- model.matrix(estimatedSalary~normTitle - 1, data = bzn_jobs)
```

```
model_anova<- "model{
  # Likelihood
  for (i in 1:n){
    y[i] ~ dnorm(beta1 * x1[i] + beta2 * x2[i] + beta3 * x3[i], 1/sigma^2)
  }

  # Prior
  beta1 ~ dnorm(50000, 1/25000^2)
  beta2 ~ dnorm(50000, 1/25000^2)
  beta3 ~ dnorm(50000, 1/25000^2)
  sigma ~ dunif(0, 1000000000)
}"
```

f. (4 points) Using your JAGS code to fit the Posterior Distribution for this Model and print the results

```
dataList = list(y = bzn_jobs$estimatedSalary, n = nrow(bzn_jobs),
                x1 = as.numeric(indicator_data[,1]),
                x2 = as.numeric(indicator_data[,2]),
                x3 = as.numeric(indicator_data[,3]))
```

```r
model <- jags.model(file = textConnection(model_anova), data = dataList)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 30
##    Unobserved stochastic nodes: 4
##    Total graph size: 144
##
## Initializing model
```

```r
update(model, n.iter = 5000) # warmup


# take samples
posterior_sample <- coda.samples(model,
                       variable.names = c("beta1", 'beta2','beta3', 'sigma'),
                       n.iter = 10000)

summary(posterior_sample)
```

```
##
## Iterations = 6001:16000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##        Mean   SD Naive SE Time-series SE
## beta1 61467 1814    18.14          18.14
## beta2 22522 1988    19.88          19.88
## beta3 38926 2217    22.17          22.17
## sigma  6218  918     9.18          11.67
##
## 2. Quantiles for each variable:
##
##        2.5%   25%   50%   75% 97.5%
## beta1 57878 60295 61494 62640 65069
## beta2 18518 21239 22534 23820 26491
## beta3 34573 37460 38929 40384 43300
## sigma  4726  5574  6113  6747  8296
```

```r
library(bayestestR)
library(knitr)
hdi(posterior_sample) %>% kable()
```

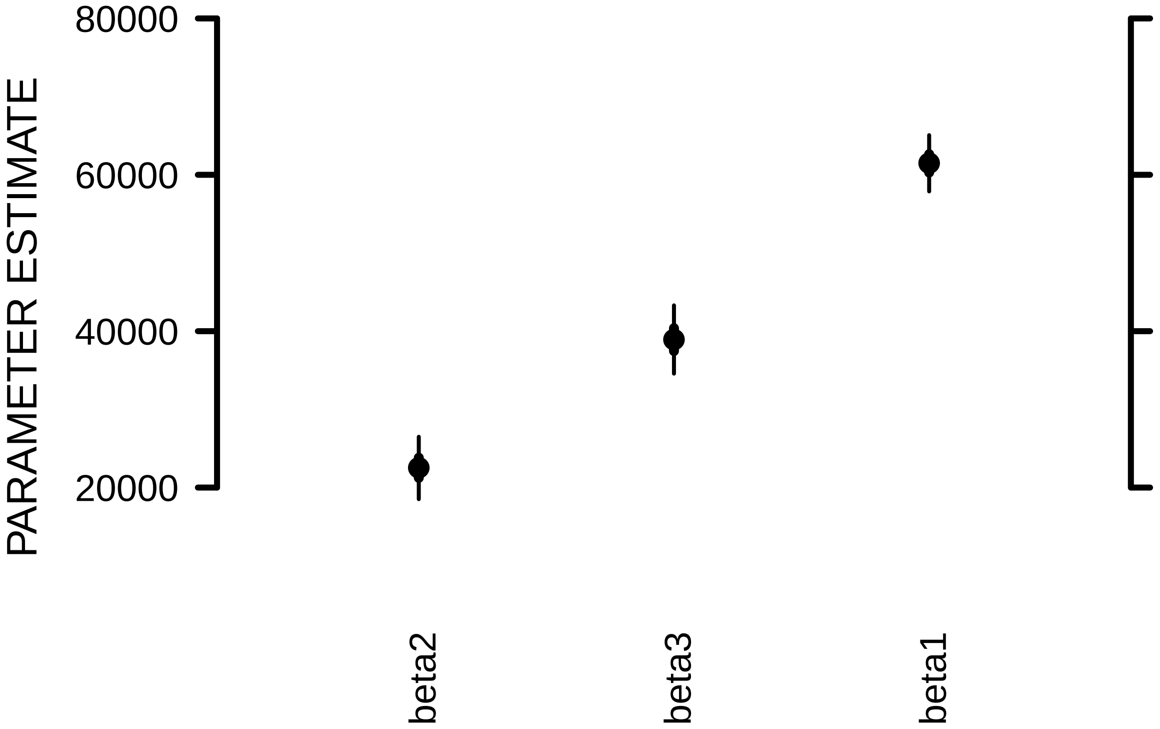
Parameter	CI	CI_low	CI_high
beta1	0.95	58018.694	65167.817
beta2	0.95	18453.765	26380.004
beta3	0.95	34745.106	43447.309
sigma	0.95	4532.581	8003.793

We have estimates for all four parameters in our model. The following table contains 95% credible intervals.

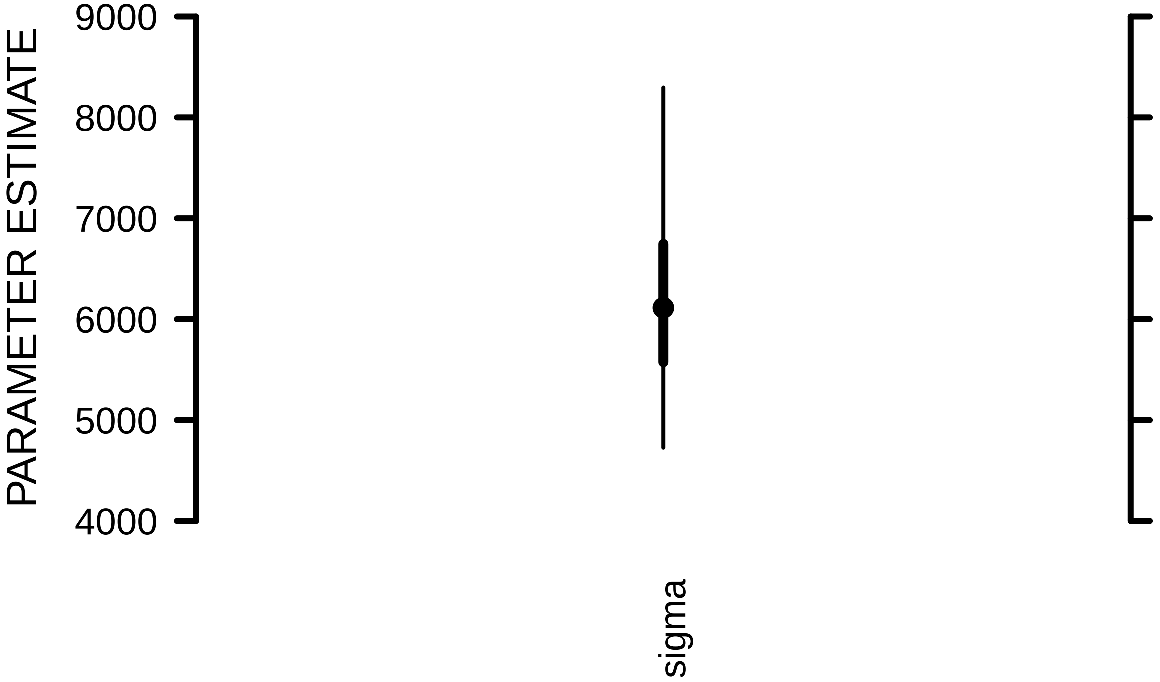
g. Visualize your results in some fashion - this can look similar to your figure in part a.

```
library(MCMCvis)

MCMCplot(posterior_sample,
         params = c('beta1', 'beta2', 'beta3'),
         rank = TRUE,
         horiz = FALSE,
         ylab = 'PARAMETER ESTIMATE')
```



```
MCMCplot(posterior_sample,
         params = c('sigma'),
         rank = TRUE,
         horiz = FALSE,
         ylab = 'PARAMETER ESTIMATE')
```



h. Compare your interval results, in part f, with those from part b

The intervals values are generally in agreement between the two models, with differences in the dollars of tens of dollars. However, the first model does not have an interval estimate for σ^2 but just a point estimate, while the second model has an interval for σ.

i. Explain the results of this model - you can assume you talking to a freshman in high school.

Hi Kid,

As you will soon understand, life is expensive. We've identified that in general nurses tend to make more money than truck drivers who tend to make more money that sales associates. Nevertheless, chase your dreams, but give some thought to finances along the way.