EEEB UN3005/GR5005
Homework - Week 06 - Due 12 Mar 2019

Xun Zhao, xz2827

Homework Instructions: Complete this assignment by writing code in the code chunks provided. If required, provide written explanations below the relevant code chunks. Replace "USE YOUR NAME HERE" with your name in the document header. When complete, knit this document within RStudio to generate a pdf. Please review the resulting pdf to ensure that all content relevant for grading (i.e., code, code output, and written explanations) appears in the document. Rename your pdf document according to the following format: hw_week_06_firstname_lastname.pdf. Upload this final homework document to CourseWorks by 5 pm on the due date.

Problem 1 (4 points)

In lab this week you used the `simulated_trees.csv` dataset to specify a linear regression model with tree age (years) as the outcome variable and tree height (centimeters) as the predictor variable. Using the same dataset, do the following:

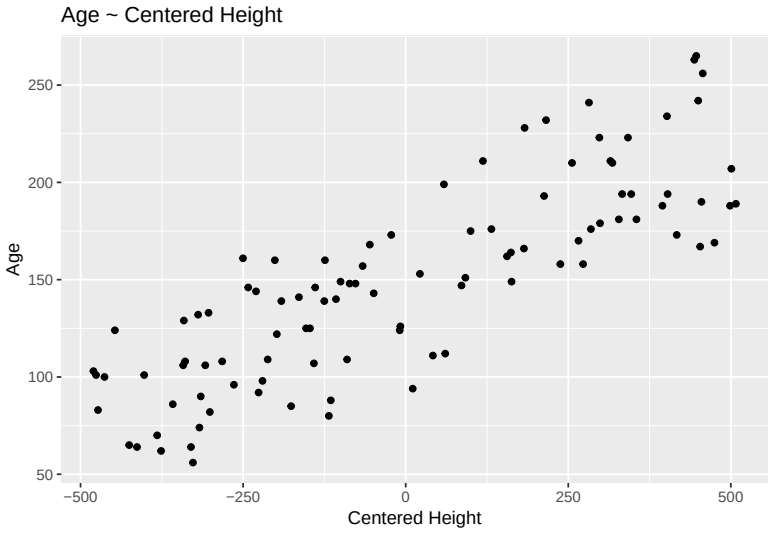
- create a centered tree height variable
- plot tree age (y-axis) vs. the centered tree height variable (x-axis)
- fit a linear regression model for tree age using centered tree height as the predictor variable

Assume a prior of `dnorm(0, 50)` for both the intercept and slope parameters and a prior of `dcauchy(0, 5)` for the standard deviation parameter. Also note, you'll need to use start values as follows to ensure a good model fit: intercept parameter = 50, slope parameter = 0, standard deviation parameter = 50.

Summarize your fit model parameters using 99% PIs. How do the parameter posteriors in this model compare to the linear regression you fit during lab? How do you interpret the intercept parameter in this model?

```
d = read.csv('simulated_trees.csv')
centered.height = d$height - mean(d$height)
centered.d = data.frame(centered.height, d$age)
colnames(centered.d) = c('centered.height', 'age')
ggplot(centered.d, aes(x = centered.height, y = d$age)) +
    geom_point() +
    xlab('Centered Height') +
```

```
    ylab('Age') +
    ggtitle('Age ~ Centered Height')
```




```
model = map(
    alist(
        age ~ dnorm(mu, sigma),
        mu <- a + b * centered.height,
        a ~ dnorm(0, 50),
        b ~ dnorm(0, 50),
        sigma ~ dcauchy(0, 5)),
    start = list(a = 50, b = 0, sigma = 50),
    data = centered.d)
precis(model, prob = 0.99)
```

```
##          Mean StdDev   0.5%  99.5%
## a     148.07   2.74 141.01 155.13
## b       0.14   0.01   0.12   0.17
## sigma  27.44   1.92  22.49  32.40
```

Answer:

Compared with the parameters in the lab, the intercept `a` changes a lot (from 7 to 148), while slope `b` and standard deviation `sigma` are almost unchanged.

The change of `a` can be calculated as follows, where $E()$ is the expectation function.

$$
\begin{aligned}
\text{age} &= a + b \cdot \text{height} \\
E(\text{age}) &= E(a + b \cdot \text{height}) \\
E(\text{age}) &= E(a) + E(b)E(\text{height}) \\
E(a) &= E(\text{age}) - E(b)E(\text{height})
\end{aligned}
$$

When height is centered, $E(\text{height})$ becomes 0, so $E(a) = E(age)$.

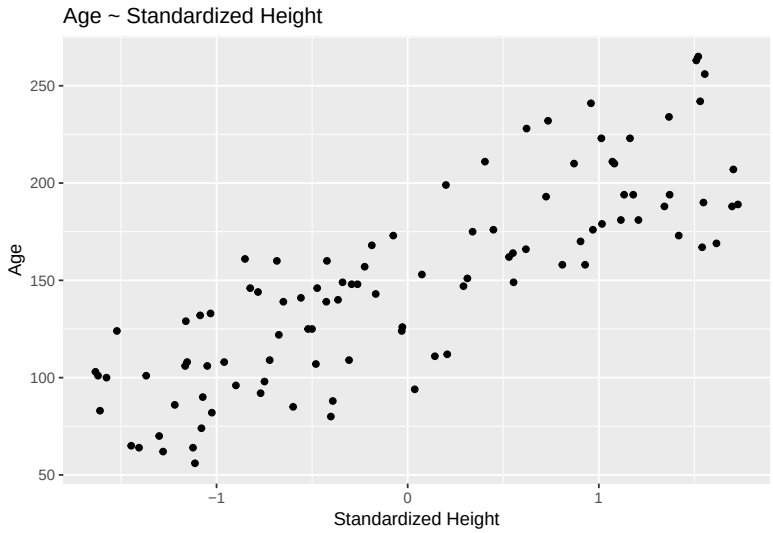
Problem 2 (4 points)

Now:

- create a standardized tree height variable
- plot tree age (y-axis) vs. the standardized tree height variable (x-axis)
- fit a linear regression model for tree age using standardized tree height as the predictor variable (use the same priors and start values as in Problem 1)

Summarize your fit model parameters using 99% PIs. How do the parameter posteriors in this model compare to the linear regression fit with the centered tree height variable? How do you interpret the slope parameter in this model?

```
d = read.csv('simulated_trees.csv')
std.height = (d$height - mean(d$height)) / sd(d$height)
std.d = data.frame(std.height, d$age)
colnames(std.d) = c('std.height', 'age')
ggplot(std.d, aes(x = std.height, y = d$age)) +
    geom_point() +
    xlab('Standardized Height') +
    ylab('Age') +
    ggtitle('Age ~ Standardized Height')
```




```
model = map(
    alist(
        age ~ dnorm(mu, sigma),
        mu <- a + b * std.height,
        a ~ dnorm(0, 50),
        b ~ dnorm(0, 50),
        sigma ~ dcauchy(0, 5)),
    start = list(a = 50, b = 0, sigma = 50),
    data = std.d)
precis(model, prob = 0.99)
```

```
##          Mean StdDev   0.5%  99.5%
## a     148.07   2.74 141.01 155.14
## b      41.69   2.75  34.59  48.78
## sigma  27.44   1.92  22.49  32.39
```

Answer:

The change of `a` has been explained above.

The change of `b` can be calculated below, where $Var()$ is the variance function.

$$\begin{aligned} \text{age} &= a + b \cdot \text{height} \\ Var(\text{age}) &= Var(a + b \cdot \text{height}) \\ Var(\text{age}) &= Var(a) + Var(b)Var(\text{height}) \\ Var(b) &= \frac{Var(\text{age}) - Var(a)}{Var(\text{height})} \end{aligned}$$

When `height` is standardized, $Var(\text{height})$ becomes 1. So does the standard deviation of

height. So $Var(b)$ changes. In addition, $Var(a)$ comes from the modeling, which makes $Var(a)$ much smaller than $Var(\text{age})$.

Thus, $Var(\text{height}) = 1$ and $Var(a) \ll Var(\text{age})$.

Then, $Var(b) \approx Var(\text{age})$.

Problem 3 (2 points)

Using the model you fit in Problem 2, generate 10,000 tree age predictions for a tree of average height (i.e., the average in the `simulated_trees` dataset). Report the mean and 50% HPDI of these predictions, and visualize the predictions using a density plot.

```
sample = extract.samples(model, 10000)
preds = rnorm(10000, mean = sample$a + sample$b * mean(std.height), sd = sample$sigma)
print(HPDI(preds, 0.5))
```

```
##     |0.5     0.5|
## 130.1015 166.9743
```

```
dens(preds)
```