Bayesian Analysis - Assignment 2

Chris Hayduk

February 9, 2019

1 Chapter 4 (4M4-4M5-4M6) Extensions

1. Calculate the minimum, maximum, median, mean, and standard deviation of heights for people 18 and younger, separately for males and females. Would you change your priors based on this information? Justify.

Solution:

```
under_18 <- data[data$age <= 18, ]

under_18_male <- data[data$male == 1, ]

under_18_female <- data[data$male == 0, ]

summary <-
  list("Height for Males Under 18" =
       list("min" = ~ min(.data$height),
            "median" = ~ median(.data$height),
            "max" = ~ max(.data$height),
            "mean (sd)" = ~ qwraps2::mean_sd(.data$height)))

table1 <- summary_table(under_18_male, summary)

#print(table1, markup = "latex")

summary <-
  list("Height for Females Under 18" =
       list("min" = ~ min(.data$height),
            "median" = ~ median(.data$height),
            "max" = ~ max(.data$height),
            "mean (sd)" = ~ qwraps2::mean_sd(.data$height)))

table2 <- summary_table(under_18_female, summary)

#print(table2, markup = "latex")
```

	Under 18 Males (N = 257)
Height for Males Under 18	
min	60.452
median	157.48
max	179.07
mean (sd)	142.32 ± 28.87

	Under 18 Females (N = 287)
Height for Females Under 18	
min	53.975
median	146.05
max	162.56
mean (sd)	134.63 ± 25.93

After examining the data in the two tables above, I would change the priors to something like the following:

$$\text{height}_\text{i} \sim \text{Normal}(\mu_\text{i}, \sigma)$$
$$\mu_\text{i} = \alpha + \beta * \text{year}_\text{i}$$
$$\alpha \sim \text{Normal}(57, 25)$$
$$\beta \sim \text{Normal}(38, 20)$$
$$\sigma \sim \text{Uniform}(0, 50)$$

I selected the mean for α by taking the weighted average of the minimum heights for males and females. I selected the mean for β by subtracting the weighted average of minimum heights from the weighted average of maximum heights. I then divided this number by 3 in order to derive the average amount that an individual must grow over the course of 3 years in order to reach the weighted average maximum height after starting from the weighted average minimum height. I also increased the standard deviation for β to reflect the uncertainty in the growth rate depending upon age and sex.

2. Plot height (y-axis) against age (x-axis) with the points color-coded by male/female, just for those 18 and younger. What type of relationship do you see? Can you use this information to set up a model for the type of data described in 4M4 (ie., each student measured for three years)? Justify.

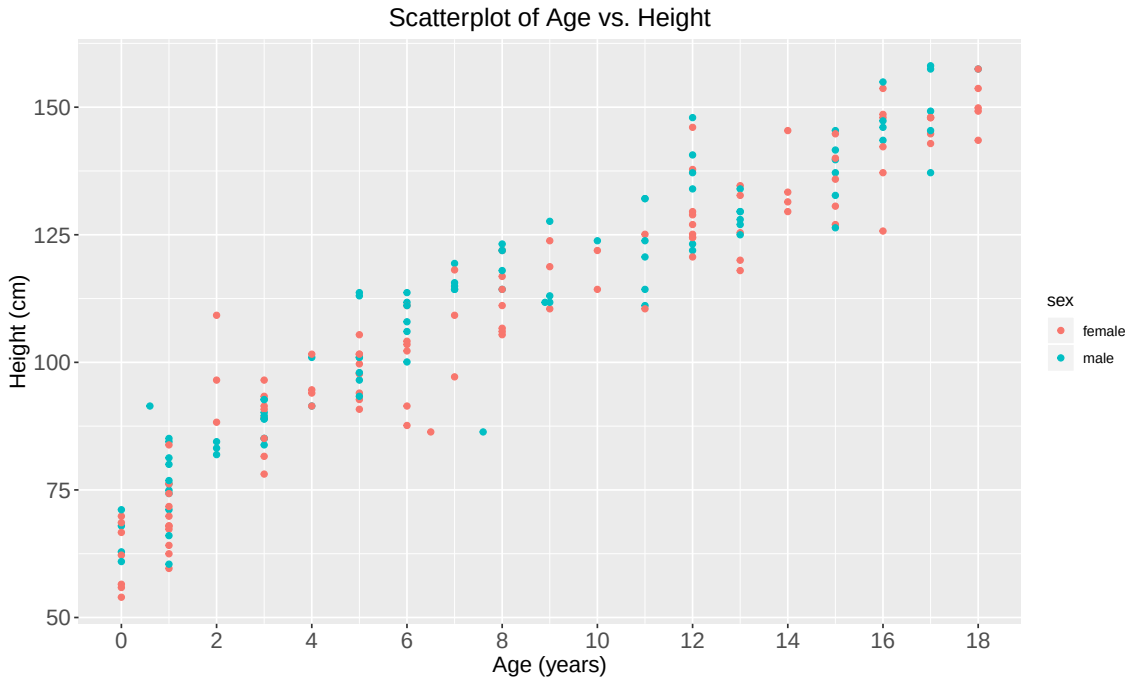
Solution:

```r
under_18$male <- ifelse(under_18$male == 1, "male", "female")

names(under_18) <- c("height", "weight", "age", "sex")

sp <- ggplot(under_18, aes(x=age, y=height)) +
  geom_point(aes(color=sex)) +
  ggtitle("Scatterplot of Age vs. Height") +
  labs(x = "Age (years)", y = "Height (cm)") +
  scale_x_continuous(breaks = pretty(under_18$age, n = 7)) +
  theme.info

sp
```



The relationship between age and height appears to be roughly quadratic. Thus, we may want to modify the model as follows:

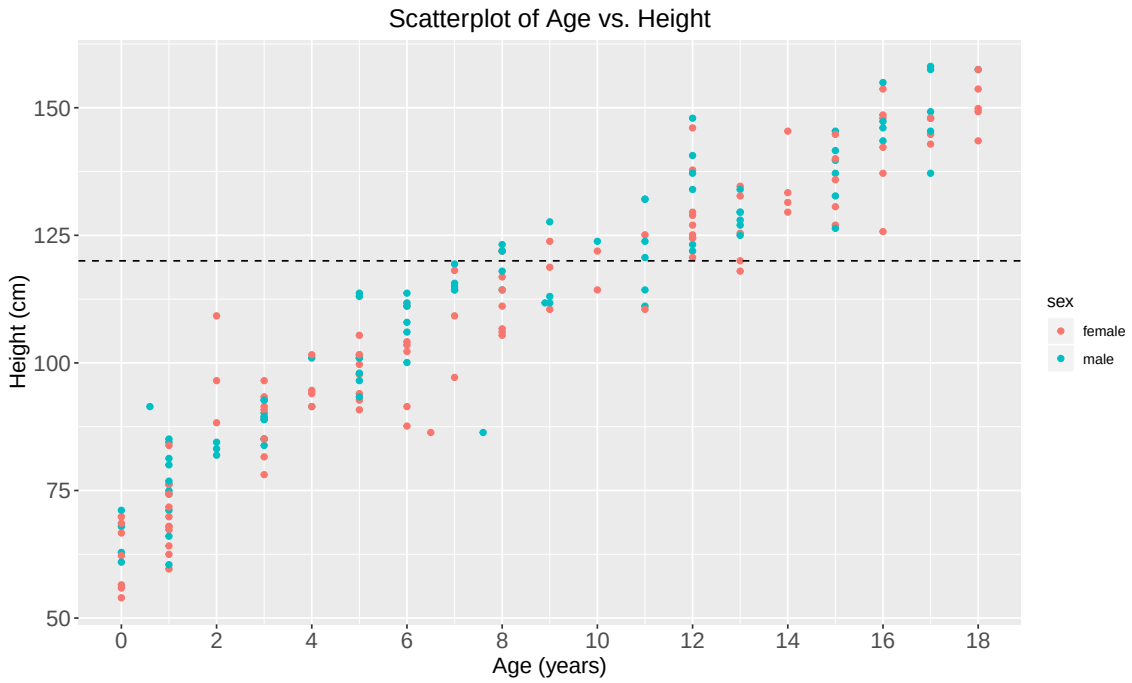
$$\begin{aligned}
\text{height}_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_1 * \text{year}_i + \beta_2 * \text{year}_i^2 \\
\alpha &\sim \text{Normal}(57, 25) \\
\beta_1 &\sim \text{Normal}(0, 50) \\
\beta_2 &\sim \text{Normal}(0, 50) \\
\sigma &\sim \text{Uniform}(0, 50)
\end{aligned}$$

Now that the relationship between the coefficients and the outcome is not so clear, I have centered the priors for both β terms at 0 and increased their uncertainty.

3. Add a horizontal line at 120 cm. on your height vs. age plot. In 4M5, you learn that the average height in the 1st year was 120 cm. Can you guess roughly how old the students in such a data set would be? Would it be reasonable to assume the students in the data set are all the same age?

Solution:

```
sp + geom_hline(yintercept = 120, linetype = "dashed")
```



Judging by the plot, it appears that an average height of 120 cm is attained somewhere between 9 - 11 years old. Let's check the numbers:

```
nine_years_old <- under_18[under_18$age == 9,]
print(mean(nine_years_old$height))

## [1] 117.5808

ten_years_old <- under_18[under_18$age == 10,]
print(mean(ten_years_old$height))

## [1] 120.015

eleven_years_old <- under_18[under_18$age == 11,]
print(mean(eleven_years_old$height))

## [1] 121.2056
```

We can see based upon the output above that, if all students in the data set are the same age, they are most likely starting at 10 years old. It is reasonable to assume that the students are all the same age because the growth rate would not be very meaningful if we were looking at students from a wide range of ages. Furthermore, the average height for students that are 10 years old nearly exactly matches the average given in the question.

4. The Centers for Disease Control (CDC) have height and weight charts by age for boys and girls. Find and download this information. Based on those results, write a suitable Bayesian model for predicting height from age. Justify your choices.

Solution:

```
boys_height_chart <- read_csv("Males Age 2-20.csv")
girls_height_chart <- read_csv("Females Age 2-20.csv")

head(boys_height_chart)
```

```
## # A tibble: 6 x 10
##   `Age (in months~ `3rd Percentile~ `5th Percentile~ `10th Percentil~
##              <dbl>            <dbl>            <dbl>            <dbl>
## 1             24               79.9             80.7             82.0
## 2             24.5             80.3             81.1             82.4
## 3             25.5             81.0             81.8             83.1
## 4             26.5             81.7             82.6             83.8
## 5             27.5             82.4             83.3             84.6
## 6             28.5             83.1             84.0             85.3
## # ... with 6 more variables: `25th Percentile Stature (in
## #   centimeters)` <dbl>, `50th Percentile Stature (in centimeters)` <dbl>,
## #   `75th Percentile Stature (in centimeters)` <dbl>, `90th Percentile
## #   Stature (in centimeters)` <dbl>, `95th Percentile Stature (in
## #   centimeters)` <dbl>, `97th Percentile Stature (in centimeters)` <dbl>
```

```
head(girls_height_chart)

## # A tibble: 6 x 10
##   `Age (in months~ `3rd Percentile~ `5th Percentile~ `10th Percentil~
##              <dbl>            <dbl>            <dbl>            <dbl>
## 1             24               78.4             79.3             80.5
## 2             24.5             78.8             79.6             80.9
## 3             25.5             79.6             80.4             81.7
## 4             26.5             80.4             81.2             82.5
## 5             27.5             81.1             82.0             83.3
## 6             28.5             81.9             82.7             84.1
## # ... with 6 more variables: `25th Percentile Stature (in
## #   centimeters)` <dbl>, `50th Percentile Stature (in centimeters)` <dbl>,
## #   `75th Percentile Stature (in centimeters)` <dbl>, `90th Percentile
## #   Stature (in centimeters)` <dbl>, `95th Percentile Stature (in
## #   centimeters)` <dbl>, `97th Percentile Stature (in centimeters)` <dbl>
```

5. Use the Howell1 data to fit the model from part 4. Interpret your parameter estimates.

Solution:

6. Fit a suitable (frequentist) regression model to the Howell1 data. Interpret your parameter estimates.

Solution:

7. Replace the 0/1 representing male and female in the Howell1 data to the character strings "male" and "female". Fit a suitable (frequentist) regression model using both age and sex as predictors. Interpret your parameter estimates. Is sex a statistically significant variable?

Solution:

8. Plot the models from questions 5, 6, and 7 on a scatter plot with height vs. age. Which model fits better? Justify your answer.

Solution:

9. Create a new scatter plot with all of the Howell1 data, color-coded by male/female. Would the models from questions 5, 6, and 7 be appropriate to apply to the full data set? Explain. Connect to what we discussed about prediction in Lecture 3 of the Applied Regression class.

Solution: