Bayesian Analysis - Assignment 2

Chris Hayduk

February 24, 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:

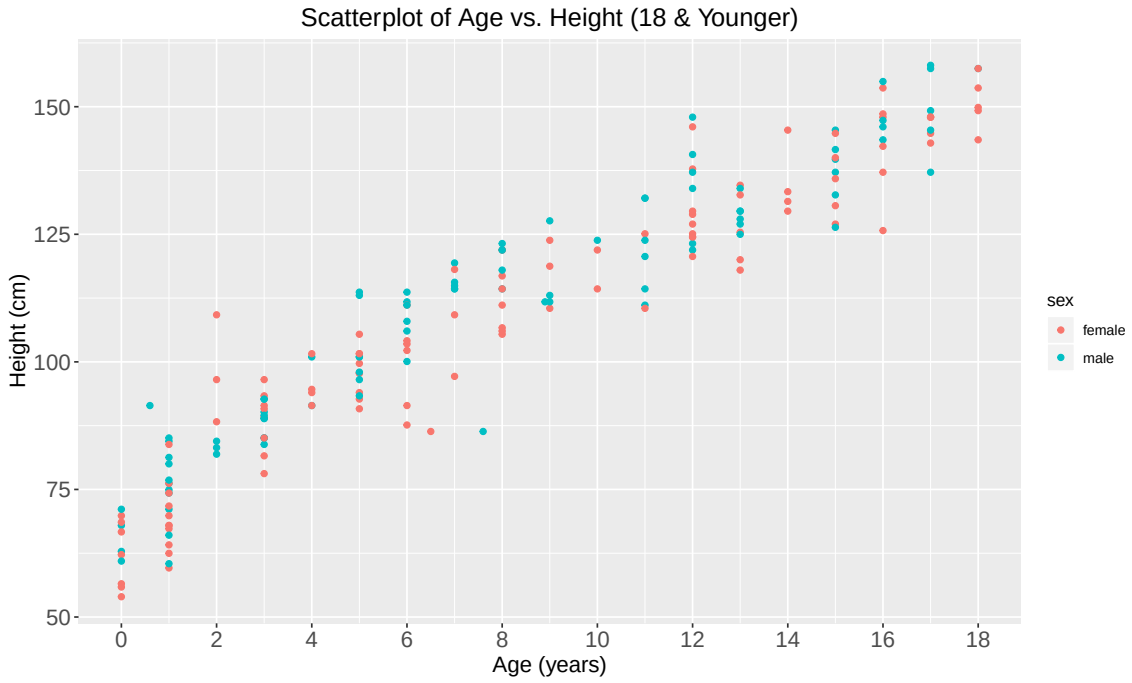
$$\text{height}_\text{i} \sim \text{Normal}(\mu_\text{i}, \sigma)$$
$$\mu_\text{i} = \alpha + \beta * \text{year}_\text{i}$$
$$\alpha \sim \text{Normal}(57, 25)$$
$$\beta \sim \text{Normal}(8.4, 4)$$
$$\sigma \sim \text{Uniform}(0, 50)$$

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 18 in order to derive the average amount that an individual must grow per year in order to reach the weighted average maximum height after starting from the weighted average minimum height. I also set the standard deviation for β to reflect our knowledge that $\beta \geq 0$.

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:

```
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 (18 & Younger)") +
  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 linear. Thus, we can still use the model as described in question 1.

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("Male Heights.csv")
girls_height_chart <- read_csv("Female Heights.csv")

names(boys_height_chart) <- c("age", "3rd_perc", "5th_perc",
                              "10th_perc", "25th_perc", "50th_perc",
                              "75th_perc", "90th_perc", "95th_perc",
                              "97th_perc")

names(girls_height_chart) <- c("age", "3rd_perc", "5th_perc",
                              "10th_perc", "25th_perc", "50th_perc",
                              "75th_perc", "90th_perc", "95th_perc",
                              "97th_perc")

#Assuming start age = 10 again
boys_height_at_0 <- boys_height_chart[boys_height_chart$age == 0.0,]$`50th_perc`
print(boys_height_at_0)

## [1] 49.98888

ages <- seq(from = 0.5, to = 216.5, by = 12)
heights <- rep(0, length(ages))
differences <- rep(NA, length(ages))
for(i in 1:length(ages)){
  height <- boys_height_chart[boys_height_chart$age == ages[i],]$`50th_perc`
  heights[i] <- height

  if(i != 1){
    differences[i] <- heights[i] - heights[i-1]
  }
}

boys_data <- data.frame("age" = ages, "height" = heights, "diff" = differences)

boys_growth_rate <- mean(boys_data$diff, na.rm = TRUE)

#Average growth rate for boys the three years from 10 - 13
print(boys_growth_rate)

## [1] 6.860501
```

Performing a similar analysis for girls:

```r
head(girls_height_chart)
```

```
## # A tibble: 6 x 10
##     age `3rd_perc` `5th_perc` `10th_perc` `25th_perc` `50th_perc`
##   <dbl>      <dbl>      <dbl>       <dbl>       <dbl>       <dbl>
## 1   0         45.1       45.6        46.3        47.7        49.3
## 2   0.5       47.5       48.0        48.7        50.1        51.7
## 3   1.5       51.0       51.5        52.3        53.7        55.3
## 4   2.5       53.6       54.2        55.0        56.5        58.1
## 5   3.5       55.9       56.4        57.3        58.8        60.5
## 6   4.5       57.8       58.4        59.3        60.8        62.5
## # ... with 4 more variables: `75th_perc` <dbl>, `90th_perc` <dbl>,
## #   `95th_perc` <dbl>, `97th_perc` <dbl>
```

```r
#Assuming start age = 10 again
girls_height_at_0 <- girls_height_chart[girls_height_chart$age == 0,]$`50th_perc`
print(girls_height_at_0)
```

```
## [1] 49.2864
```

```r
heights <- rep(0, length(ages))
differences <- rep(NA, length(ages))
for(i in 1:length(ages)){
  height <- girls_height_chart[girls_height_chart$age == ages[i],]$`50th_perc`
  heights[i] <- height

  if(i != 1){
    differences[i] <- heights[i] - heights[i-1]
  }

}

girls_data <- data.frame("age" = ages, "height" = heights, "diff" = differences)

girls_growth_rate <- mean(girls_data$diff, na.rm = TRUE)

#Average growth rate for girls the three years from 10 - 13
print(girls_growth_rate)
```

```
## [1] 6.191512
```

Now let's average the growth rate and initial height numbers:

```r
mean_growth_rate <- mean(c(girls_growth_rate, boys_growth_rate))

print(mean_growth_rate)

## [1] 6.526007

mean_starting_height <- mean(c(girls_height_at_0, boys_height_at_0))

print(mean_starting_height)

## [1] 49.63764
```

Thus, in order to predict height using age, we will use the following model:

$$\text{height}_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta * \text{year}_i$$
$$\alpha \sim \text{Normal}(49.6, 5)$$
$$\beta \sim \text{Normal}(6.5, 3)$$
$$\sigma \sim \text{Uniform}(0, 50)$$

The 3rd percentile for girls at 0 years old (ie. as a newborn) measured 45.1 cm. Thus, it is unlikely that α will fall below 45.1 cm. As a result, I chose the standard deviation of the prior to be 5 cm. Since the prior is normally distributed, this means that I expect the α term to be within (39.6 cm, 59.6 cm) with about 95% certainty.

I chose the standard deviation of the prior for β to be 3 cm. The correlation between age and height is clearly positive. Thus, it is highly unlikely that β should be negative. A standard deviation of 3 cm means that I expect β to be within (0.5 cm, 12.5 cm) with about 95% certainty.

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

Solution:

```
m1 <- map(
  alist(
    height ~ dnorm(a + b*age, sigma),
    a ~ dnorm(49.6, 5),
    b ~ dnorm(6.5, 3),
    sigma ~ dunif(0, 50)
  ), data = under_18)

print(m1)
```

```
##
## Maximum a posteriori (MAP) model fit
##
## Formula:
## height ~ dnorm(a + b * age, sigma)
## a ~ dnorm(49.6, 5)
## b ~ dnorm(6.5, 3)
## sigma ~ dunif(0, 50)
##
## MAP values:
##         a         b     sigma
## 72.470862  4.585416  8.471086
##
## Log-likelihood: -704.01
```

The intercept term, a, represents the average height when age = 0. In this case, the intercept is abou 72.5 cm. The slope, b, represents the average increase in height for a 1 year increase in age. In this case, height increases by about 4.5 cm for every 1 year of age.

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

Solution:

```
linear.model1 <- lm(height ~ age, data = under_18)

summary(linear.model1)

##
## Call:
## lm(formula = height ~ age, data = under_18)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3301  -5.0108  -0.2407   5.5884  26.7113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.5153     1.0610   69.29   <2e-16 ***
## age           4.4967     0.1086   41.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.493 on 196 degrees of freedom
## Multiple R-squared:  0.8974,Adjusted R-squared:  0.8969
## F-statistic:  1714 on 1 and 196 DF,  p-value: < 2.2e-16
```

The intercept (when age = 0) here is about 73.5 cm. The slope is about 4.5 cm. These numbers are very similar to the values obtained using our Bayesian model above. Thus, it seems that the priors did not do much to inform the model.

Now let's attempt fitting the model using matrices:

```
y <- under_18$height

intercept <- matrix(rep(1, length(under_18$age)), ncol = 1)

x_var <- matrix(under_18$age, ncol = 1)

x <- cbind(intercept, x_var)

beta_hat <- solve((t(x) %*% x)) %*% t(x) %*% y

print(beta_hat)

##           [,1]
## [1,] 73.515325
## [2,]  4.496675
```

As you can see, matrix algebra leads to the same result for the slope and intercept as the lm() function.

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:

```
linear.model2 <- lm(height ~ age + sex, data = under_18)

summary(linear.model2)

##
## Call:
## lm(formula = height ~ age + sex, data = under_18)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2347  -4.9338  -0.3577   5.4779  28.4815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.7192     1.1958  59.978  < 2e-16 ***
## age           4.5096     0.1065  42.348  < 2e-16 ***
## sexmale       3.6023     1.1858   3.038  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.321 on 195 degrees of freedom
## Multiple R-squared:  0.902,Adjusted R-squared:  0.901
## F-statistic: 897.6 on 2 and 195 DF,  p-value: < 2.2e-16
```

In this case, the intercept represents the average height of females at age = 0 (about 71.72 cm). When considering males, we add the coefficient for sexmale to the intercept. Thus, the average height of males at age = 0 is 71.7192 cm + 3.6023 cm = 75.3215 cm. The slope has the same interpretation as before: the average increase in height for a 1 year increase in age. In this case, height increases by about 4.5 cm for every 1 year of age. All three terms in this model are statistically significant at significance level $\alpha = 0.01$.

Now let's try using an interaction term:

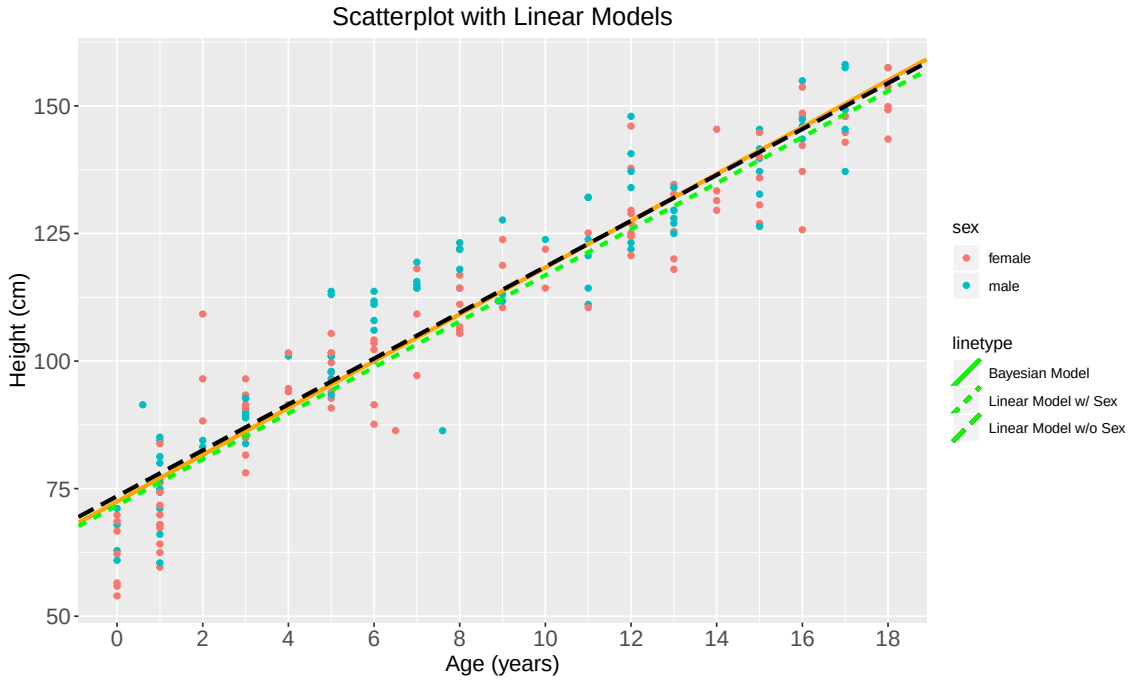
```
linear.model3 <- lm(height ~ age*sex, data = under_18)

summary(linear.model3)

##
## Call:
## lm(formula = height ~ age * sex, data = under_18)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2392  -4.9959  -0.4058   5.5187  28.5904
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.57539    1.42731  50.147   <2e-16 ***
## age          4.52708    0.14224  31.828   <2e-16 ***
## sexmale      3.92163    2.09090   1.876   0.0622 .
## age:sexmale -0.03996    0.21524  -0.186   0.8529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.341 on 194 degrees of freedom
## Multiple R-squared:  0.902,Adjusted R-squared:  0.9005
## F-statistic: 595.4 on 3 and 194 DF,  p-value: < 2.2e-16
```

The estimates for the intercept, age, and sexmale are all quite similar to the previous model. Furthermore, sexmale is no longer statistically significant and the interaction term both is not statisticall significant either. The previous model also had a slightly higher adjusted R-squared value. As a result, we should favor the previous model.

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:

```
sp + geom_abline(aes(intercept= coef(m1)[1],
                     slope=coef(m1)[2],
                     linetype = "Bayesian Model"),
                     color="orange",
                     size = 1.25) +
  geom_abline(aes(intercept= coef(linear.model1)[1],
                     slope=coef(linear.model1)[2],
                     linetype = "Linear Model w/o Sex"),
                     color="black",
                     size = 1.25) +
  geom_abline(aes(intercept= coef(linear.model2)[1],
                     slope=coef(linear.model2)[2],
                     linetype = "Linear Model w/ Sex"),
                     color="green",
                     size = 1.25) +
  ggtitle("Scatterplot with Linear Models")
```

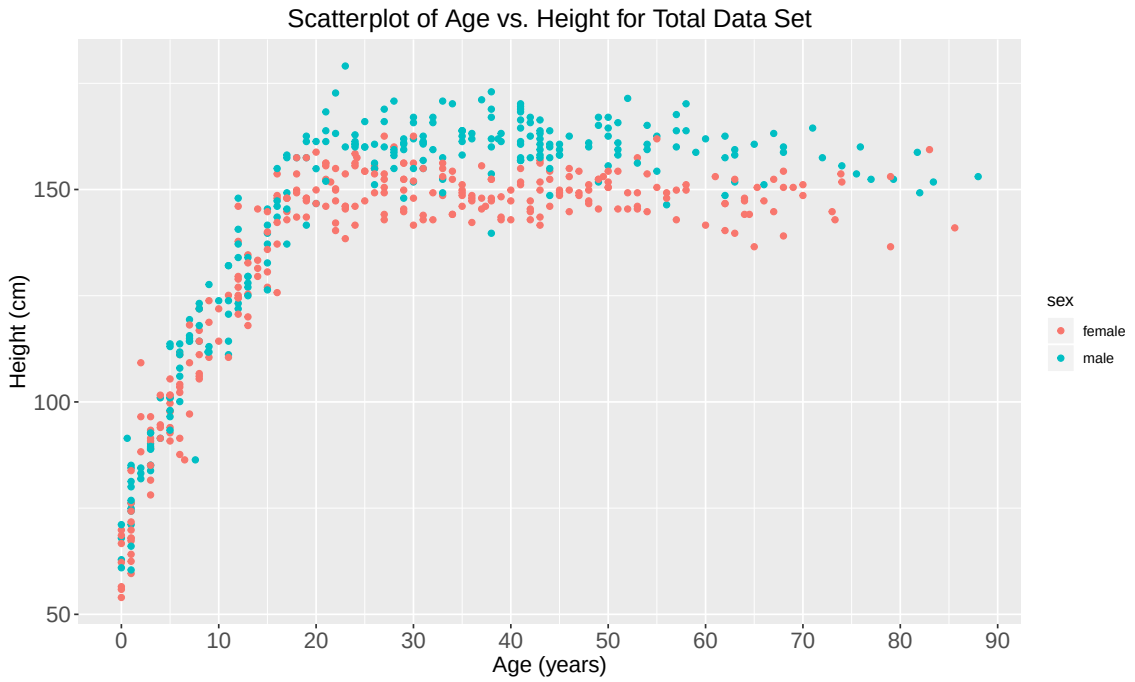


The three linear models fit the data almost identically.

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.

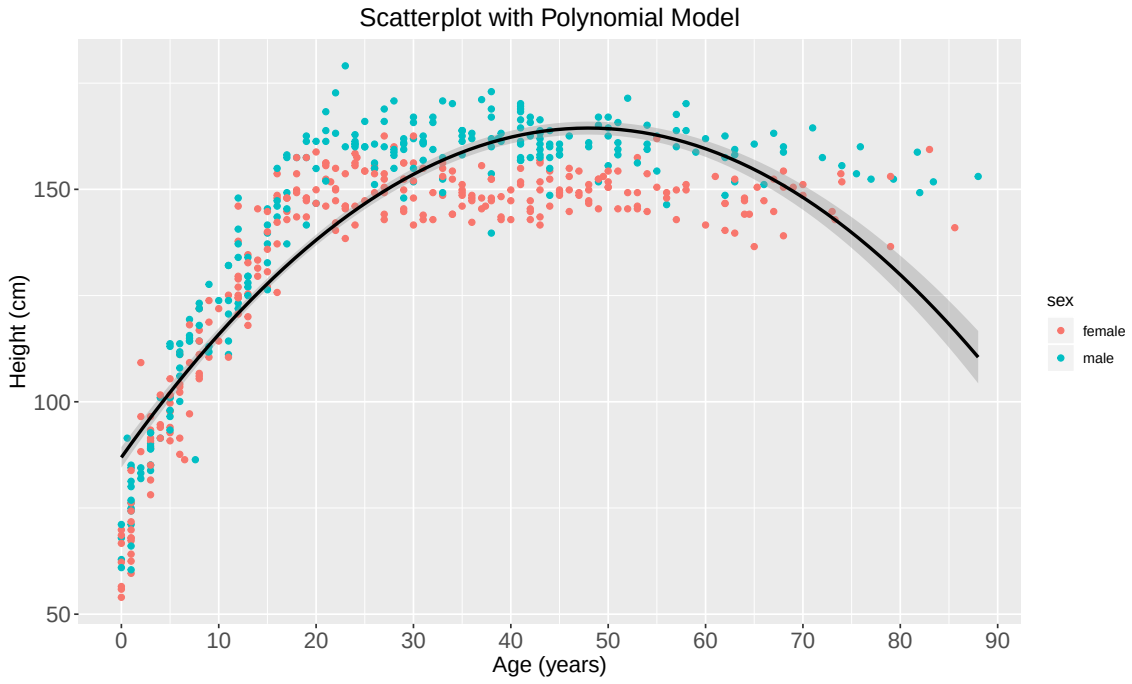
Solution:

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

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

sp2 <- ggplot(data, aes(x=age, y=height)) +
  geom_point(aes(color=sex)) +
  ggtitle("Scatterplot of Age vs. Height for Total Data Set") +
  labs(x = "Age (years)", y = "Height (cm)") +
  scale_x_continuous(breaks = pretty(data$age, n = 7)) +
  theme.info

sp2
```



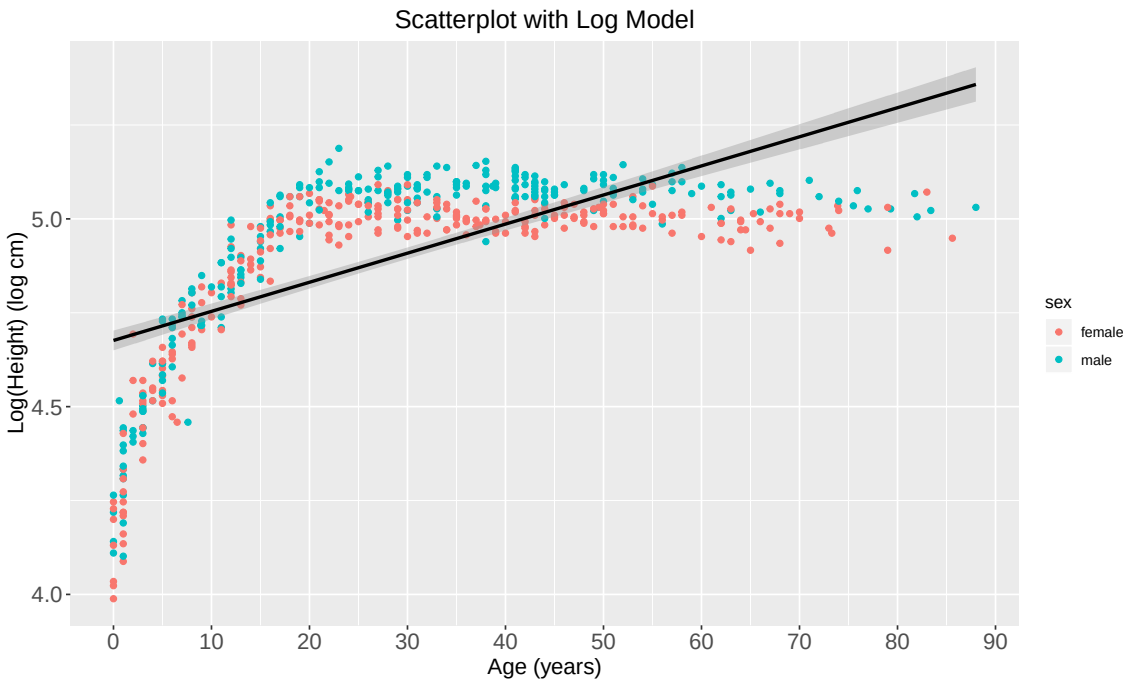
The three linear models would not be appropriate to apply to the full data set because the data is highly non-linear since people stop growing after a certain age. We could add a quadratic term in order to attempt to adequately fit the full data set. Here is an example of this:

```r
poly_model <- lm(height ~ age + I(age^2), data = data)

summary(poly_model)

##
## Call:
## lm(formula = height ~ age + I(age^2), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.931  -8.836   0.070   8.109  42.535
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 86.905857   1.233587   70.45   <2e-16 ***
## age          3.231271   0.081550   39.62   <2e-16 ***
## I(age^2)    -0.033672   0.001123  -29.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.37 on 541 degrees of freedom
## Multiple R-squared:  0.7999,Adjusted R-squared:  0.7992
## F-statistic:  1081 on 2 and 541 DF,  p-value: < 2.2e-16

sp2 + stat_smooth(aes(y=height), method = "lm",
                  formula = y ~ x + I(x^2), size = 1,
                  color="black") +
  ggtitle("Scatterplot with Polynomial Model")
```



This still does not appear to fit the data properly. Let's try a log-model instead:

```
data$log_height <- log(data$height)
log_model <- lm(log_height ~ age, data = data)

summary(log_model)

##
## Call:
## lm(formula = log_height ~ age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68777 -0.09971  0.03904  0.13462  0.33331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.6762931  0.0134346  348.08   <2e-16 ***
## age         0.0077467  0.0003739   20.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1808 on 542 degrees of freedom
## Multiple R-squared:  0.4419,Adjusted R-squared:  0.4409
## F-statistic: 429.2 on 1 and 542 DF,  p-value: < 2.2e-16

ggplot(data, aes(x= age, y=log_height)) +
  geom_point(aes(color=sex)) +
  ggtitle("Scatterplot with Log Model") +
  labs(x = "Age (years)", y = "Log(Height) (log cm)") +
  scale_x_continuous(breaks = pretty(data$age, n = 7)) +
  theme.info +
  stat_smooth(aes(y=log_height), method = "lm",
                  formula = y ~ x, size = 1,
                  color="black")
```



It seems we may have been better off with the quadratic model judging by the graphs and the Adjusted R-squared values.