Galton’s Height Data

- a. (1 point)
- b. (2 points)
- c. (5 points)
- d. (2 points)
- e. (4 points)
- f. (4 points)
- g. (4 points)
- h. (4 points)
- i. (2 points)
- j. (8 points)
- k. (4 points)
- l. (4 points)
- m. (2 points)

Note: There are several additional analyses and calculations in this write-up (begin with **Aside**:…). You are not required to do these calculations, but are strongly encouraged to look over them.

Sir Francis Galton (1822–1911) was an English statistician. He founded many concepts in statistics, such as correlation, quartile, percentile and regression, that are still being used today.

In this R markdown exercise, you are going to analyze the famous Galton data on the heights of parents and their children. The data were collected in the late 19th century in England. He coined the term regression towards mediocrity (http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf) to describe the result of his linear model. (Note that the paper was written in 1886. The “computer” mentioned in the paper was actually a person whose job was to do number crunching.) Surprisingly, Galton’s analysis is still useful today (see e.g. Predicting height: the Victorian approach beats modern genomics (https://www.wired.com/2009/03/predicting-height-the-victorian-approach-beats-modern-genomics/), Predicting human height by Victorian and genomic methods (https://www.nature.com/articles/ejhg20095)).

Galton’s height data can be download here (Galton.txt) (right click and choose Save Link As…). The description of the data can be found on this webpage (GaltonData.html). Note that this is not a csv file. You need to use the `read.table()` function with appropriate parameters to load the data correctly to R.

a. (1 point)

Calculate the correlation matrix between `Height` , `Father` and `Mother` .

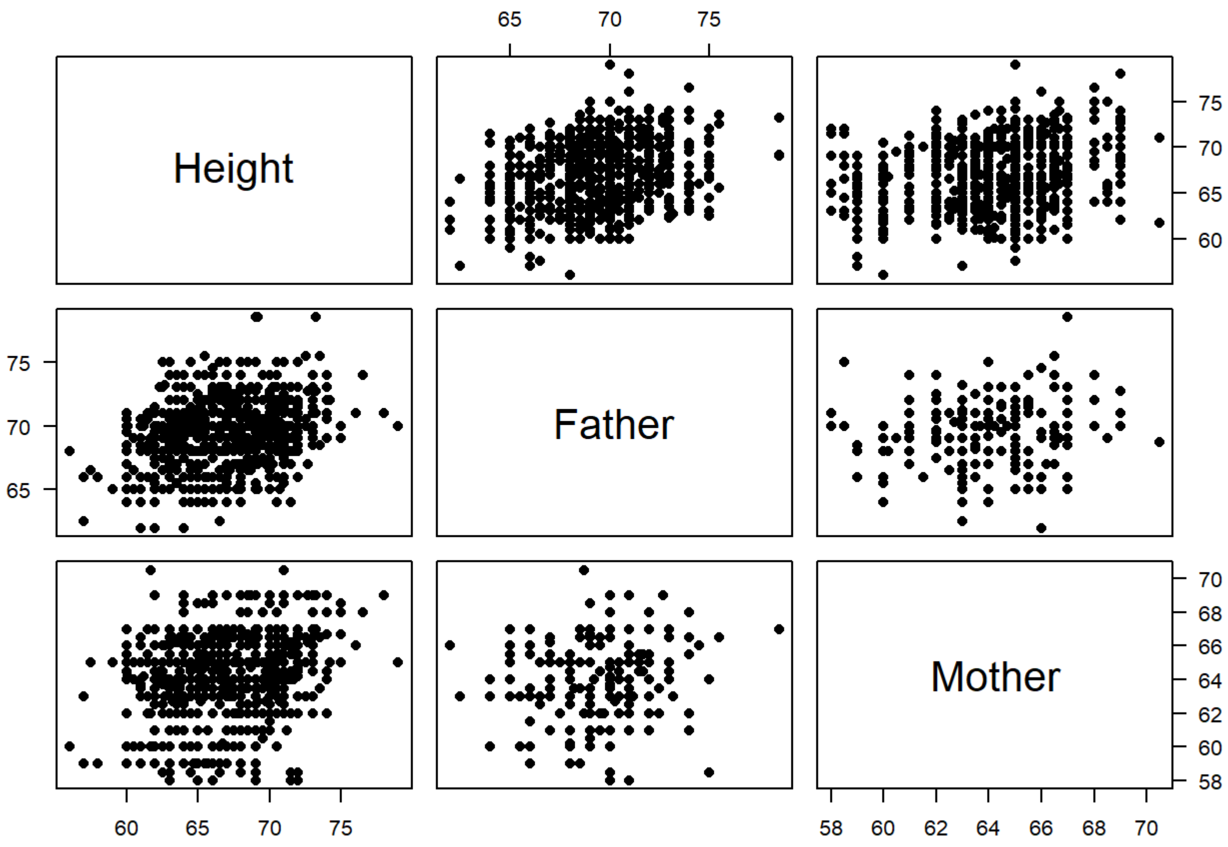
```
# Load data
galton <- read.table("Galton.txt", header=TRUE)

# correlation matrix
cor(galton[,c("Height","Father","Mother")])
```

```
          Height     Father     Mother
Height 1.0000000 0.27535483 0.20165489
Father 0.2753548 1.00000000 0.07366461
Mother 0.2016549 0.07366461 1.00000000
```

b. (2 points)

Use the `pairs()` function to create a matrix of scatterplots of the columns `Height`, `Father` and `Mother`. This is a graphical representation of the correlation matrix calculated above. (Hint: You need to subset the data frame to pull the 3 columns and then pass them to the `pairs()` function.)

```
# Matrix of scatterplots
pairs(galton[,c("Height","Father","Mother")], pch=16, las=1)
```



c. (5 points)

Fit a multiple regression model predicting children’s height (`Height`) from father’s height (`Father`), mother’s height (`Mother`), and gender (`Gender`). In other words, the model should contain the following terms:

$$\hat{H}_{children} = \beta_0 + \beta_1 f_{gender} + \beta_2 H_{father} + \beta_3 H_{mother},$$

where $\hat{H}_{children}$ is the predicted height (in inches) of the adult children, H_{father} and H_{mother} are the height (in inches) of the father and mother, respectively. f_{gender} is a binary variable: f_{gender}=0 for males and f_{gender}=1 for females. (4 pts)

`galton$Gender` is a factor variable with the reference level set to “F” (you can type `levels(galton$Gender)` to confirm it). We want to set the reference level to “M” to be consistent with the f_{gender} variable.

Note: Starting from R 4.0, the `stringsAsFactors` option in `read.table()` defaults to FALSE, so the above statement is no longer true for R 4.0 and later. `galton$Gender` is now a character factor and we need to turn it to a factor variable.

```
# Turn 'Gender' column into a factor
galton$Gender <- factor(galton$Gender, levels = c('M','F'))

# Fit model
fit_mult <- lm(Height ~ Gender+Father+Mother, data=galton)

summary(fit_mult)
```

```

Call:
lm(formula = Height ~ Gender + Father + Mother, data = galton)

Residuals:
   Min     1Q Median     3Q    Max 
-9.523 -1.440  0.117  1.473  9.114 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.57071    2.74067   7.506 1.48e-13 ***
GenderF     -5.22595    0.14401 -36.289  < 2e-16 ***
Father       0.40598    0.02921  13.900  < 2e-16 ***
Mother       0.32150    0.03128  10.277  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.154 on 894 degrees of freedom
Multiple R-squared:  0.6397,	Adjusted R-squared:  0.6385 
F-statistic:   529 on 3 and 894 DF,  p-value: < 2.2e-16
```

Hence, the regression equation is

$$\hat{H}_{children} = 20.57 -5.226\mathrm{f}_{\mathrm{gender}} + 0.406\mathrm{H}_{\mathrm{father}} + 0.3215\mathrm{H}_{\mathrm{mother}}$$

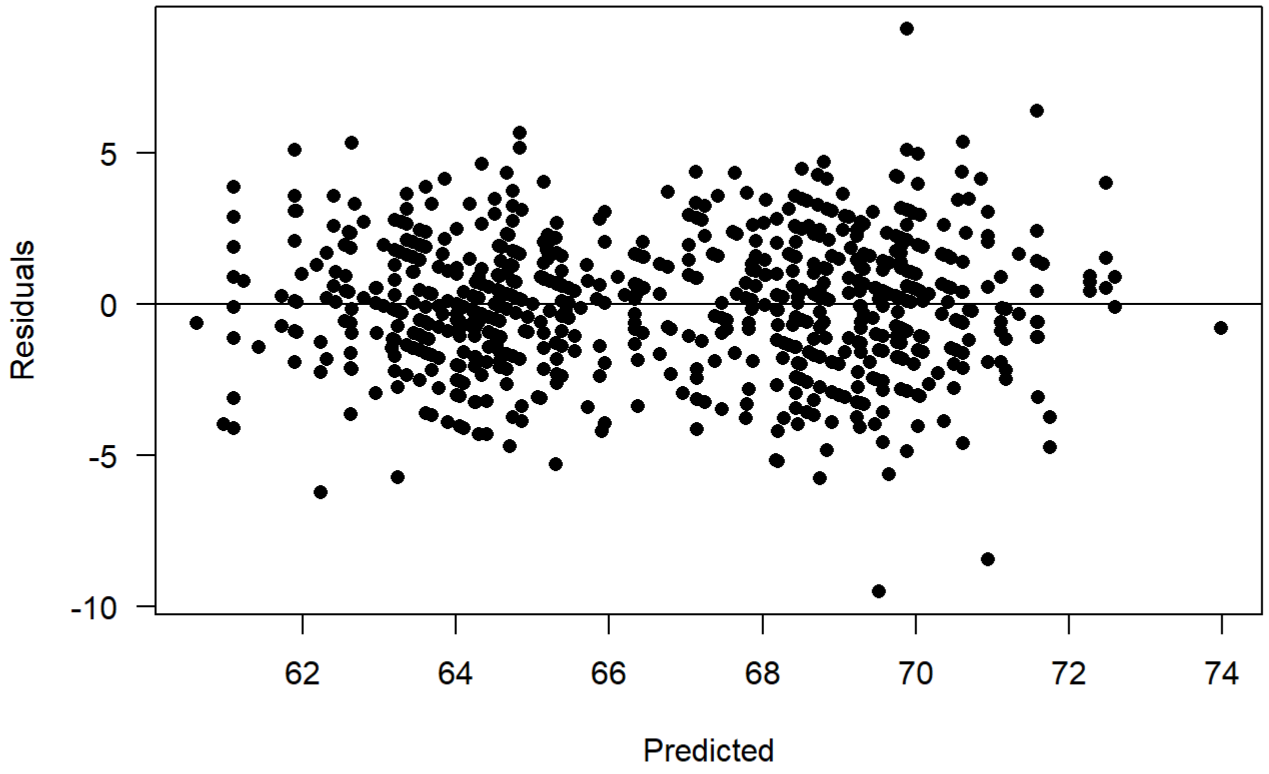
Which slopes are significant (at the 5% level)? (1 pt)

From the summary, we see that all slopes have p-values less than 2×10^{-16}. So all slopes are highly significant.

d. (2 points)

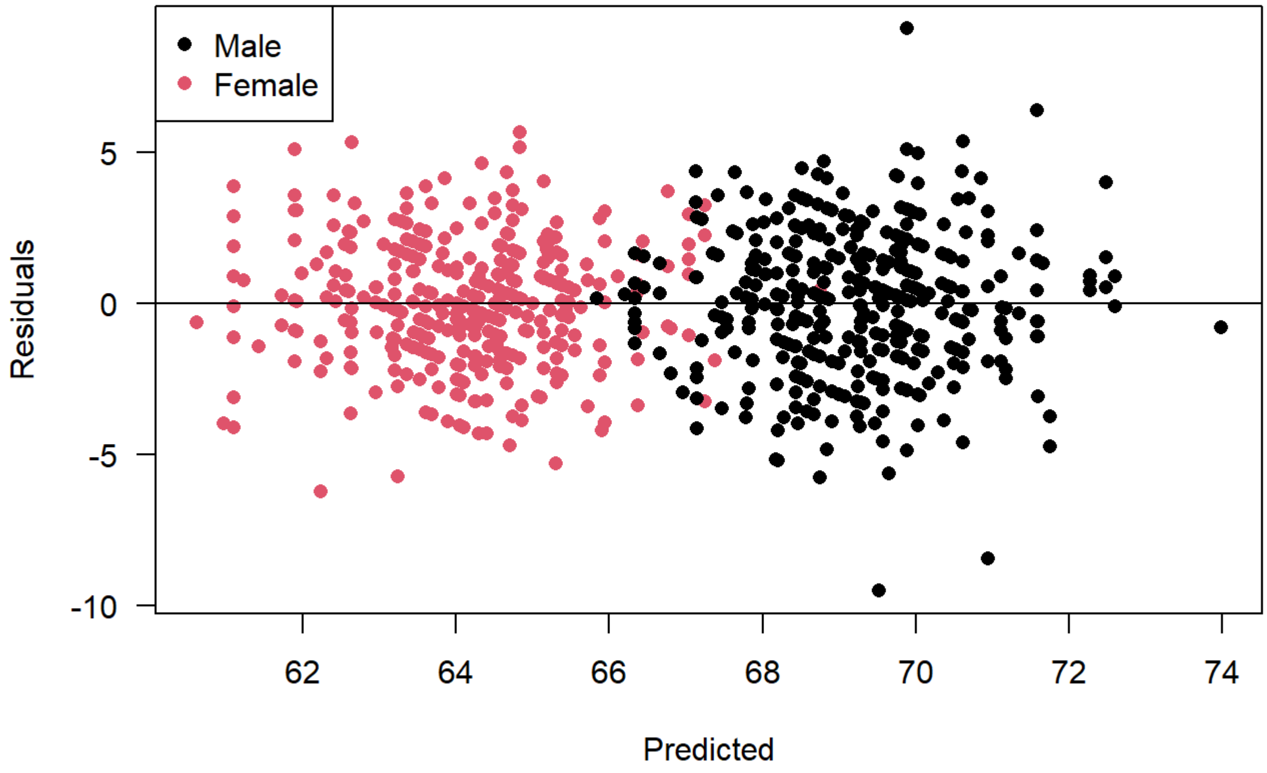
Plot the residuals versus the fitted values for the multiple regression model above.

```
plot(fit_mult$residuals ~ fit_mult$fitted.values,xlab="Predicted", ylab="Resid
uals",
     pch=16, las=1)
abline(h=0)
```



Aside: We can clearly see two clusters of points here. The cluster of points with smaller predicted heights belong to the female children and the other cluster of points belong to the male children. To see this, we can color-code the residuals by gender:

```
plot(fit_mult$residuals ~ fit_mult$fitted.values, col=galton$Gender,
     xlab="Predicted", ylab="Residuals", pch=16, las=1)
abline(h=0)
legend("topleft",c("Male","Female"),pch=16,col=1:2)
```



Instead of fitting a multiple regression model, Galton constructed a simple model predicting children's height from parents' heights. However, he first had to deal with the gender difference between male and female heights.

e. (4 points)

Calculate the means of Father's and Mother's heights in the data set. Then show that Father's mean height is about 8% higher than Mother's mean height. (2 pts)

We can use the `colMeans()` function (see Section 18.8 of Peng's textook):

```
(parents_avg <- colMeans(galton[,c("Father","Mother")]))
```

```
  Father   Mother 
69.23285 64.08441 
```

We see that Father's mean height is 69.2 inches and Mother's mean height is 64.1 inches.

Ratio of Father's mean height and Mother's mean height:

```
parents_avg["Father"]/parents_avg["Mother"]
```

```
  Father 
1.080338 
```

This shows that Father's mean height is about 8% higher than Mother's mean height.

Calculate the mean heights of the adult male and female children in the data set. Then show that male children's mean height is also about 8% higher than female children's mean height. (2 pts)

We can use the `tapply()` function to compute the group means:

```
(children_avg <- tapply(galton$Height, galton$Gender,mean))
```

```
       M        F 
69.22882 64.11016 
```

We see that male children's mean height is 69.2 inches and female children's mean height is 64.1 inches.

Ratio of male children's mean height and female children's mean height:

```
children_avg["M"]/children_avg["F"]
```

```
       M 
1.079842 
```

This shows that male children's mean height is about 8% higher than female children's mean height.

Aside: I use `colMeans()` and `tapply()` to compute the means because I find them convenient. You can of course calculate the means one by one: `mean(galton$Father)`, `mean(galton$Mother)`, `mean(galton$Height[galton$Gender=="M"])`, `mean(galton$Height[galton$Gender=="F"])`.

---

f. (4 points)

Calculate the medians of Father's and Mother's heights in the data set. Then show that Father's median height is about 8% higher than Mother's median height. (2 pts)

There is no `colMedians()` function, so we use `apply()` on the columns (margin=2):

```
(parents_med <- apply(galton[,c("Father","Mother")], 2, median))
```

```
Father Mother
    69     64
```

We see that Father’s median height is 69 inches and Mother’s median height is 64 inches.

Ratio of Father’s median height and Mother’s median height:

```
parents_med["Father"]/parents_med["Mother"]
```

```
  Father
1.078125
```

This shows that Father’s median height is about 8% higher than Mother’s median height.

Calculate the median heights of the adult male and female children in the data set. Then show that male children’s median height is also about 8% higher than female children’s median height. (2 pts)

We can use the `tapply()` function to compute the group medians:

```
(children_med <- tapply(galton$Height, galton$Gender,median))
```

```
   M    F
69.2 64.0
```

We see that male children’s median height is 69.2 inches and female children’s median height is 64 inches.

Ratio of male children’s median height and female children’s median height:

```
children_med["M"]/children_med["F"]
```

```
      M
1.08125
```

This shows that male children’s median height is about 8% higher than female children’s median height.

Another way to get the means and medians is to use the `summary()` function:

```
summary(galton[,c("Father","Mother")])
```

```
     Father          Mother     
 Min.   :62.00   Min.   :58.00  
 1st Qu.:68.00   1st Qu.:63.00  
 Median :69.00   Median :64.00  
 Mean   :69.23   Mean   :64.08  
 3rd Qu.:71.00   3rd Qu.:65.50  
 Max.   :78.50   Max.   :70.50  
```

```
tapply(galton$Height, galton$Gender, summary)
```

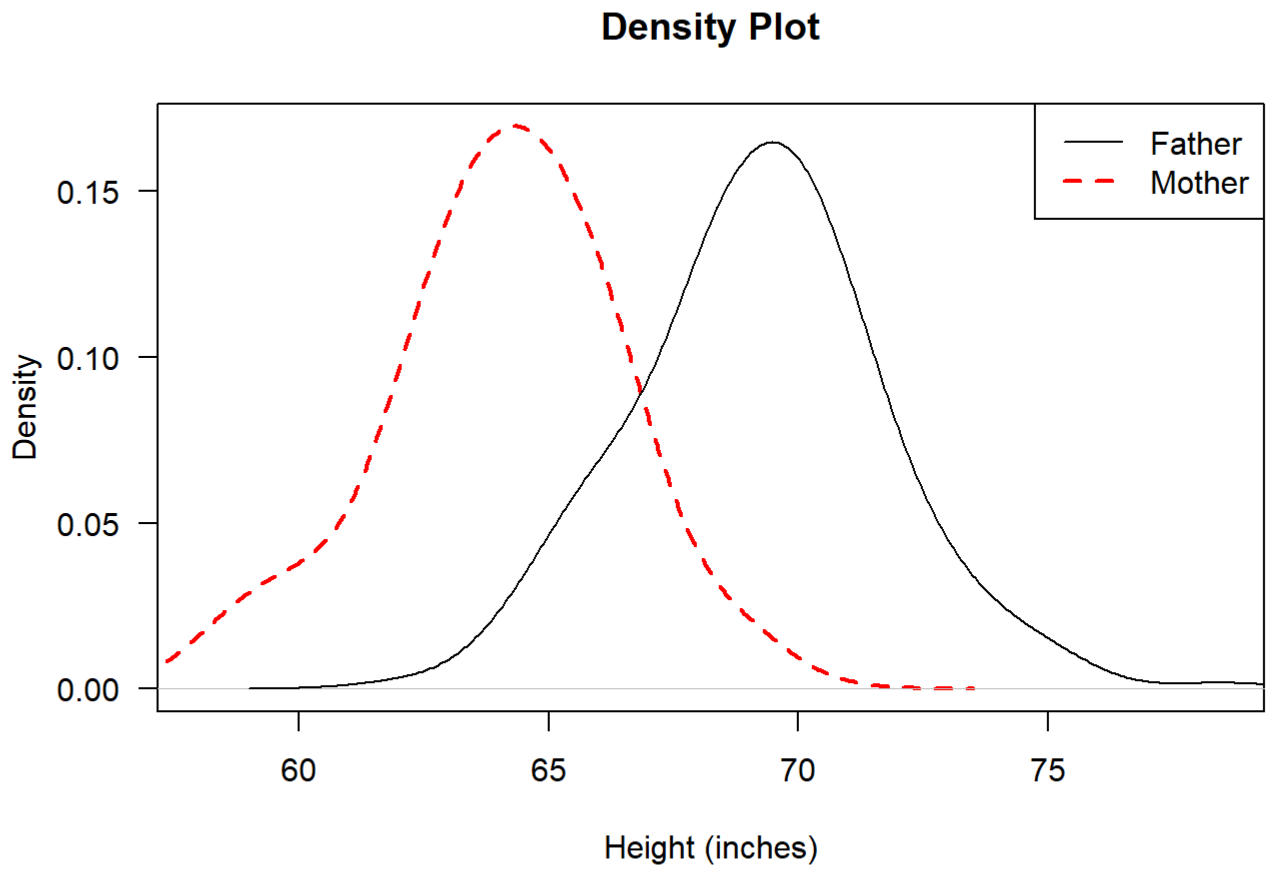
```
$M
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  60.00   67.50   69.20   69.23   71.00   79.00 

$F
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  56.00   62.50   64.00   64.11   65.50   70.50 
```

You also see the other quartiles.

Aside: In addition to looking at the averages of the male and female heights in the data, another good way to look at the gender difference is to look at some plots.

First, we look at the density plots of Father’s and Mother’s heights:

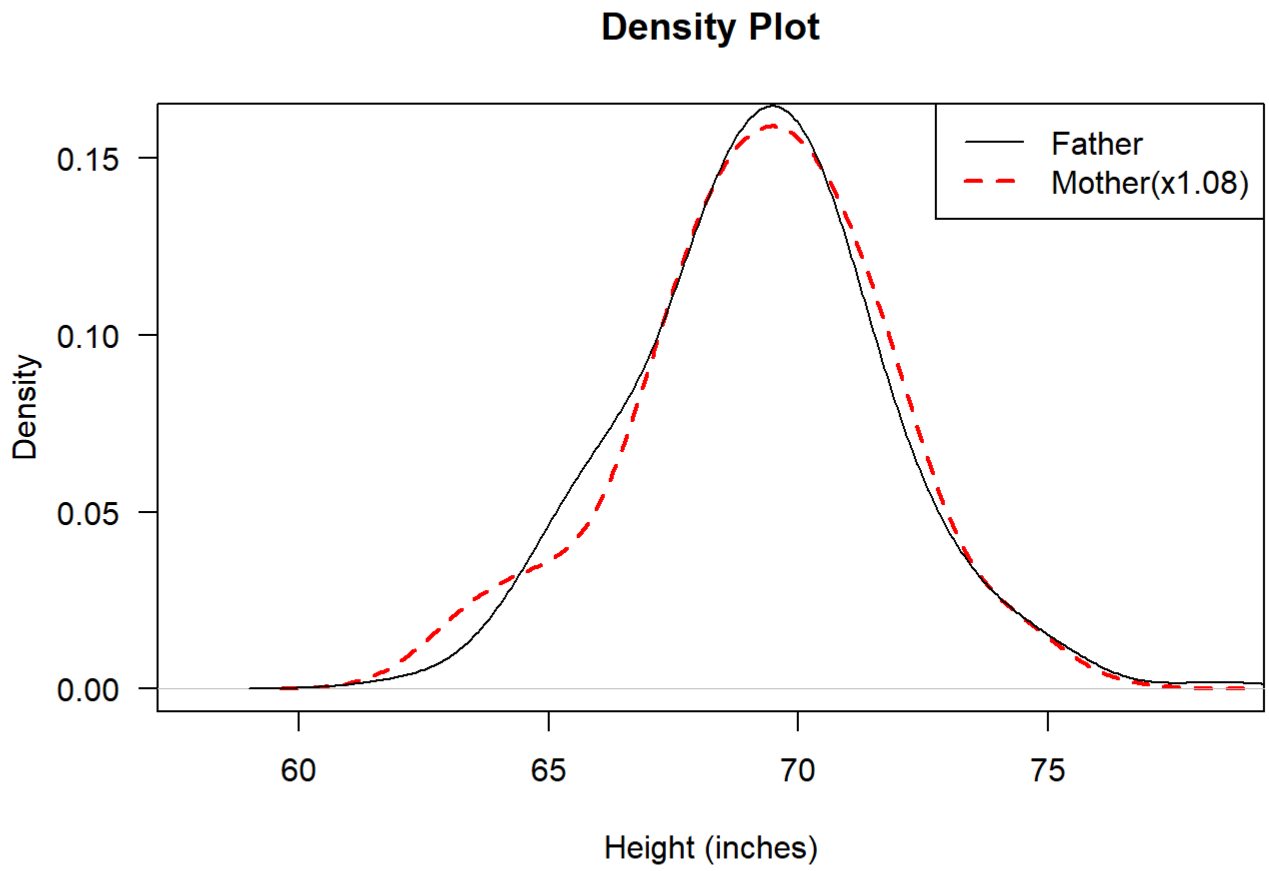
```
plot(density(galton$Mother, bw=1), xlab="Height (inches)", main="Density Plot"
, las=1,
     xlim=c(min(galton$Mother),max(galton$Father)),lty=2,lwd=2, col="red")
lines(density(galton$Father, bw=1))
legend("topright", c("Father","Mother"), col=c("black","red"), lty=1:2, lwd=1:
2)
```



Note that I set the smoothing parameter `bw=1` to smooth the density curve on the scale of an inch since the height data are accurate to only about an inch. The command `lines()` is used to add a line on an existing plot.

The difference between Father's and Mother's heights is apparent in the plot above. Now let's see what happens if we multiply Mother's height by 1.08.

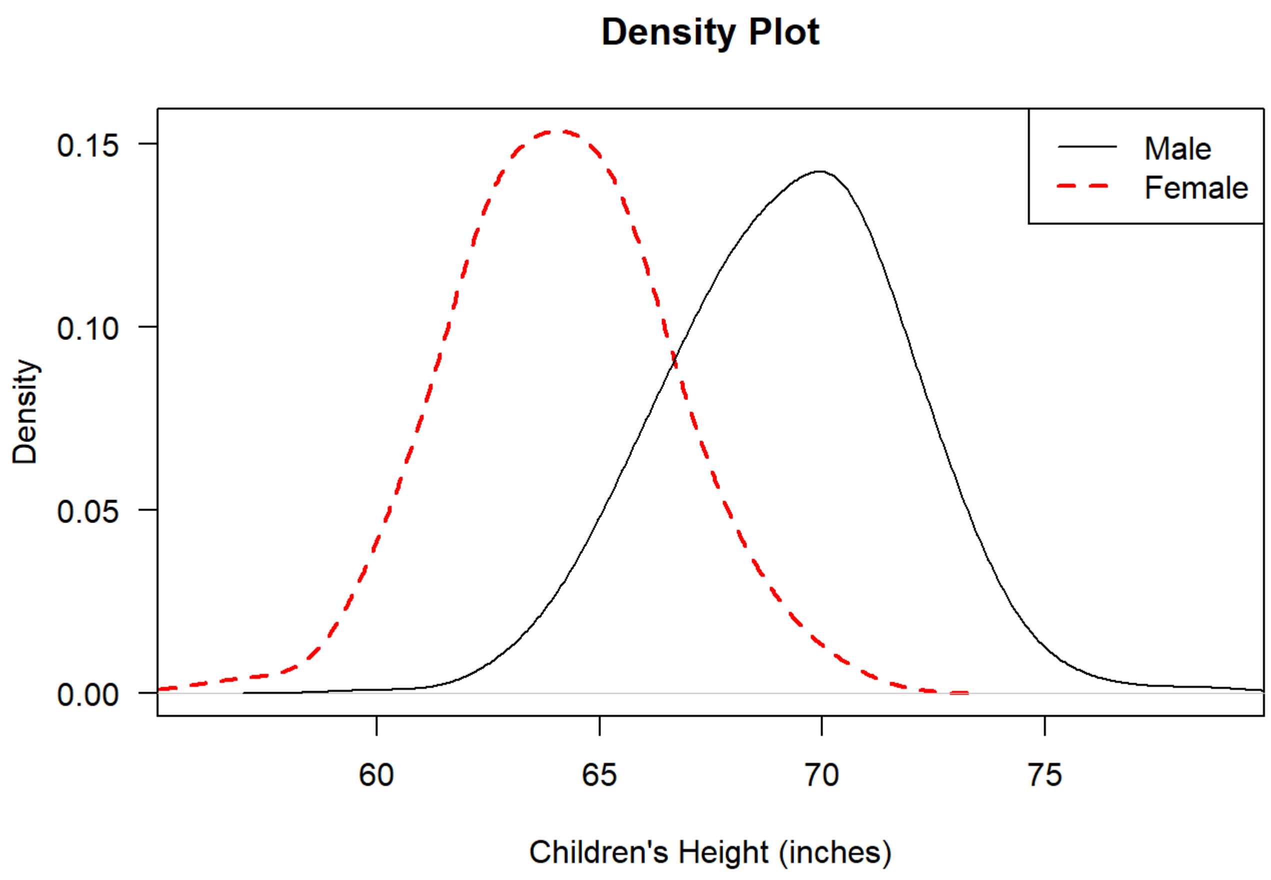
```
plot(density(galton$Mother*1.08, bw=1), xlab="Height (inches)", main="Density
 Plot", las=1,
     xlim=c(min(galton$Mother),max(galton$Father)),lty=2,lwd=2, col="red")
lines(density(galton$Father, bw=1))
legend("topright", c("Father","Mother(x1.08)"), col=c("black","red"), lty=1:2,
lwd=1:2)
```



Now, the gender difference is much reduced.

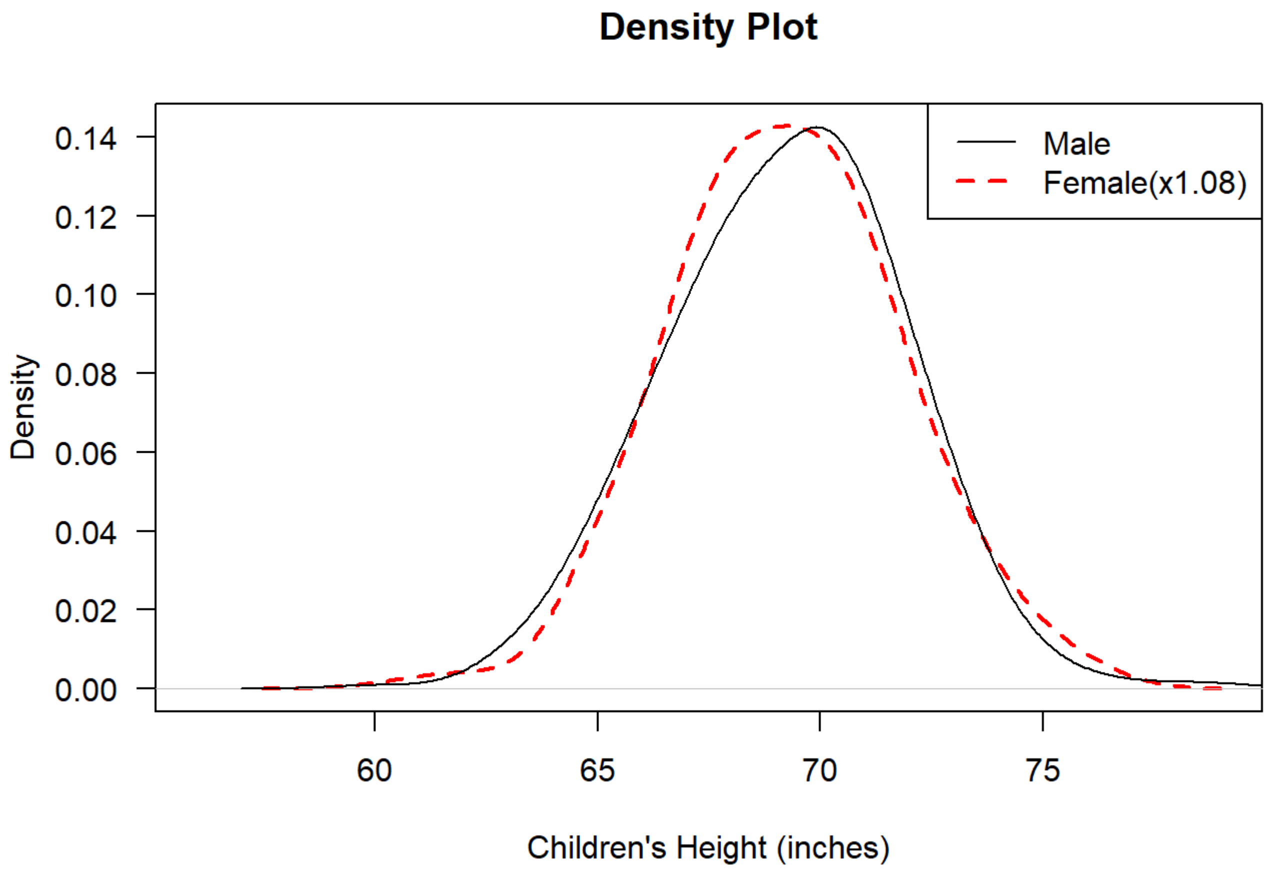
Next, we look at the male and female heights of the children.

```
plot(density(galton$Height[galton$Gender=="F"], bw=1), xlab="Children's Height (inches)",
     main="Density Plot", las=1, xlim=c(min(galton$Height),max(galton$Height)),
     lty=2,lwd=2, col="red")
lines(density(galton$Height[galton$Gender=="M"], bw=1))
legend("topright", c("Male","Female"), col=c("black","red"), lty=1:2, lwd=1:2)
```



The gender difference is clearly shown. Now multiply the female heights by 1.08:

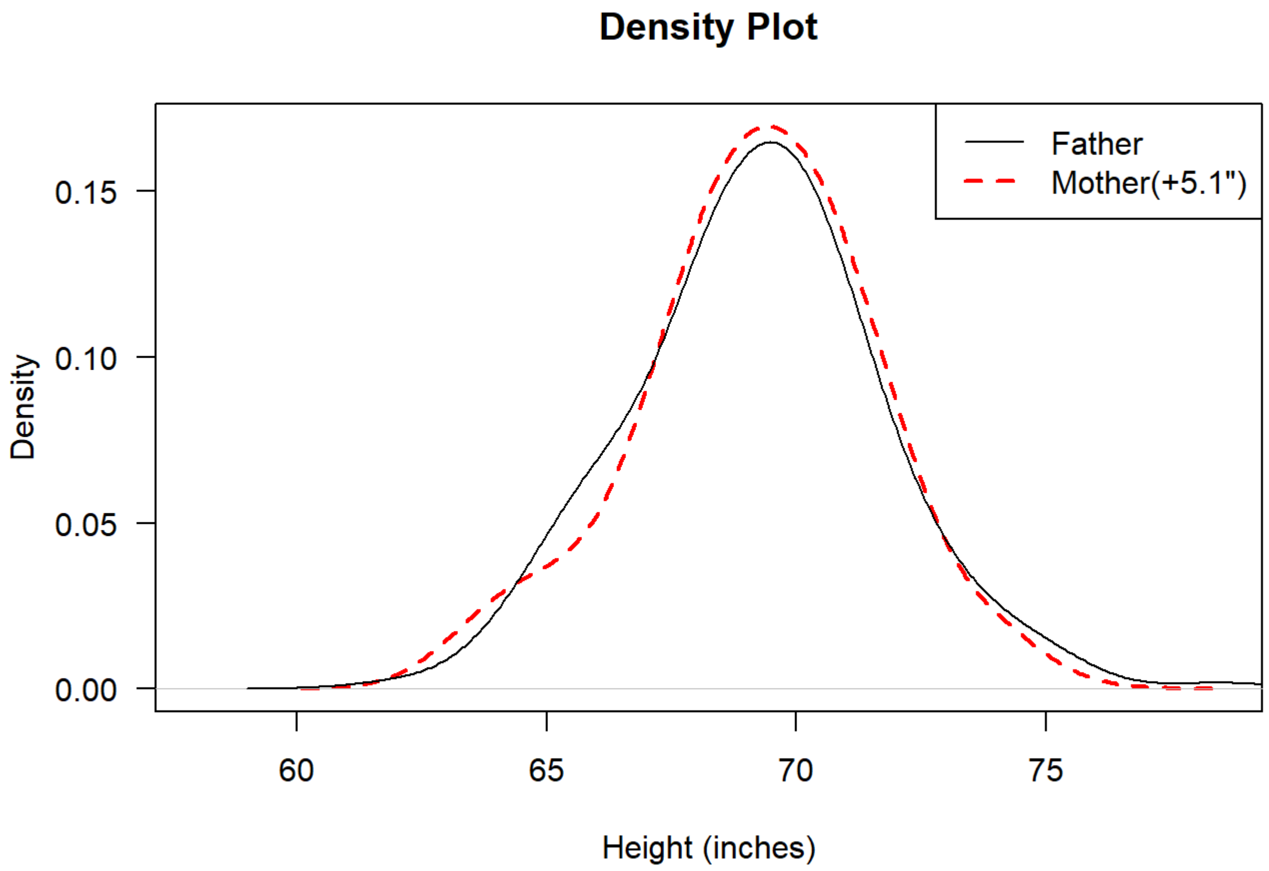
```
plot(density(galton$Height[galton$Gender=="F"]*1.08, bw=1),
     xlab="Children's Height (inches)", main="Density Plot",
     las=1, xlim=c(min(galton$Height),max(galton$Height)),
     lty=2,lwd=2, col="red")
lines(density(galton$Height[galton$Gender=="M"], bw=1))
legend("topright", c("Male","Female(x1.08)"), col=c("black","red"), lty=1:2, l
wd=1:2)
```



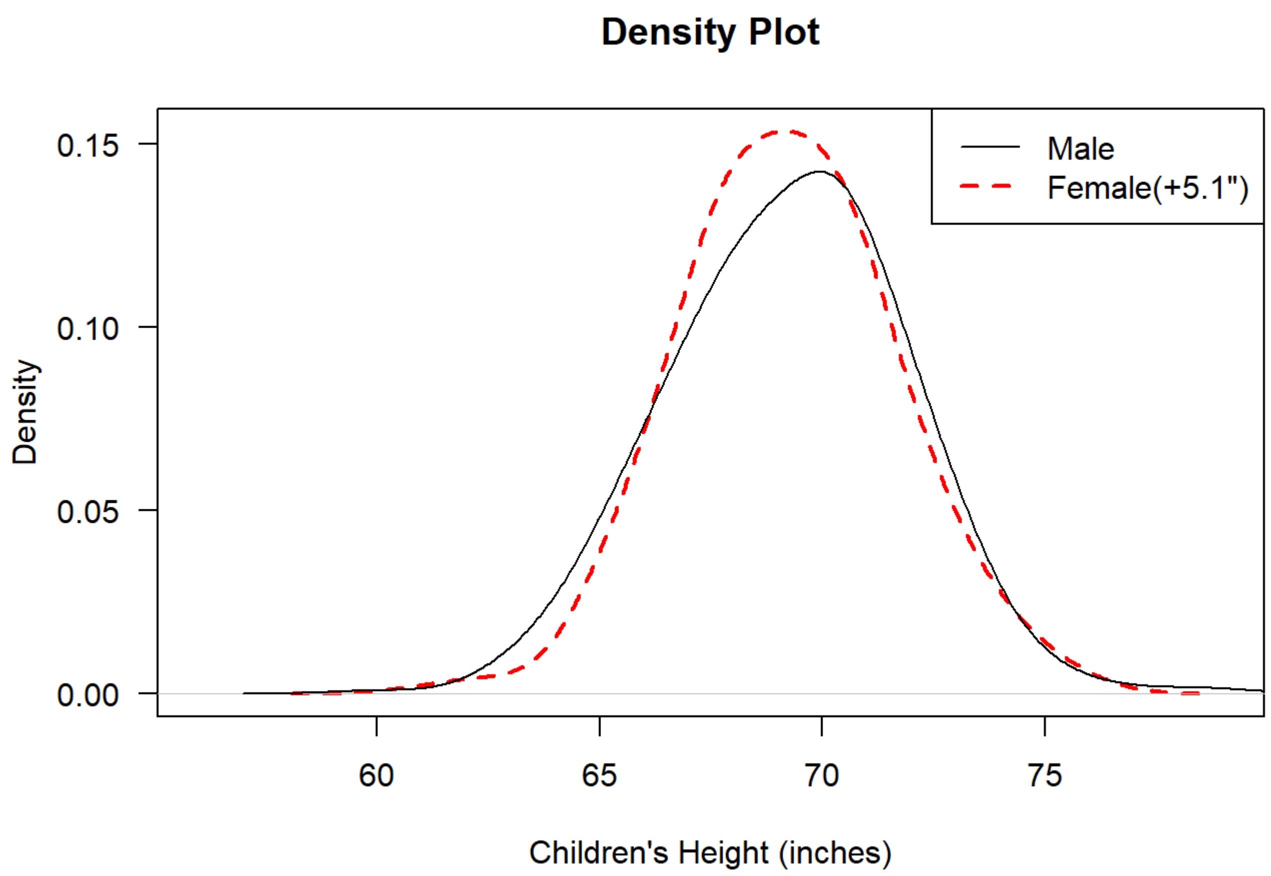
Again, the gender difference is much reduced.

Galton chose to multiply all the female heights by 1.08, but we can see from the calculations that we can also reduce the gender difference by adding 5.1 inches to all the female heights.

```
plot(density(galton$Mother+5.1, bw=1), xlab="Height (inches)", main="Density P
lot", las=1,
     xlim=c(min(galton$Mother),max(galton$Father)),lty=2,lwd=2, col="red")
lines(density(galton$Father, bw=1))
legend("topright", c("Father",'Mother(+5.1")'), col=c("black","red"), lty=1:2,
lwd=1:2)
```



```
plot(density(galton$Height[galton$Gender=="F"]+5.1, bw=1),
     xlab="Children's Height (inches)", main="Density Plot",
     las=1,xlim=c(min(galton$Height),max(galton$Height)),
     lty=2,lwd=2, col="red")
lines(density(galton$Height[galton$Gender=="M"], bw=1))
legend("topright", c("Male",'Female(+5.1")'), col=c("black","red"), lty=1:2, l
wd=1:2)
```



Galton defined the mid-parental height as the average of the Father's and Mother's height:

$$H_{midparental} = \frac{1}{2}(H_{father} + 1.08H_{mother}),$$

where the factor 1.08 was introduced to account for the gender difference. He also "transmuted" the heights of all female children to the male equivalents by multiplying the female heights by 1.08. He then fitted a model predicting children's adjusted height from the mid-parental height.

g. (4 points)

Add a column to the data frame that stores the mid-parental heights. (1 pt)

Name the new column as `MP` .

```
galton$MP <- (galton$Father + 1.08*galton$Mother)/2
```

Add another column to the data frame that stores the adjusted heights of the children: the adjusted heights of the male children are the same as their heights; the adjusted heights of the female children are equal to their heights times 1.08. (2 pts)

Name the new column as `AH`.

```
galton$AH <- galton$Height
galton$AH[galton$Gender=="F"] <- galton$Height[galton$Gender=="F"]*1.08
```

Calculate the correlation coefficient between the children's adjusted height and the mid-parental height (1 pt)

```
cor(galton$AH,galton$MP)
```

```
[1] 0.5105247
```

The correlation is $\boxed{0.51}$.

h. (4 points)

Fit a simple regression model predicting children's adjusted height from the mid-parental height. (2 pts)

```
fit <- lm(AH ~ MP, data=galton)
summary(fit)
```

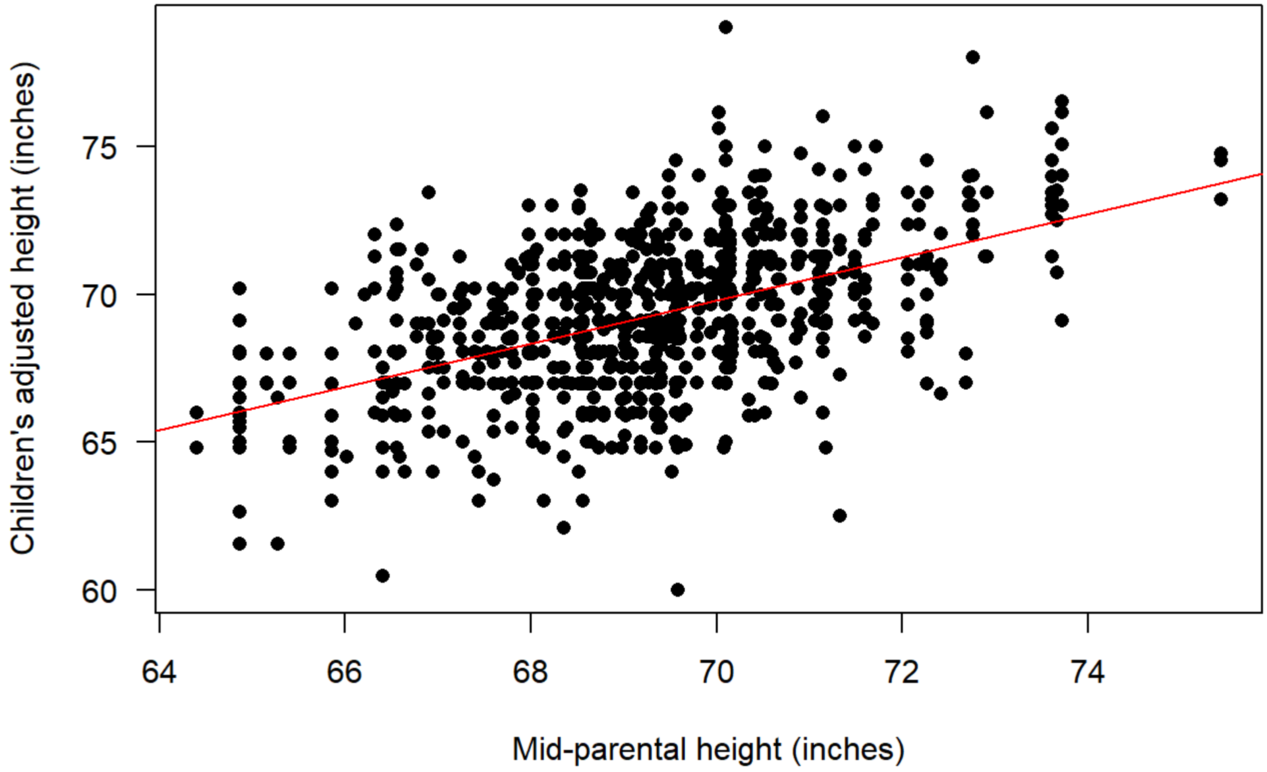
```

Call:
lm(formula = AH ~ MP, data = galton)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.4947 -1.4779  0.0995  1.5175  9.1262 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.76698    2.84062   6.607 6.74e-11 ***
MP           0.72906    0.04102  17.772  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.233 on 896 degrees of freedom
Multiple R-squared:  0.2606,	Adjusted R-squared:  0.2598 
F-statistic: 315.9 on 1 and 896 DF,  p-value: < 2.2e-16
```

The regression equation is

$$\boxed{\hat{AH} = 18.77 + 0.7291H_{\text{midparental}}}$$

Make a scatter plot of children's adjusted height vs the mid-parental height and then add the regression line on the plot. (2 pts)

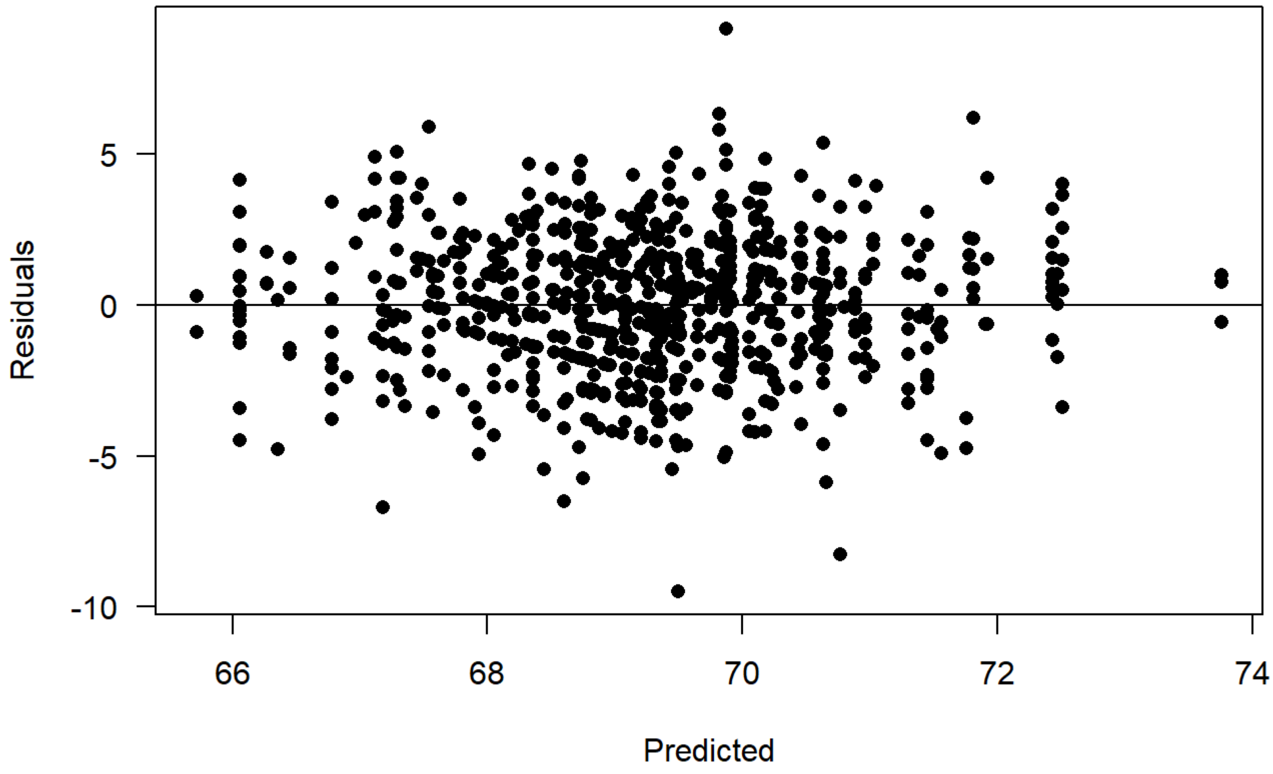
```
plot(AH ~ MP, data=galton, pch=16, las=1, xlab="Mid-parental height (inches)",
     ylab="Children's adjusted height (inches)")
abline(fit, col="red")
```



i. (2 points)

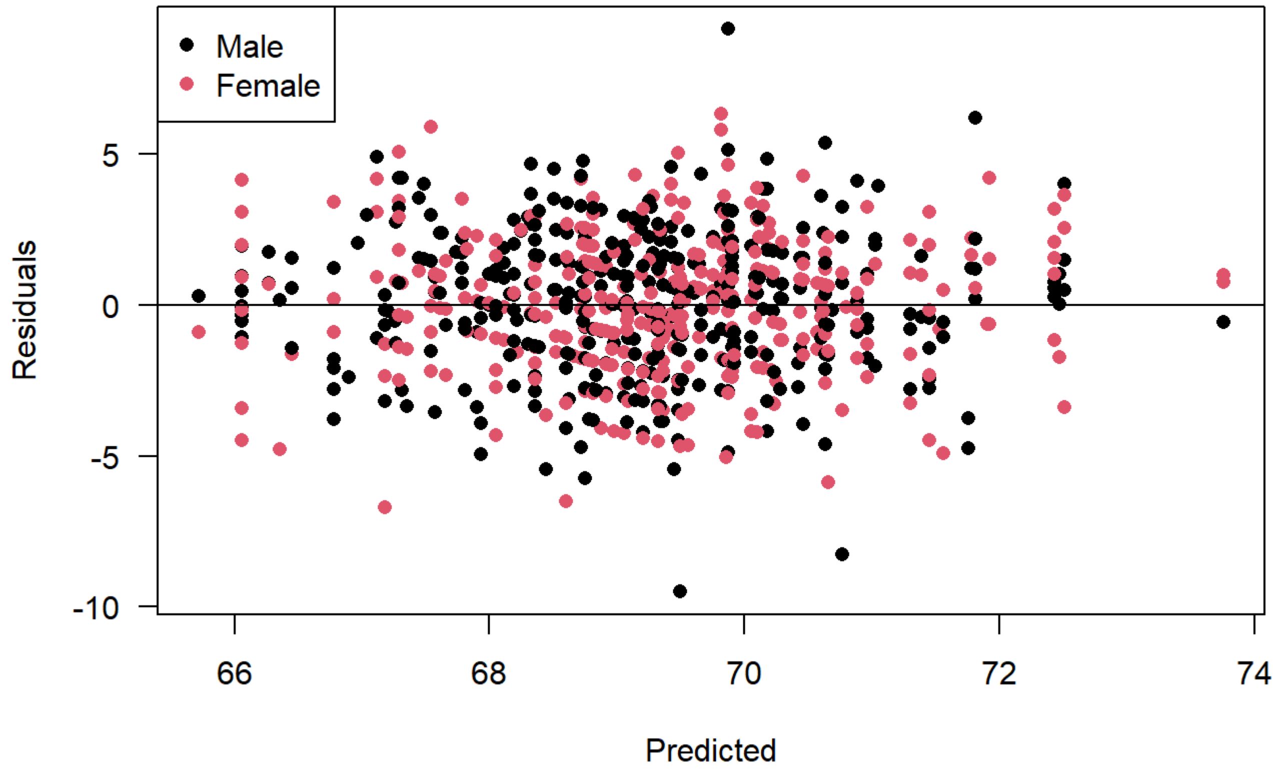
Plot the residuals versus the fitted values for the simple regression model above.

```
plot(fit$residuals ~ fit$fitted.values,xlab="Predicted", ylab="Residuals",
     pch=16, las=1)
abline(h=0)
```



Aside: After “transmuting” the female heights, we no longer see the two clusters of points. When we color-code the gender in the residual plot, male and female points are mixed together:

```
plot(fit$residuals ~ fit$fitted.values, col=galton$Gender,
     xlab="Predicted", ylab="Residuals", pch=16, las=1)
abline(h=0)
legend("topleft",c("Male","Female"),pch=16,col=1:2)
```



How does the simple regression model in (h) compare with the multiple regression model in (c) One measure of the "goodness of fit" is R^2. However, comparing R^2 returned by the model in (c) and R^2 of the model in (h) is misleading because their predicted variables are different. In (c), the predicted variable is children's height, whereas in (h) the predicted variable is children's adjusted height. To have a fair comparison, we want to calculate the R^2 of the model in (c) for the adjusted height and then compare it with the R^2 in (h).

j. (8 points)

1. (2 points) Calculate the predicted values of children's adjusted height from the multiple regression model by multiplying the predicted heights by 1.08 for female children and keeping the predicted heights of the male children unchanged. Store the result in a new variable.

Below we store the predicted value of AH to the variable `predicted_AH`.

```
predicted_AH <- fit_mult$fitted.values
predicted_AH[galton$Gender=="F"] <- 1.08*predicted_AH[galton$Gender=="F"]
```

2. (5 points) Calculate R^2 for the adjusted heights of the model in (c) by $R^2_{AH} = 1 - SSE_{AH}/SST_{AH}$, where $SSE_{AH} = \sum(AH - \hat{AH})^2$ and $SST_{AH} = \sum(AH - \overline{AH})^2 = (n-1)s^2_{AH}$. Here AH is the actual adjusted heights of the Galton children calculated in (h) above, $\hat{AH}$ is the predicted adjusted heights calculated in (j1) above, $\overline{AH}$ is the mean of the adjusted height, s^2_{AH} is the sample variance of the adjusted height, and n is the total number of observations in the dataset.

```
SSE_AH <- sum( (galton$AH - predicted_AH)^2 )
SST_AH <- (nrow(galton)-1)*var(galton$AH)
(Rsq_AH <- 1-SSE_AH/SST_AH)
```

```
[1] 0.2665998
```

Hence, R^2_{AH} of the model in (c) is about $\boxed{0.27}$.

3. (1 point) Based on the values of the R^2 for the adjusted height, is the multiple regression model in (c) much better than the simple regression model in (h)?

The R^2_{AH} of the multiple regression model is 0.27 as computed above. This is only slightly larger than 0.26, the R^2_{AH} of the simple regression model. Based on this result the multiple regression model is not much better than the simple regression model.

Aside: You might wonder why the R^2 for the adjusted height ($R^2_{AH} = 1 - SSE_{AH}/SST_{AH}$) of the multiple regression model is so much smaller than the R^2 for the height ($R^2_H = 1 - SSE_H/SST_H$). The main reason is that SST_H is 1.9 times SST_{AH}.

Let's compare all the terms. First, compare SSE_{AH} with SSE_H:

```
SSE_H <- sum( (galton$Height-fit_mult$fitted.values)^2 )
(SSE_AH/SSE_H)
```

```
[1] 1.068398
```

So SSE_{AH} is only about 7% larger than SSE_H. On the other hand, $SST_H/SST_{AH} = s^2_{AH}/s^2_H$ =

```
var(galton$Height)/var(galton$AH)
```

```
[1] 1.905082
```

As a result, $SSE_{AH}/SST_{AH} = (1.068 \times 1.905)SSE_H/SST_H = 2.035SSE_H/SST_H$. Even though the prediction errors are about the same, SSE_{AH} is divided by a smaller denominator, making $1 - R^2_{AH} = 2.035(1 - R^2_H)$. That is, the fractional error associated with the predicted adjusted height is about twice of that associated with the predicted height. With $R^2_H = 0.64$ (see the summary of the multiple regression model in question (c) above), $R^2_{AH} = 1 - 2.035 \times (1 - 0.64) = 0.27$, the value we calculated above. This explains why R^2_{AH} is so much smaller than R^2_H.

The reason that SST_{AH} is smaller than SST_H is easy to explain: one measures the variance of the adjusted height and the other measures the variance of the height. The variance of the height is larger because of the gender difference in height. When the gender difference is taken into account in the adjust height, its variance is reduced.

Alan is a boy born in Guatemala. Carly is a girl born in India. They are both two years old. The heights of Alan’s father and mother are 62 inches and 58 inches, respectively. The heights of Carly’s father and mother are 68 inches and 65 inches, respectively.

k. (4 points)

Use the multiple regression model above to predict the height of Alan and Carly when they become adults.

First, create a data frame to store the information of Alan and Carly.

```
data_new <- data.frame(Gender=c("M","F"), Father=c(62,68), Mother=c(58,65))
rownames(data_new) <- c("Alan","Carly")
```

Compute the mid-parental height (for the next question):

```
data_new$MP <- (data_new$Father + 1.08*data_new$Mother)/2

# Look at the new data frame
data_new
```

```
      Gender Father Mother    MP
Alan       M     62     58 62.32
Carly      F     68     65 69.10
```

```
# predict the adult heights
(pred_mult <- predict(fit_mult, newdata=data_new))
```

```
    Alan    Carly 
64.38807 63.84845 
```

The predicted adult heights for Alan and Carly are 64.4 inches and 63.8 inches, respectively.

l. (4 points)

Use the simple regression model above to predict the height of Alan and Carly when they become adults.
Note: You’ll need to convert the predicted adjusted height back to height for Carly.

```
pred_simple <- predict(fit, newdata=data_new)
pred_simple["Carly"] <- pred_simple["Carly"]/1.08
pred_simple
```

```
    Alan    Carly 
64.20176 64.02293 
```

OR

```
pred_simple <- predict(fit, newdata=data_new)
pred_simple[data_new$Gender=="F"] <- pred_simple[data_new$Gender=="F"]/1.08
pred_simple
```

```
    Alan    Carly 
64.20176 64.02293 
```

The second method is more convenient if `data_new` contains many female data points.

The predicted adult heights for Alan and Carly are 64.2 inches and 64 inches, respectively.

m. (2 points)

Explain why the multiple and simple regression models above may not be suitable for predicting the adult heights for Alan and Carly.
Hint: Watch this video (https://youtu.be/C_NTLtM-f2I#t=6m54s) for a similar question, or read the bottom of P.35 in the Fall 2017 Stat 200 notebook for two other similar questions.

As stated at the beginning, Galton’s height data were collected in late 19th century in England. The regression models based on the data apply to people in that area around that time. We can’t assume the models could apply to a wider population without further study. Alan was born in Guatemala in the 21th century, and Carly was born in India in the 21st century. Hence the regression models constructed from the Galton height data may not apply to them. A study indicates that the average adult height not only differs in regions but also changes in the past 100 years (https://elifesciences.org/articles/13410).