Regression_HW4

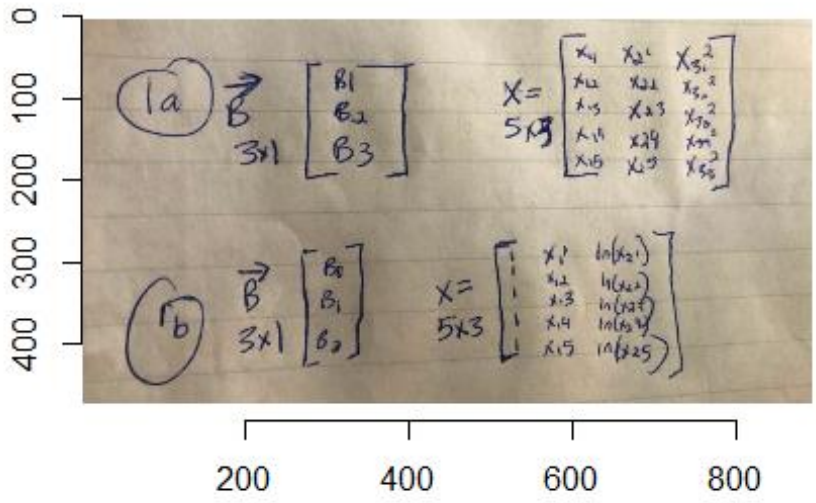
Blake Tindol

11/2/2020

Question: 1A & B

For each of the following regression models, carefully provide the design matrix X and vector of regression coefficients B

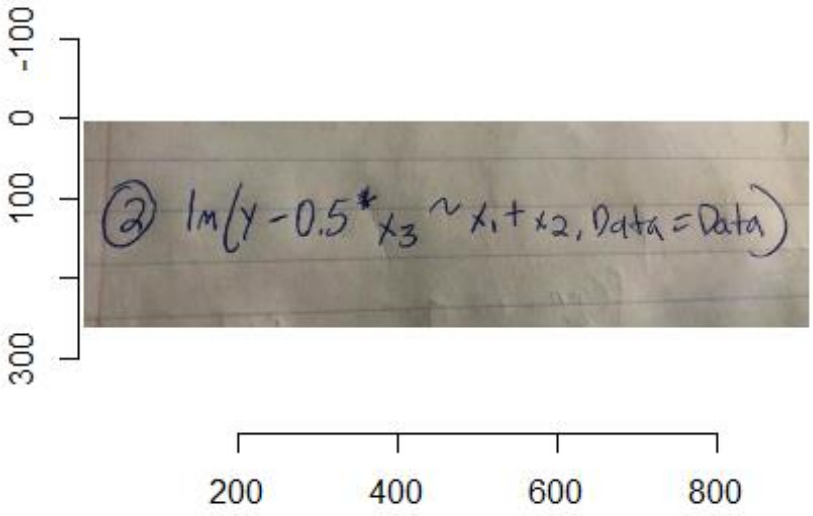
```
plot(myimg)
```



Question: 2

Supose you want to fit the multiple regression model such as where i=1,….,n, by the method of least squares when you are given that b3 = 0.5. Precisely explain how you can obtain the

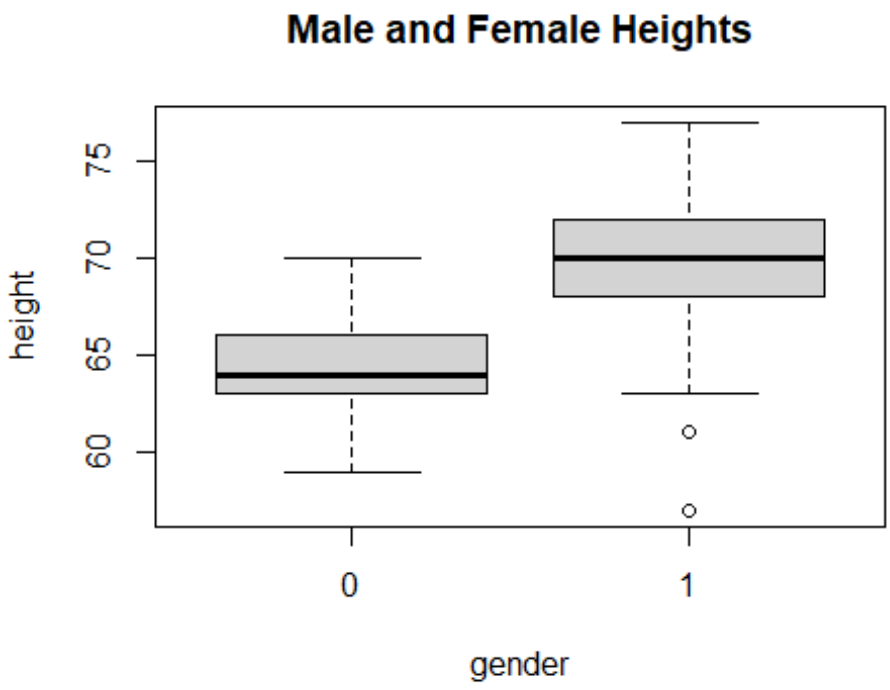
desired fitted multiple regression from a computer program.



Question: 3a

Provide side by side boxplots of heights of female and male children. Comment on the plot.

```
boxplot(height~male, data = childData, main = "Male and Female Heights",xlab
= "gender")
```

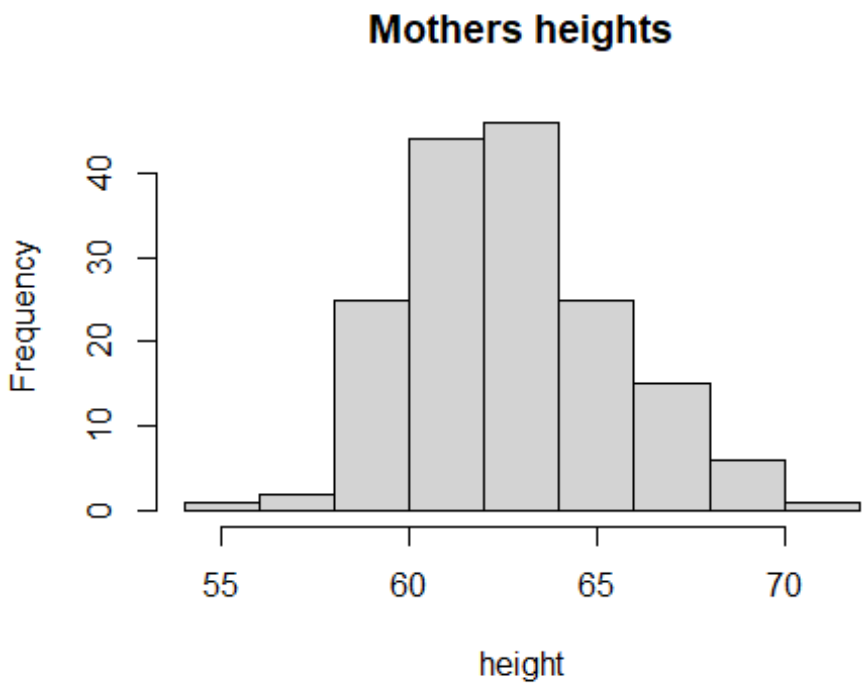



```
# We can see that male has a few outlier data points and the average hight of
male is much larger than the average height of female
```

Question: 3b

Provide a histogram of mothers' heights. Comment on the plot. Are there any missing values or extreme values (outliers)?

```
hist(childData$momht, main = "Mothers heights", xlab = "height")
```




```
summary(childData)
```

```
##        ID            alcohol        exercise         height
##  Min.   :  1.00   Min.   : 0.00   Min.   : 0.000   Min.   :57.00
##  1st Qu.: 44.00   1st Qu.: 0.00   1st Qu.: 2.000   1st Qu.:64.00
##  Median : 86.00   Median : 1.00   Median : 3.000   Median :66.00
##  Mean   : 86.56   Mean   : 4.27   Mean   : 4.698   Mean   :66.84
##  3rd Qu.:131.00   3rd Qu.: 4.75   3rd Qu.: 6.000   3rd Qu.:69.50
##  Max.   :173.00   Max.   :55.00   Max.   :30.000   Max.   :77.00
##                   NA's   :6       NA's   :1
##       male            dadht           momht
##  Min.   :0.0000   Min.   :55.00   Min.   :54.00
##  1st Qu.:0.0000   1st Qu.:67.00   1st Qu.:61.00
##  Median :0.0000   Median :69.00   Median :63.00
##  Mean   :0.4545   Mean   :69.17   Mean   :63.27
##  3rd Qu.:1.0000   3rd Qu.:72.00   3rd Qu.:65.00
##  Max.   :1.0000   Max.   :78.00   Max.   :71.00
##
```

```
IQR(childData$momht)*1.5 + 71
```

```
## [1] 77
```

```
54 - IQR(childData$momht)*1.5
```

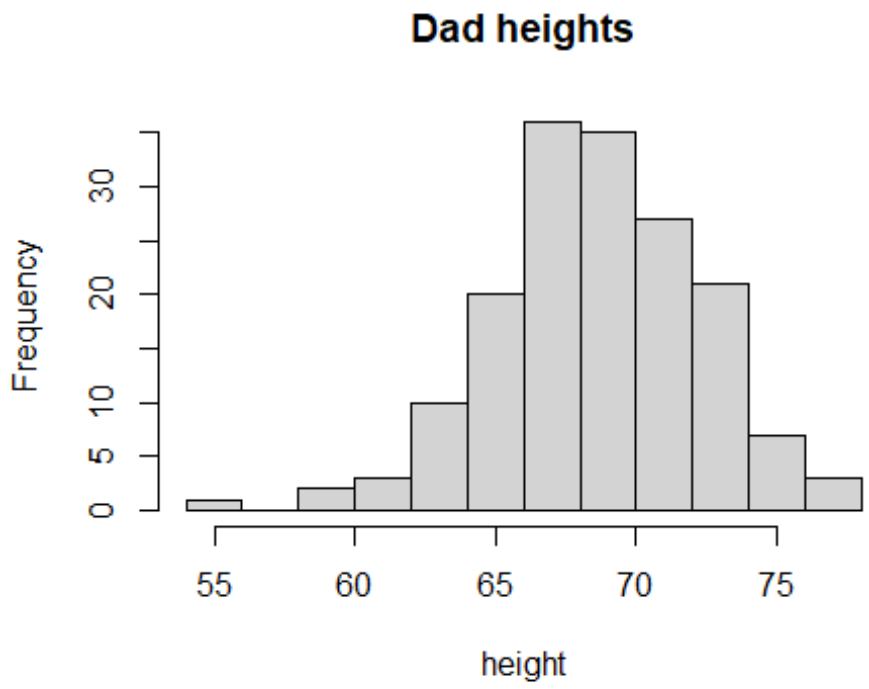
```
## [1] 48
```

```
# We can see that the histogram just like the box plot doesnt show any data
below 48 or above 77 indicating that there is no missing or outlier
datapoints for mother height
```

Question: 3c

Provide a histogram of fathers' heights. Comment on the plot. Are there any missing values or extreme values (outliers)?

```
hist(childData$dadht, main = "Dad heights", xlab = "height")
```




```
summary(childData)
```

```
##        ID            alcohol         exercise          height
##  Min.   :  1.00   Min.   : 0.00   Min.   : 0.000   Min.   :57.00
##  1st Qu.: 44.00   1st Qu.: 0.00   1st Qu.: 2.000   1st Qu.:64.00
##  Median : 86.00   Median : 1.00   Median : 3.000   Median :66.00
##  Mean   : 86.56   Mean   : 4.27   Mean   : 4.698   Mean   :66.84
##  3rd Qu.:131.00   3rd Qu.: 4.75   3rd Qu.: 6.000   3rd Qu.:69.50
##  Max.   :173.00   Max.   :55.00   Max.   :30.000   Max.   :77.00
##                   NA's   :6       NA's   :1
##       male            dadht           momht
##  Min.   :0.0000   Min.   :55.00   Min.   :54.00
##  1st Qu.:0.0000   1st Qu.:67.00   1st Qu.:61.00
##  Median :0.0000   Median :69.00   Median :63.00
##  Mean   :0.4545   Mean   :69.17   Mean   :63.27
##  3rd Qu.:1.0000   3rd Qu.:72.00   3rd Qu.:65.00
```

```
##  Max.   :1.0000   Max.   :78.00   Max.   :71.00
##
```

```
IQR(childData$dadht)*1.5 + 78 # finding max outlier range
```

```
## [1] 85.5
```

```
55 - IQR(childData$dadht)*1.5  # finding min
```

```
## [1] 47.5
```

```
childData[childData$dadht > 85.5 | childData$dadht < 47.5,"dadht"] # There are no extreme outliers
```

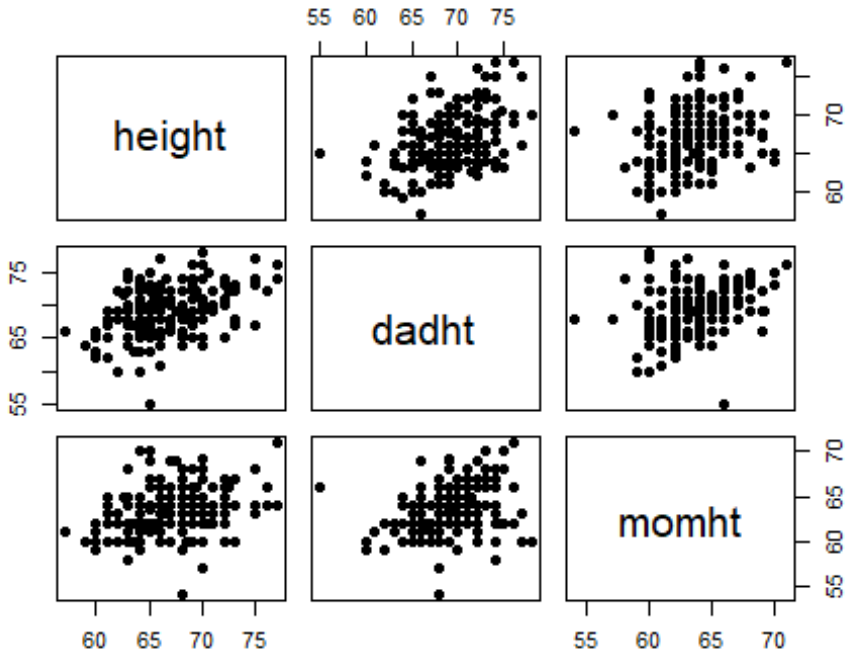
```
## numeric(0)
```

```
# We can see that the histogram just like the box plot there are outliers but the outliers are not above the range establish by 1.5 * the IQR
```

Question: 3d

Provide a scatter plot matrix (as in Lecture 8 page 9) for 3 variables height, momht and dadht. Comment on the plot

```
pairs(childData[,c(4,6,7)], pch = 19)
```

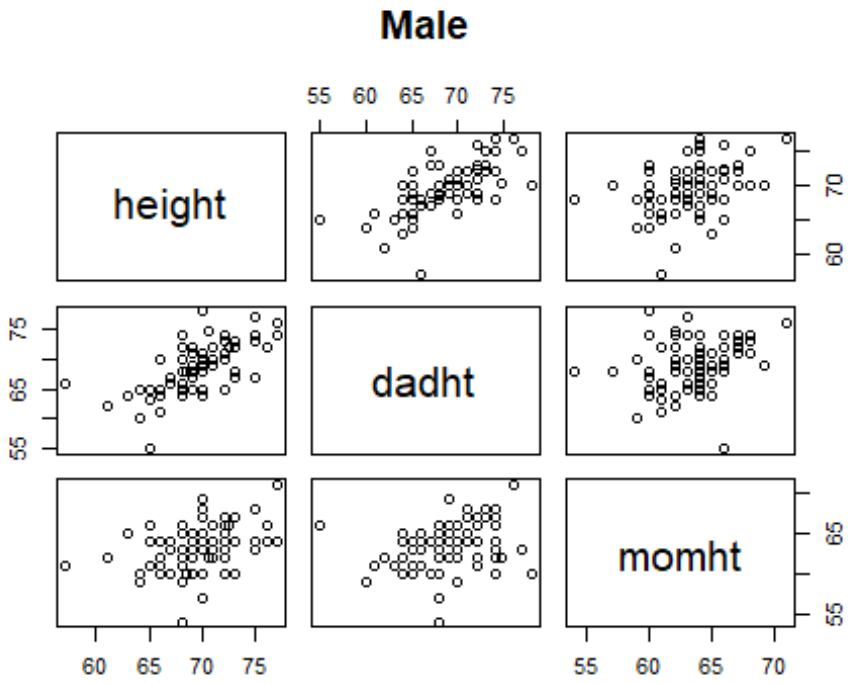



```
# It does not look like there is any strong relationship between any of the independent variables indicating we dont need to investigate further
```

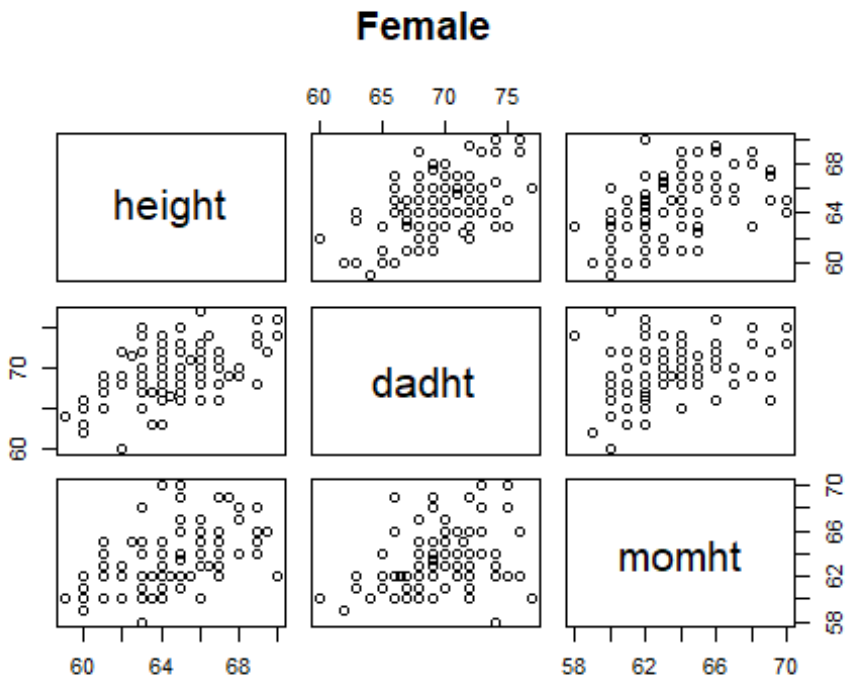
Question: 3e

Now provide scatter plot matrix in part (d) separately for male and female children. Comment on the plot.

```
pairs(childData[which(childData$male == 1),c('height',"dadht",'momht')],main
= "Male")
```




```
pairs(childData[which(childData$male == 0),c('height',"dadht",'momht')],main
= "Female")
```




```
# for the male matrix it looks like there is a stronger relationship between
height and dad height when you compare the height and mom height relationship
# for the female matrix there does not seem to be a distinct difference in
the strength of relationship from height to dad height or mother height
```

Question: 3f

Show the summary and provide a precise interpretation of the coefficient β3 that can be understood by a layman

```
model1 <- lm(height ~ male + momht + dadht,data = childData)
summary(model1)
```

```
##
## Call:
## lm(formula = height ~ male + momht + dadht, data = childData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7431  -1.4537   0.0191   1.5299   5.9459
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.96746    4.65831   3.642 0.000364 ***
## male         5.29822    0.36377  14.565  < 2e-16 ***
## momht        0.29962    0.06876   4.357 2.34e-05 ***
## dadht        0.41213    0.05107   8.069 1.54e-13 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 161 degrees of freedom
## Multiple R-squared:  0.6604, Adjusted R-squared:  0.6541
## F-statistic: 104.4 on 3 and 161 DF,  p-value: < 2.2e-16

# Here b3 represents fathers height and the .41213 means this. For every unit
increase in dad height there will be a .41213 unit increase in height
```

Question: 3g

Show the summary and provide a precise interpretation of the coefficient β1 that can be understood by a layman.

```
model2 <- lm(height ~ male + momht ,data = childData)
summary(model2)

##
## Call:
## lm(formula = height ~ male + momht, data = childData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4918  -1.4455   0.0776   1.6008   7.0776
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.38884    4.87674   7.052 4.85e-11 ***
## male         5.01498    0.42777  11.724  < 2e-16 ***
## momht        0.47685    0.07698   6.194 4.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.736 on 162 degrees of freedom
## Multiple R-squared:  0.5231, Adjusted R-squared:  0.5172
## F-statistic: 88.85 on 2 and 162 DF,  p-value: < 2.2e-16

# Here b1 is 5.014 and reads as such means that a person who is a Male is
5.0149 units taller than a female on average this is with mother height as
part of the model as a predictor
```

Question: 3h

Show the summary and provide a precise interpretation of the coefficient β1 that can be understood by a layman.

```
model3 <- lm(height ~ male + dadht ,data = childData)
summary(model3)
```

```
##
## Call:
## lm(formula = height ~ male + dadht, data = childData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2355  -1.6516  -0.1012   1.4491   6.2813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97678    3.55308   8.718  3.2e-15 ***
## male         5.36620    0.38308  14.008  < 2e-16 ***
## dadht        0.48322    0.05101   9.472  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.441 on 162 degrees of freedom
## Multiple R-squared:  0.6204, Adjusted R-squared:  0.6157
## F-statistic: 132.4 on 2 and 162 DF,  p-value: < 2.2e-16
```

```
# Here b1 is 5.366 and reads as such means that a person who is a Male is
5.366 units taller than a female on average this is with dad height as part
of the model as a predictor
```

Question: 3i

Based on R2, which model is best? Which model is second best?

```
summary(model1)$r.squared # best
```

```
## [1] 0.660445
```

```
summary(model2)$r.squared # worst
```

```
## [1] 0.523112
```

```
summary(model3)$r.squared #second best
```

```
## [1] 0.6204024
```