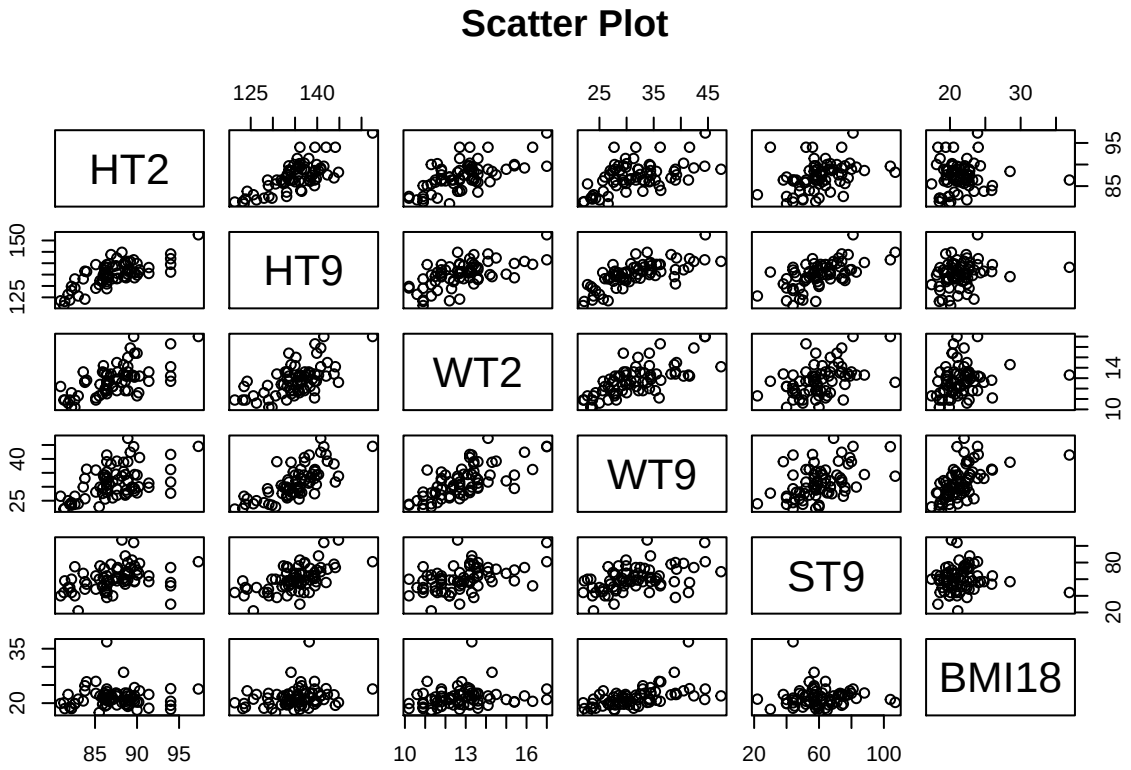
STAT 504: Linear Regression

Pratima K C

Question 1a (1 point) Draw the scatterplot matrix of HT2, HT9, WT2, WT9, ST9 and BMI18. Compare the scatterplot matrix with the matrix of sample correlations for these variables. What do you observe? Compare the relationship of the predictors with the response and the pairwise predictor relationships.

Answer: The scatter plot matrix and the matrix of sample correlations for the above variables both shows the similar results. The BMI18 has very low correlation with other metrics. The ST9 has higher correlation with HT9 than other metrics. However, the HT9 has higher correlation with HT2 than ST9 because the height at age 9 is mostly likely to be realted with height at age 2 than strength at age 9. Similarly, the WT9 has higher correlation with HT9 than other metrices. The scatter polot and correaltions of these metrics are given below:




```
## [1] "Correlation of different metrices"
```

```
##         WT2    HT2   WT9   HT9    ST9  BMI18
## WT2   1.000 0.6445 0.693 0.607 0.4516 0.1909
## HT2   0.645 1.0000 0.523 0.738 0.3617 0.0426
## WT9   0.693 0.5229 1.000 0.728 0.4530 0.5459
## HT9   0.607 0.7384 0.728 1.000 0.6034 0.2369
## ST9   0.452 0.3617 0.453 0.603 1.0000 0.0056
## BMI18 0.191 0.0426 0.546 0.237 0.0056 1.0000
```

Question 1 b (2 points) Fit two linear models:

1. ***E[BMI18 | WT9, ST9]***
2. ***E[BMI18 | HT2, WT2, HT9, WT9, ST9]***

Print their summaries and comment on the output. Which of the estimates that are in model 2 and not model 1 are significant at the 5% level (α = 0:05)?

Answer: In model 1 both the predictor variables (WT9 & ST9) are significant as p-values are smaller than alpha=0.05. However, in model 2 three predictor varilabes (HT2, WT2, HT9)are not significant as their p-values are larger than α=0.05. But other two variables (WT9 & ST9) both variables are significant for model 2 too. When comapring non of the estimates that are in model 2 and not in model 1 are significant. The summary for both models are given below:

```
##
## Call:
## lm(formula = BMI18 ~ WT9 + ST9, data = BGSgirls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1736 -1.2146 -0.2474  1.1231 11.2834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.63878    1.54661   9.465 5.66e-14 ***
## WT9          0.32418    0.05151   6.293 2.72e-08 ***
## ST9         -0.05552    0.01983  -2.799  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.222 on 67 degrees of freedom
## Multiple R-squared:  0.3715, Adjusted R-squared:  0.3528
## F-statistic:  19.8 on 2 and 67 DF,  p-value: 1.747e-07
```

```
##
## Call:
## lm(formula = BMI18 ~ HT2 + WT2 + HT9 + WT9 + ST9, data = BGSgirls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0948 -1.2186 -0.2533  1.0090 10.4951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.855335   8.781156   3.514 0.000817 ***
## HT2         -0.193997   0.130819  -1.483 0.142996
## WT2         -0.317779   0.278736  -1.140 0.258505
## HT9          0.008057   0.096344   0.084 0.933613
## WT9          0.419762   0.075211   5.581  5.2e-07 ***
## ST9         -0.044416   0.022219  -1.999 0.049853 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.14 on 64 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.3996
```

```
## F-statistic: 10.19 on 5 and 64 DF,  p-value: 3.294e-07
```

Question 1c (4 points) Conduct an anova comparison of the above 2 models.Test the hypothesis H0 : (βHT2; βWT2; βHT9) = (0; 0; 0) at the 5% level.

Answer:

```
## Analysis of Variance Table
##
## Model 1: BMI18 ~ WT9 + ST9
## Model 2: BMI18 ~ HT2 + WT2 + HT9 + WT9 + ST9
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     67 330.70
## 2     64 293.04  3    37.666 2.7421 0.05037 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 1c (i)(1 point) What is the test statitic for this test?

Answer: The test statistics for this test is 2.7421

Question 1c (ii)(1 point) What distribution does this test statistic follow under the null hypothesis (specity the degrees of freedom)

Answer: RSS(model1 & model2) and df, here test statistic follow under 64 & 3.

Question 1c (iii) (1 point) Do you reject the null hypothesis?

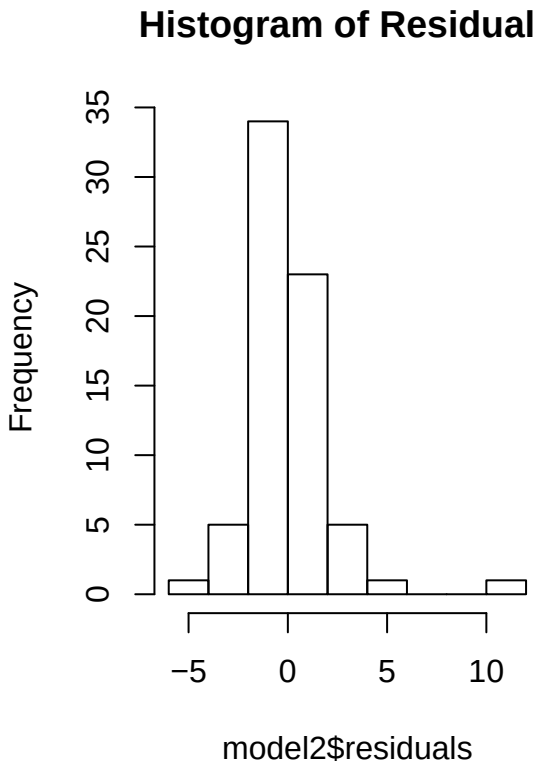
Answer: No, we cannot reject the null hypothesis because the p-value is higher than significance level at $\alpha = 0.05$***.***

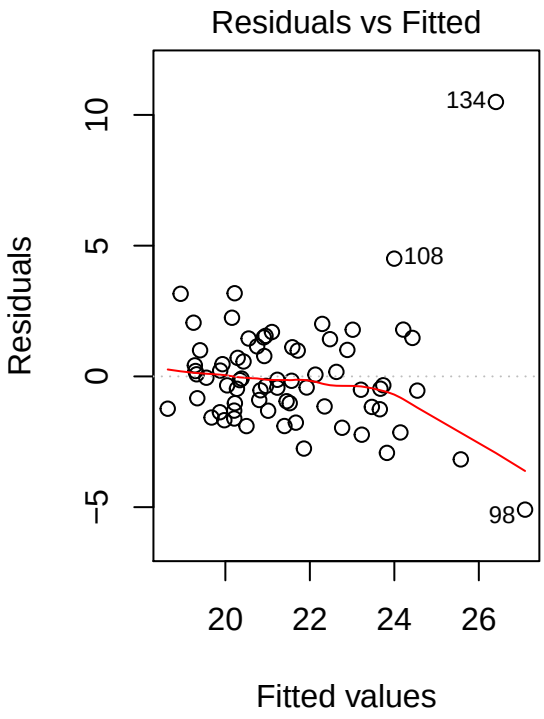
Question 1c (iv) (1 point) Based on the result of the hypothesis test, which model would you choose?

Answer: I would choose to use the smaller model (model1) as we cannot reject the null hypothesis in bigger model (model2).

Question 1d (2 point) Plot the histogram of the residuals and the TA plot from the model 2 above. Do the normality and the constant variance assumptions appear to hold?

Answer: The histogram seems to be right skewed so it does not holds the normality assumption. The TA plots shows that data mostly have the connstant variable through out but it gets spreaded towards the right side with higher fitted values and this side have higher variance. Therefore the constant variance assumptions to be voilated slightly.





Question 1e (2 point) What are the $\hat{\sigma}$ and the in-sample MSE of model 2 above?

Answer: For $\hat{\sigma}$ I used the following equation: $\hat{\sigma} = \sqrt{(RSS/(n-p+1))}$

$\hat{\sigma} = 2.139801$ ***and*** $MSE = RSS/n = 293.04/70 = 4.186286$

Question 1f (5 point) Consider model 2 summary output and test null hypotheses

Answer: Different correction are given below:

```
## [1] "p-values of model2"
```

```
##          HT2          WT2          HT9          WT9          ST9
## 1.429963e-01 2.585048e-01 9.336128e-01 5.198467e-07 4.985317e-02
```

```
## [1] "Bonferroni Correction"
```

```
## [1] "Parameters where p-value < 0.1"
```

```
##   HT2   WT2   HT9   WT9   ST9
## FALSE FALSE FALSE  TRUE FALSE
```

Therefore, we only reject H_0^4 because it has p-value <0.1

Holm correction

```
## [1] "Parameters where p-value < adjusted significant level"
```

```
##   WT9   ST9   HT2   WT2   HT9
##  TRUE FALSE FALSE FALSE FALSE
```

Therefore, we only reject H_0^4 because it has p-value $< \alpha$

Benjamini-Hochberg correction

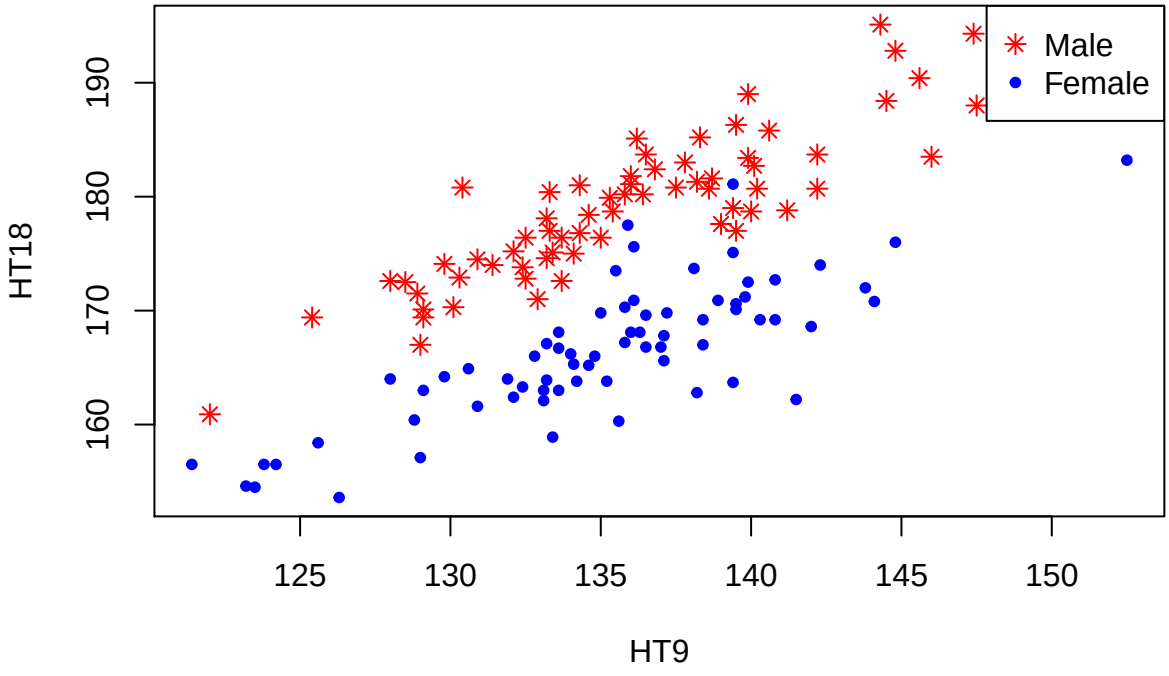
```
## [1] "Parameters where p-value < adjusted significant level"
```

```
##   WT9   ST9   HT2   WT2   HT9
##  TRUE  TRUE FALSE FALSE FALSE
```

Therefore, we only reject H_0^4 because it has p-value $< \alpha$

Question 2 a (1 point) Consider the regression of HT18 (response) on HT9 and Sex (predictors). Draw the scatterplot of HT18 vs HT9 using a different symbol or color for men and women (with a legend). Comment on the appropriate mean function for the data. Looking at the scatterplot, do you think including the factor Sex in the linear model is justified? Explain.

Answer: Yes, I think including the factor Sex in the linear model is jusitifies as we can see the scatter plot, it seems to have two linear model. As per the graph male tend to be taller than female in both age group. The mean function is given by $E[Y|X = x] = \beta_0 + \beta_1 * X_{HT9} + \beta_2 * X_{Sex}$, where β_0 is expected value of y (HT18), when HT9=0, and Sex=0 that mean male.



Question 2b (6 points) Fit the linear model (call it fit.height) with HT18 as a response and HT9 and Sex as predictors.

Answer:

```
##
## Call:
## lm(formula = HT18 ~ HT9 + as.factor(Sex), data = BGSall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4694  -2.0952  -0.0136   1.7101  10.4467
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      48.51731    7.33385   6.616 8.27e-10 ***
## HT9               0.96006    0.05388  17.819  < 2e-16 ***
## as.factor(Sex)1 -11.69584    0.59036 -19.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.432 on 133 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8494
## F-statistic: 381.7 on 2 and 133 DF,  p-value: < 2.2e-16
```

Question 2b (i) (1 point) Consider the fitted regression line for women. What is the intercept of this fitted line?

Answer: The intercept of this fitted line is 48.51731+ (-11.69584)= 36.82147

Question 2b (ii) (1 point) What is the predicted height at 18 years of age for a 135cm tall 9-year-old girl (heights given in the data set are in centimeters - cm)?

Answer: The predicted height at 18 years of age for a 135 cm tall 9-year old girl is 166.4291 cm

```
##        1
## 166.4291
```

Question 2b (iii) (2 points) What would be the predicted average change in height at age 18, for the same girl if in fact her height at age 9 was 137cm, but was measured wrongly as 135cm?

Answer: The predicted average change in height at age 18 would be: (137-135)**0.96006= 1.92012cm***

Question 2b (iv) (1 point) Based on this model, what is the 95% confidence interval for the difference in height between men and women?

Answer: Based on this model the 95% confidence interval for the difference in the height between men and women is:

$= -11.69584 \pm (0.59036 * 1.96)$

$= -11.69584 \pm 1.157106$

$= (-12.8524, -10.5396)$

Question 2b (v) (1 point) Was your suspicion from part (a) correct? Test the hypothesis H0 : β Sex = 0 at the 5% level. Do you reject the null hypothesis? Explain.

Answer: The null hypothesis is rejected as the p-value (2e-16) is $<$ 0.05. The suspicion from part a is correct. The inclusion of sex does make difference in the model.

Question 2c (5.5 + 4 Bonus points) Consider the following three models in addition to the above model in fit.height. (These are written in Wilkinson-Rogers notation, 1 indicates that the intercept is present in the model.) 1. HT18 | 1 + HT2 + HT9 + Sex 2. HT18 | 1 + HT2 + HT9 + Sex + Sex:HT2 + Sex:HT9 3. HT18 | 1 + HT2 + HT9 + HT2:HT9 + Sex + Sex:HT2 + Sex:HT9 + Sex:HT2:HT9

Answer:

```
##
## Call:
## lm(formula = HT18 ~ HT2 + HT9 + Sex, data = BGSall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4508  -2.0960  -0.1031   1.7497  10.4690
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47.26295    8.26503   5.718 6.85e-08 ***
## HT2           0.04305    0.12917   0.333    0.739
## HT9           0.94129    0.07806  12.059  < 2e-16 ***
## Sex         -11.66213    0.60091 -19.407  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.444 on 132 degrees of freedom
## Multiple R-squared:  0.8517, Adjusted R-squared:  0.8484
## F-statistic: 252.8 on 3 and 132 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = HT18 ~ HT2 + HT9 + Sex + Sex:HT2 + Sex:HT9, data = BGSall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8022 -1.6619  0.1103  1.8932 10.9553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.1094    11.8573   3.383 0.000949 ***
## HT2          -0.1706     0.1782  -0.958 0.340019
## HT9           1.1329     0.1099  10.311  < 2e-16 ***
## Sex           0.7723    16.2328   0.048 0.962126
## HT2:Sex       0.4389     0.2546   1.724 0.087144 .
## HT9:Sex      -0.3761     0.1540  -2.442 0.015949 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.393 on 130 degrees of freedom
## Multiple R-squared:  0.8582, Adjusted R-squared:  0.8528
## F-statistic: 157.4 on 5 and 130 DF,  p-value: < 2.2e-16
```

```
##
## Call:
```

```
## lm(formula = HT18 ~ HT2 + HT9 + HT2:HT9 + Sex + Sex:HT2 + Sex:HT9 +
##     Sex:HT2:HT9, data = BGSall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7579 -1.6473 -0.0042  1.9225 10.9155
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -85.953158 243.203453  -0.353    0.724
## HT2           1.251244   2.745688   0.456    0.649
## HT9           2.062760   1.795162   1.149    0.253
## Sex          28.708720 299.470361   0.096    0.924
## HT2:HT9      -0.010477   0.020189  -0.519    0.605
## HT2:Sex       0.159776   3.420070   0.047    0.963
## HT9:Sex      -0.588978   2.204074  -0.267    0.790
## HT2:HT9:Sex   0.002135   0.025048   0.085    0.932
##
## Residual standard error: 3.412 on 128 degrees of freedom
## Multiple R-squared:  0.8589, Adjusted R-squared:  0.8512
## F-statistic: 111.3 on 7 and 128 DF,  p-value: < 2.2e-16
```

Question 2c(i) (2 points) How many parameters are estimated in each of the proposed models (fit.height,fit.height2, fit.height3, fit.height4)?

Answer: The number of parameter estimated in each modele are: fit.height= 3; fit.height2= 4; fit.height3= 6; fit.height4= 8

Question 2c (ii) (1.5 points) What is the predicted average height of a girl who is 135cm tall at age 9 and 90cm tall at age 2 accoriding to fit.height2, fit.height3, and fit.height4 respectively?

Answer: The prediction is given below:

```
## [1] "Prediction for fit.heighth2"

##        1
## 166.5496

## [1] "Prediction for fit.heighth3"

##        1
## 167.1903

## [1] "Prediction for fit.heighth4"

##        1
## 167.3507
```

Question 2c(iii) (Bonus 2 points) Consider a sequential procedure where you start from the smallest model (fit.height) and compare it with model fit.height2 using the anova test at the 10% level. If model fit.height is not rejected the procedure ends and you opt for model fit.height as the preferred model.

Answer: When comparing the smaller model (fit.height) with the larger model (fit.height2) we cannot reject the null hypothesis as the p-value (0.7394) > 0.1. So opt to choose the smaller model (fit.height).

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ HT9 + as.factor(Sex)
## Model 2: HT18 ~ HT2 + HT9 + Sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    133 1566.9
## 2    132 1565.6  1    1.3176 0.1111 0.7394
```

Question 2c(iv) (Bonus 2 points) Consider a converse sequential procedure from above. You start with largest model (from fit.height4) and compare it with model fit.height3 using the anova test at the 10% level. If the p-value is smaller than 10% you opt for the bigger model.

Answer: When comparing the largest model (fit.height4) iwth the another model (fit.height3) we cannot reject the null hypothesis as the p-value (0.7465) > 0.1. Then, we compared the model fit.height3 with fit.height2, here we reject the null hypothesis because p-value (0.05412) < 0.1. Therefore we opt to choose the larger model ie. fit.heigth3.

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ HT2 + HT9 + Sex + Sex:HT2 + Sex:HT9
## Model 2: HT18 ~ HT2 + HT9 + HT2:HT9 + Sex + Sex:HT2 + Sex:HT9 + Sex:HT2:HT9
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    130 1496.9
## 2    128 1490.1  2    6.8211 0.293 0.7465
```

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ HT2 + HT9 + Sex
## Model 2: HT18 ~ HT2 + HT9 + Sex + Sex:HT2 + Sex:HT9
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    132 1565.6
## 2    130 1496.9  2    68.697 2.9831 0.05412 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 2c(v) (2 points) What is the $\hat{\sigma}$ associated with models fit.height, fit.height2, fit.height3 and fit.height4? Based on the $\hat{\sigma}$, what is your preferred model?

Answer: Based on $\hat{\sigma}$ I would prefer the fit.height3 because it has smallest $\hat{\sigma}$ value.

fit.height($\hat{\sigma}$)= 3.432

fit.height2($\hat{\sigma}$)= 3.444

fit.height3($\hat{\sigma}$)= 3.393

fit.height4($\hat{\sigma}$)= 3.412