Analysis on Dataset “jhs”

by Yimi Zhao

1 Introduction of Dataset

The “jhs” dataset contains only 19 observations across 5 variables. It records the height, weight of 19 persons with their age and sex. The data format is :

jhs name, height, weight, age, sex of junior high school students [sex = 1, M; 2, F]

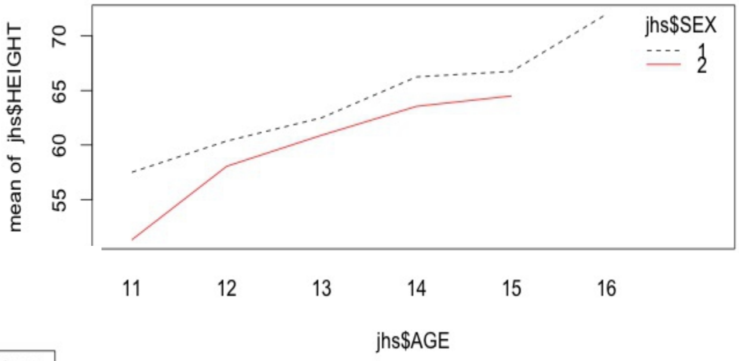
```
> head(jhs)
   NAME  HEIGHT  WEIGHT AGE SEX
1  Alfred    69.0    112.5    14    1
2  Alice     56.5    84.0     13    2
3  Barbara  65.3     98.0     13    2
4  Carol     62.8    102.5    14    2
5  Henry     63.5    102.5    14    1
6  James     57.3    83.0     12    1
```

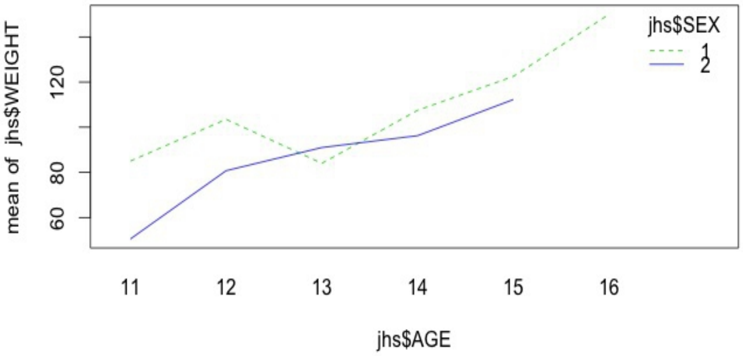
2 Data Analysis

2.1 Data Exploration

First, we explore the dataset a little bit. In total, there are ten boys and nine girls from eleven years old to sixteen years old. From the outputs as below, we can see that **<u>boys are overall taller and heavier than girls. And obviously, both height and weight increase as boys and girls getting older.</u>**

```
# mean values over AGE & SEX
> round(tapply(jhs$HEIGHT, jhs$SEX, mean),2)
   1     2
63.91 60.59
> round(tapply(jhs$WEIGHT, jhs$SEX, mean),2)
    1      2
108.95  90.11
> round(tapply(jhs$HEIGHT, jhs$AGE, mean),2)
   11    12    13    14    15    16
54.40 59.44 61.43 64.90 65.62 72.00
> round(tapply(jhs$WEIGHT, jhs$AGE, mean),2)
    11     12     13     14     15     16
67.75  94.40  88.67 101.88 117.38 150.00
```

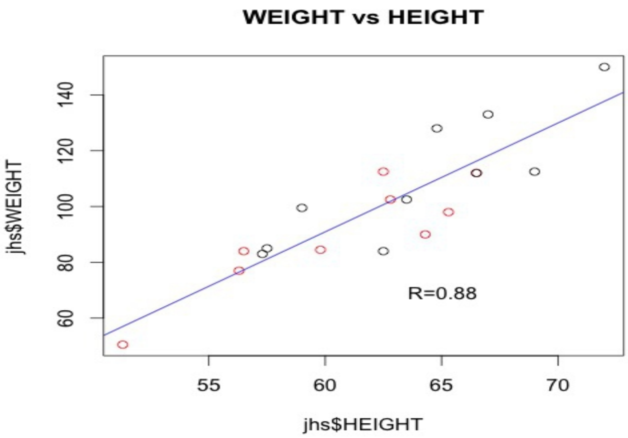




```
# plot
>interaction.plot(jhs$AGE,j hs$SEX, jhs$HEIGHT,
col=c(1,2))
> interaction.plot(jhs$AGE, jhs$SEX, jhs$WEIGHT,
col=c(3,4))
```

Moreover, we also found **<u>weight and height are highly correlated(correlation coefficient=0.88).</u>**

```
# plot WEIGHT vs HEIGHT
> plot(jhs$WEIGHT~jhs$HEIGHT, main="WEIGHT
vs HEIGHT",col=jhs$SEX)
> cor(jhs$WEIGHT, jhs$HEIGHT)
[1] 0.8777852
> abline(lm(WEIGHT~HEIGHT, jhs),col="blue")
> text(locator(1),"R=0.88")
```



2.2 Data Analysis

According to the above data exploration, it looks like that height and weight are quite different based on age and sex. In this case, <u>we want to find out that is there a statistically significant difference in height and weight of these teenagers in terms of their age and sex?</u>

```
#MANOVA
> jhs.manova<-manova(cbind(WEIGHT, HEIGHT)~AGE*SEX,
jhs)
> summary(jhs.manova,test="Pillai")
            Df  Pillai approx   F num    Df den Df     Pr(>F)
AGE          5   1.26994        2.78322   10     16     0.03309 *
SEX          1   0.36451        2.00757    2      7     0.20459
AGE:SEX      4 0.40430     0.50673        8     16     0.83407
Residuals  8
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(jhs.manova, test="Wilks")
            Df   Wilks approx  F num Df den Df Pr(>F)
AGE              5   0.10139  2.99673    10     14   0.0301 *
SEX              1   0.63549  2.00757     2      7    0.2046
AGE:SEX      4   0.61273   0.48565     8     14    0.8469
Residuals  8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(jhs.manova,test="Hotelling-Lawley")
            Df Hotelling-Lawley approx F num Df den Df  Pr(>F)
AGE              5      5.2004  3.12026     10     12   0.03291 *
SEX              1       0.5736  2.00757     2      7     0.20459
AGE:SEX      4      0.6042  0.45319      8     12     0.86621
Residuals  8
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(jhs.manova, test="Roy")
                 Df    Roy approx    F num Df den   Df   Pr(>F)
AGE              5   4.3605       6.9769     5      8   0.008579 **
SEX              1   0.5736       2.0076     2      7   0.204588
AGE:SEX      4    0.5541     1.1082      4      8    0.415920
Residuals  8
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The above results of MANOVA indicate that **<u>there is statistically significant difference in height or weight only in terms of the respondents' age</u>**. For example, by Wilk's test, F(5,8)=2.99673, p<0.05, Wilk's Lambda=0.101.

```
# adjusted manova model
> age.manova<-manova(cbind(WEIGHT, HEIGHT)~AGE, jhs)
> summary(age.manova, test="Wilks")

          Df   Wilks approx F num Df den Df   Pr(>F)
AGE        5 0.17573   3.3251     10     24  0.007685 **
Residuals 13
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary.aov(age. manova)
 Response WEIGHT :
            Df Sum Sq Mean Sq F value   Pr(>F)
AGE          5 6343.9 1268.77   5.513 0.006108 **
Residuals   13 2991.9  230.14

Response HEIGHT :
            Df Sum Sq Mean Sq F value   Pr(>F)
AGE          5 333.30  66.660  6.1957 0.003774 **
Residuals   13 139.87  10.759
```

The adjusted manova model fitting confirmed there is significant difference in both height and weight across the young respondents' age. But till now, we only know that there is at least one pair of age (from 11 to 16) has significant difference in height and weight. Next, we are going to figure out which pairs do?

Combinations of AGE groups	**p-value**	**Combinations of AGE groups**	**p-value**
C(11,12)	0.3836107	C(12,16)	0.08326466
C(11,13)	0.4973971	C(13,14)	0.3169596
C(11,14)	0.1042526	C(13,15)	0.05851187
C(11,15)	0.06446945	C(13,16)	0.1774114
C(11,16)	--	C(14,15)	0.187398
C(12,13)	0.3339918	C(14,16)	0.1174646
C(12,14)	0.0333356*	C(15,16)	0.2180216
C(12,15)	0.05030927		

```
 # pairwise comparison
> for(i in 11:15){
+     for(j in i:16){
+         if(i!=j){
+             if(!(i==11&j==16)){
+              model.s<-summary(manova(cbind(WEIGHT, HEIGHT)~AGE, jhs, subset=jhs$AGE %in% c(i, j)))
+              p<-model.s$stats[[11]]
+              if(p<0.05){
+                  print(cbind(i, j, p))
+              }
+             }
+         }
+     }
+ }
        i   j         p
[1,] 12  14 0.0333356
```

Through pairwise comparison of combinations of “AGE” groups, **it turns out that height and weight show a statistically significant difference across age 12 and age 14 groups, with significant level as 0.05.**