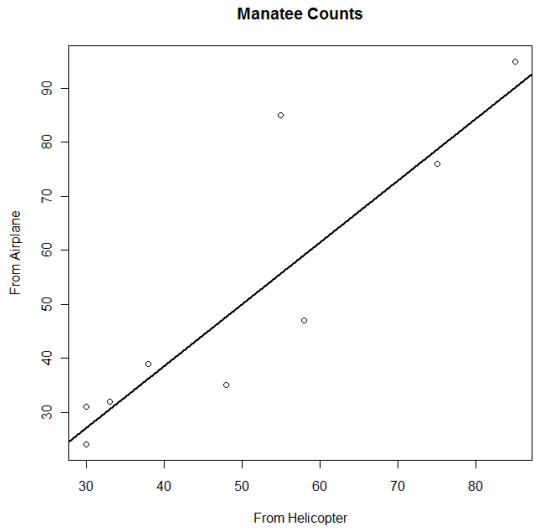
STAT 425 - Homework #2

PROBLEM 2

a) Graph airplane count (Y) versus helicopter count (X), and draw in the estimatedregression line.

```
> air=c(24, 31, 32, 39, 47, 47, 35, 76, 95, 85);  # counts from airplane
> heli=c(30, 30, 33, 38, 58, 58, 48, 75, 85, 55); # counts from helicopter
> lm1=lm(air~heli)
> plot(heli, air, main="Manatee Counts", xlab="From Helicopter", ylab="From Airplane")
> abline(lm1, lwd=2)
```



b) T-tests regarding the slope $\boldsymbol{\beta_1}$

- $\boldsymbol{H_0}: \boldsymbol{\beta_1} = 0$

Find t-value:

```
> summary(lm1)

Call:
lm(formula = air ~ heli)

Residuals:
     Min       1Q   Median       3Q      Max
-12.6551  -9.8500  -0.5449   3.7176  29.3068

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.4632    12.4957  -0.597  0.56685
heli          1.1483     0.2311   4.969  0.00109 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.13 on 8 degrees of freedom
Multiple R-squared: 0.7553,     Adjusted R-squared: 0.7247
F-statistic: 24.69 on 1 and 8 DF,  p-value: 0.001094
```

Or alternative way:

```
> RSS=sum(lm1$res^2)
> df=8
> SE_beta1=sqrt(RSS/df/SSx)
> beta1=lm1$coeff[2]
> t_stat=beta1/SE_beta1
> t_crit=qt(0.975,df)
> t_stat
    heli
4.969151
> t_crit
[1] 2.306004
> t_stat>t_crit
heli
TRUE
```

- $RSS = \sum_1^n (y_i - \hat{y}_i)^2, SSx = \sum_1^n (x_i - \bar{x})^2$
- $se(\hat{\beta}_1) = \frac{RSS/(n-2)}{SSx}$
- $t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 4.969$, $t_{0.975,8} = 2.3$
- $t > t_{0.975,8} \rightarrow$ Reject null $\boldsymbol{H_0}: \boldsymbol{\beta_1} = 0 \rightarrow \boldsymbol{\beta_1} \neq 0$
- Linear model seems suitable since coefficient significantly different than zero

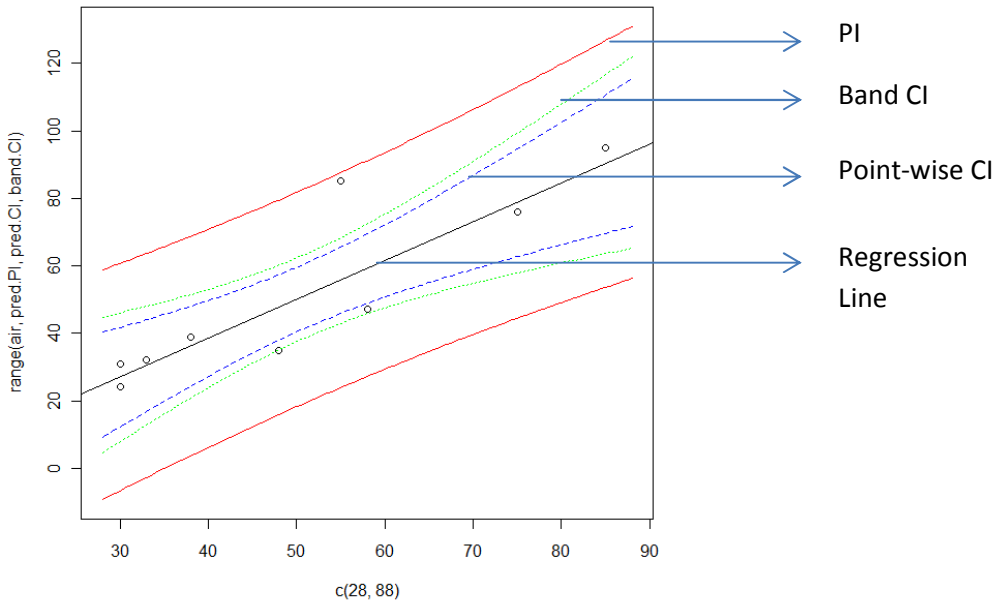
- $\boldsymbol{H_0}: \boldsymbol{\beta_1} = \mathbf{1}$

```
> g=lm(air~offset(heli));   #set coefficient of helicopter variable to 1
> anova(g,lm1)
Analysis of Variance Table

Model 1: air ~ offset(heli)
Model 2: air ~ heli
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      9 1450.9
2      8 1379.9  1    71.034 0.4118  0.539
> g=lm(air~offset(heli));   #set coefficient of helicopter variable to 1
> anova(g,lm1)
Analysis of Variance Table

Model 1: air ~ offset(heli)
Model 2: air ~ heli
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      9 1450.9
2      8 1379.9  1    71.034 0.4118  0.539
>
> t_stat1=(beta1-1)/SE_beta1
> t_stat1
    heli
0.6417426
> t_crit
[1] 2.306004
> t_stat1>t_crit
 heli
FALSE
```

- Compare original model with $y = \beta_0 + x$ (set $\beta_1 = 1$)
- P-value = 0.539 >0.05, so cannot reject null hypothesis $\beta_1 = 1$. → Keep simpler model ($\beta_1 = 1$)
- Alternatively, $t = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)}$=0.6417 <$t_{0.975,8} = 2.0306$→ Cannot reject null $\beta_1 = 1$

- Which test seems more relevant to this study?
  - The second test ($\beta_1 = 1$) seems more relevant to this study because the airplane and helicopter covered the same area, so it should be expected that the count would be similar, which means the slope of the line should be equal to 1. From the study, we can logically deduce that the slope will be different than zero because the more counts from helicopter means also a higher count from airplanes (since the area is the same).

c) Explain plotted figures. Explain why the blue intervals are shorter than the red intervals.Explain why the blue intervals are shorter than the green intervals.



- Plotted Figures
  - The plotted figure shows the original data points (x=heli, y=air) and the regression line using the linear model from part a.
  - The prediction intervals (red, solid), prediction confidence intervals (blue,dash) and confidence interval bands (green,dots) are shown for the range x=28 to x=88.
- Blue intervals are shorter than the red intervals
  - The point-wise CI (blue) is used to predict the mean value of y (dependent variable) at given x (independent variable). (Interpretation: in the long run, 95% of these intervals will include the corresponding future values of x).
  - The PI (red) is used to predict the value of one y (dependent variable) at given x (independent variable). (Interpretation: in the long run, 95% of these intervals will include the corresponding true mean).
  - The PI (red) is always wider than the point-wise CI (blue) since the PI is an interval estimate for a single future observation (uncertainty: estimated coefficients + random

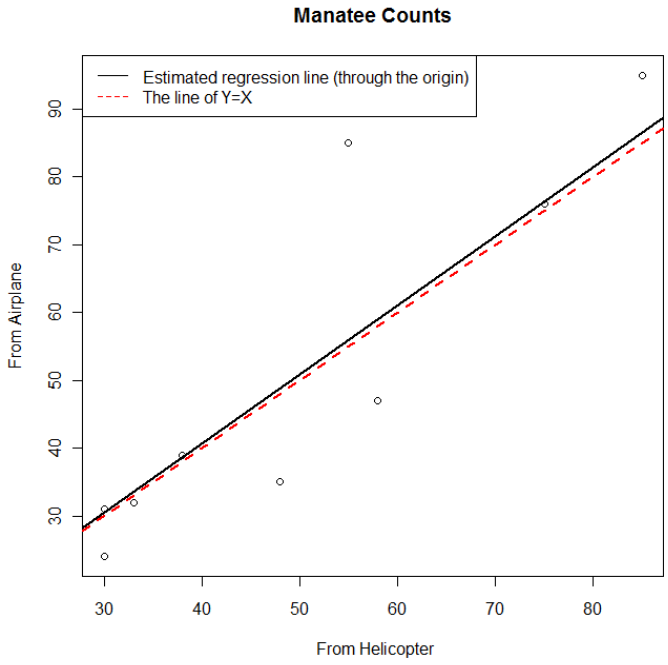
error of new observation), while the point-wise CI is an interval estimate for an average (uncertainty: estimated coefficients). From the equation, the formula used for PI has an extra 1 (under the square root) compared to that of the point-wise CI. Thus, the width of the PI (red) is always wider than the point-wise CI's (blue).

- Blue intervals are shorter than the green intervals.
  - The point-wise CI (blue) is used to predict the mean value of y (dependent variable) at given x (independent variable). Each point-wise CI separately represents the probability (with 1-α % confidence) of covering the true value for its corresponding value of x.
  - The CI band (green) represents the probability of simultaneously covering the true value (with 1-α % confidence) of all the corresponding values of x. That is, each CI in the band all cover their corresponding true values simultaneously.
  - Thus, the point-wise CI (blue) is shorter than the CI band (green) since the simultaneous coverage probability is less than the individual coverage probability. That is, for the same coverage probability (e.g. 95%), the CI band (green) needs to be wider than the point-wise CI (blue).

d) R-code

```
> lm2=lm(air~heli-1)
> plot(heli, air, main="Manatee Counts", xlab=" From Helicopter", ylab="From Airplane")
> abline(lm2, lwd=2)
> abline(0,1,lwd=2, lty=2, col="red")
> legend("topleft", lty=c(1,2), col=c("black", "red"),
+  legend=c("Estimated regression line (through the origin)", "The line of Y=X"))
```

- Plot



- Is the slope in this graph significantly different from 1?

```
> summary(lm2)

Call:
lm(formula = air ~ heli - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8700 -10.6744  -0.9788   0.4200  29.0031

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
heli  1.01813    0.07401   13.76 2.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.66 on 9 degrees of freedom
Multiple R-squared: 0.9546,     Adjusted R-squared: 0.9496
F-statistic: 189.3 on 1 and 9 DF,  p-value: 2.386e-07

> g2=lm(air~offset(heli)-1);   #set coefficient of helicopter variable to 1
> anova(g2,lm2)
Analysis of Variance Table

Model 1: air ~ offset(heli) - 1
Model 2: air ~ heli - 1
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     10 1451.0
2      9 1441.4  1    9.6067 0.06  0.812
```

  - Slope = 1.01813
  - So it is very close to 1
  - Compare linear model through origin (lm2) and linear model through origin with slope=1 (g2). Doing a F-test to test $\beta_1 = 1$, leads to the conclusion that it cannot be rejected because the p-value = 0.812 > 0.05. Thus, there is *no* evidence to suggest that the slope is significantly different from 1 (i.e. keep the null that $\beta_1 = 1$).

PROBLEM 3

R Code and Output:

```
> n=40; x=runif(n, -5,5); y=1+2*x+3*rnorm(n);
> x.quan=quantile(x); # return the 5-number summary
> id=(1:n)[x < x.quan[2]];
> id=c(id, (1:n)[x > x.quan[4]])
>
> # id: a subset of the n obs whose x-values are outside
> # the 25% and 75% quantiles of x.
> length(id)# should equal 20=n/2
[1] 20
>
> myfit0=lm(y~x);
> myfit1=lm(y[id]~x[id])
> myfit2=lm(y[-id]~x[-id])
>
> summary(myfit0)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3402 -1.8492  0.2637  2.6559  5.3049

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7811     0.5164   1.513    0.139
x             2.0678     0.1907  10.843 3.43e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.195 on 38 degrees of freedom
Multiple R-squared: 0.7557,     Adjusted R-squared: 0.7493
F-statistic: 117.6 on 1 and 38 DF,  p-value: 3.433e-13

> summary(myfit1)

Call:
lm(formula = y[id] ~ x[id])

Residuals:
   Min     1Q Median     3Q    Max
-7.296 -1.475  0.289  1.619  4.101

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8311     0.6592   1.261    0.223
x[id]         2.0884     0.1888  11.063 1.85e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.911 on 18 degrees of freedom
Multiple R-squared: 0.8718,     Adjusted R-squared: 0.8647
F-statistic: 122.4 on 1 and 18 DF,  p-value: 1.846e-09

> summary(myfit2)

Call:
lm(formula = y[-id] ~ x[-id])

Residuals:
    Min      1Q  Median      3Q     Max
-6.1156 -2.6331  0.1579  3.3002  5.2117

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6768     0.8671   0.781  0.44524
x[-id]        1.9533     0.5513   3.543  0.00232 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.611 on 18 degrees of freedom
Multiple R-squared: 0.4108,     Adjusted R-squared: 0.3781
F-statistic: 12.55 on 1 and 18 DF,  p-value: 0.002324
```

```
> # produce 4 plots arranged in a 2x2 table
> par(mfrow=c(2,2))
> plot(x,y, xlab="", ylab="", main='myfit0');
> abline(myfit0);
>
> plot(x,y,xlab="", ylab="", type="n" ,main='myfit1');
> points(x[id],y[id],col=2);
> abline(myfit1, lty=2, col=2);
>
> plot(x,y,xlab="", ylab="", type="n", main='myfit2');
> points(x[-id],y[-id],col=3);
> abline(myfit2, lty=3, col=3);
>
> plot(x,y, xlab="", ylab="",  main='all models');
> abline(myfit0);
> abline(myfit1, lty=2, col=2);
> abline(myfit2, lty=3, col=3);
>
> # comapre RSS and TSS
>
> Beta0 = round(c(myfit0$coeff[1],myfit1$coeff[1],myfit2$coeff[1]),3)
> Beta1 = round(c(myfit0$coeff[2],myfit1$coeff[2],myfit2$coeff[2]),3)
> RSS = round(c(sum(myfit0$res^2),sum(myfit1$res^2),sum(myfit2$res^2)),3)
> TSS = round(c(sum((y-mean(y))^2),sum((y[id]-mean(y[id]))^2),sum((y[-id]-mean(y[-id]))^2)),3)
> Ave_RSS = round(c(RSS[1]/length(y),RSS[2]/length(id), RSS[3]/length(id)),3)
> Ave_TSS = round(c(TSS[1]/length(y),TSS[2]/length(id), TSS[3]/length(id)),3)
> RSS_TSS_ratio = round(RSS/TSS,3)
> Rsq = round(1-RSS/TSS,3)
> report=rbind(RSS,TSS,Ave_RSS,Ave_TSS,RSS_TSS_ratio,Rsq)
> Beta0
(Intercept) (Intercept) (Intercept)
      0.781       0.831       0.677
> Beta1
     x  x[id] x[-id]
 2.068  2.088  1.953
> report
                [,1]     [,2]    [,3]
RSS          387.952  152.546 234.684
TSS         1588.294 1189.789 398.343
Ave_RSS        9.699    7.627  11.734
Ave_TSS       39.707   59.489  19.917
RSS_TSS_ratio  0.244    0.128   0.589
Rsq            0.756    0.872   0.411
```

a) Report the LS estimates of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ from myfit0, myfit1, and myfit2. Are theysimilar?

	myfit0	myfit1	myfit2
$\boldsymbol{\beta_0}$=	0.781	0.831	0.677
$\boldsymbol{\beta_1}$ =	2.068	2.088	1.953

- Yes, β_0 (range: 0.677 to 0.831) and β_1(range: 1.953 to 2.088) from myfit0, myfit1, and myfit2 are very similar.

b) Report $\boldsymbol{R^2}$ from myfit0, myfit1, and myfit2. Are they similar? If not, explain why.

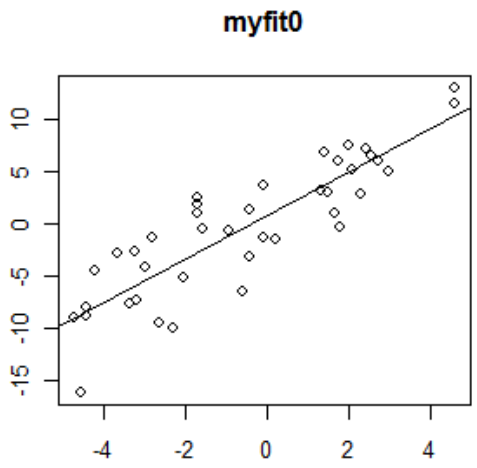
	myfit0	myfit1	myfit2
$\boldsymbol{R^2}$	0.756	0.872	0.411

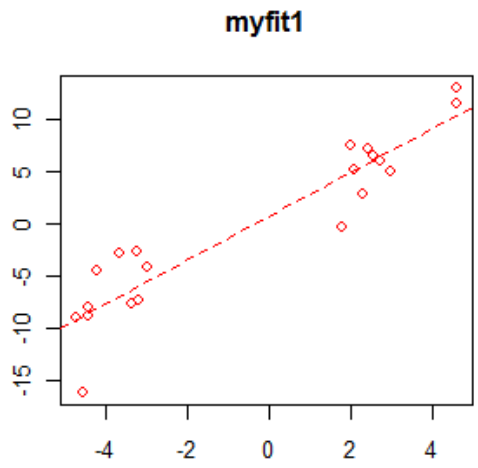
- The values are not similar, since difference is at least 0.1 (10%).
- Although the linear regression lines are similar (subsets of the same data are used), the coefficients of determination are not.

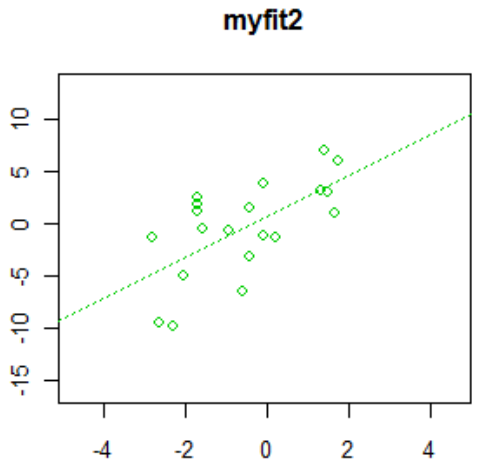
	myfit0	myfit1	myfit2
RSS	387.952	152.546	234.684
TSS	1588.294	1189.789	398.343
Ave_RSS	9.699	7.627	11.734
Ave_TSS	39.707	59.489	19.917
RSS_TSS_ratio	0.244	0.128	0.589
R^2	0.756	0.872	0.411

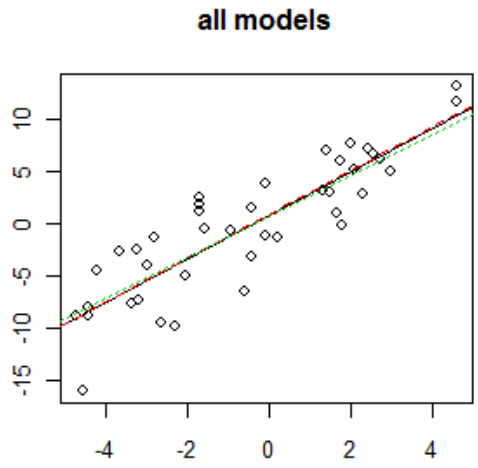
- Looking at the report table, we can note the difference between RSS.
  - RSS: myfit0 > myfit2 > myfit1
  - The RSS will be lower for myfit1 and myfit2 since fewer residuals (40 vs 20) are summed up in the RSS, even though the individual corresponding terms (residuals) in the sum will be similar.
  - However, looking at a scaled RSS (Ave_RSS) found by dividing the RSS by its corresponding number of observations (note this is not the mean RSS which is divided by the deg. of freedom), suggests that the scaled RSS are very similar for the 3 models.
- Looking at the report table, we can note the difference between RSS:
  - TSS: myfit0 > myfit1 > myfit2
  - The TSS will be lower for myfit1 and myfit2 since fewer terms (40 vs 20) are summed up in the TSS, even though the individual corresponding terms in the sum will be similar. The TSS for myfit1 is greater than myfit2 since the data points in myfit1 (bottom 25% and top 75%) are in the extremes, and hence a farther from the mean.
  - However, looking at a scaled TSS (Ave_RSS) found by dividing the TSS by its corresponding number of observations (note this is not the mean TSS which is divided by the degree of freedom), suggests that the scaled TSS is much greater for myfit1, followed by myfit 0 and myfit2. Thus, taking into account the number of observations, the TSS (scaled) is much greater for myfit1 (since points are further from the mean) and much smaller for myfit2 (since points are closer to the mean).
- Therefore, the effective ratio RSS/TSS, which is the same as Ave_RSS/Ave_TSS, is ranked as myfit2 > myfit0 > myfit1, since the numerators (Ave_RSS) is about the same but the denominators (Ave_TSS) are ranked as myfit2 < myfit0 < myfit1
- Thus, $\boldsymbol{R^2}$ will be larger for smaller ratios, and thus the ranking will be: myfit1>myfit0>myfit1
- Thus $\boldsymbol{R^2}$is artificially inflated by selective sampling the extreme points (i.e. use the bottom and top data points), while $\boldsymbol{R^2}$ is artificially deflated by selective sampling the middle points.
- $RSS = \sum_1^n (y_i - \hat{y}_i)^2, TSS = \sum_1^n (y_i - \bar{y})^2, R^2 = 1 - RSS/TSS$
- Looking at the plots confirms the points made above:
  - Scaled RSS (squared distance from regression line to observation) will be similar for all models
  - Scaled TSS (squared distance from observation to mean of data) will be greater for the model with extreme points (myfit1) while smaller for the model with middle points (myfit0)

- Plots:


myfit0
10
5
0
-5
-10
-15
-4
-2
0
2
4


myfit1
10
5
0
-5
-10
-15
-4
-2
0
2
4


myfit2
10
5
0
-5
-10
-15
-4
-2
0
2
4


all models
10
5
0
-5
-10
-15
-4
-2
0
2
4

PROBLEM 4

Variables:

- Sex: 0 = males, 1 = females
- WT2: Age 2 weight (kg)
- HT2: Age 2 height (cm)
- WT9: Age 9 weight (kg)
- HT9: Age 9 height (cm)
- LG9: Age 9 leg circumference (cm)
- ST9: Age 9 strength (higher value = stronger)
- WT18: Age 18 weight (kg)
- HT18: Age 18 height (cm)
- LG18: Age 18 leg circumference (cm)
- ST18: Age 18 strength (higher value = stronger)
- Soma: Somatotype, a seven-point scale, as a measure of fatness (1 = slender, 7 = fat), determined using a photograph taken at age 18

a) For boys: $\boldsymbol{SOMA = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9 + e}$

- R-Code and Output:

```
> lm1.boy=lm(Soma~HT2 + WT2 + HT9 + WT9 , subset=(Sex==0), data=BGS)
> summary(lm1.boy)

Call:
lm(formula = Soma ~ HT2 + WT2 + HT9 + WT9, data = BGS, subset = (Sex ==
    0))

Residuals:
    Min      1Q  Median      3Q     Max
-2.2896 -0.8326 -0.0671  0.7614  3.3530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.75692    5.08000   3.102  0.00291 **
HT2         -0.01500    0.07227  -0.208  0.83627
WT2         -0.41382    0.13278  -3.117  0.00279 **
HT9         -0.08070    0.04677  -1.725  0.08955 .
WT9          0.16642    0.03628   4.587 2.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.252 on 61 degrees of freedom
Multiple R-squared: 0.299,      Adjusted R-squared: 0.253
F-statistic: 6.503 on 4 and 61 DF,  p-value: 0.0001997

> lm1.boy.rsquared=summary(lm1.boy)$r.squared
> lm1.boy.coeff=lm1.boy$coeff
> lm1.boy.sigma=summary(lm1.boy)$sigma
> round(lm1.boy.rsquared,4)
[1] 0.299
> round(lm1.boy.coeff,4)
(Intercept)         HT2         WT2         HT9         WT9
    15.7569     -0.0150     -0.4138     -0.0807      0.1664
> round(lm1.boy.sigma,4)
[1] 1.2523
> cor(BGS$WT2[BGS$Sex==0],BGS$HT2[BGS$Sex==0])
[1] 0.6087297
> cor(BGS$WT2[BGS$Sex==0],BGS$HT9[BGS$Sex==0])
[1] 0.5201342
> cor(BGS$WT2[BGS$Sex==0],BGS$WT9[BGS$Sex==0])
[1] 0.5911935
> cor(BGS$WT9[BGS$Sex==0],BGS$HT2[BGS$Sex==0])
[1] 0.4871889
> cor(BGS$WT9[BGS$Sex==0],BGS$HT9[BGS$Sex==0])
[1] 0.6503651
```

- Report:

$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\sigma}$	R^2
15.7569	-0.0150	-0.4138	-0.0807	0.1664	1.2523	0.299

- According to the summary in R, the only p-values of the predictor's coefficients smaller than 0.05 correspond to WT2 and WT9. Thus, WT2 and WT9 are the significant variables in this model.
- The overall F-test determines whether any of the predictors have significance in the model.
  - $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  - $H_1: at\ least\ on\ of\ \beta_1, \beta_2,\ or\ \beta_3 is\ \neq 0$
  - From the R summary, the p-value is very small (0.0001997). Thus, reject the null hypothesis and conclude the response is linearly related to one or more of the predictors (when controlling for the rest).

- The sign of some of $\hat{\beta}$ seems to be unexpected. For example, heavier boys at age 2 tend to be thinner (have lower SOMA) at age 18. Explain
  - The problem of the unexpected sign in the coefficient can be attributed to the existence of multicollinearity, that is the existence of highly correlated independent variables. This may distort the size and sign of coefficient estimates. Therefore, the sign of the coefficients may become meaningless and interpretation should be done with caution due to the problems in estimating parameter when highly correlated variables are present. In this case, there is a high positive correlation for (0.59) between WT2 and WT9 for boys. The latter should dominate the model since weight at age 9 should be more relevant than weight at age 2.
  - Note: an interpretation of the coefficient WT2 is as follows: given the same weight at age 9 (and also controlling for the other variables), a heavier boy at age 2 will tend to be thinner (have lower SOMA) at age 18. If they have the same weight at age 9, that means that the heavier boy lost more weight from age 2 to 9. Thus, it some sense, the negative sign of $\hat{\beta}$ makes sense in that a heavier boy at age 2 (with the same weight at age 9) is probably more likely to lose more weight from age 9 to 18 since he already lost more weight up until age 9 (e.g. maybe due to faster metabolism or genetics).

b) For boys: $\boldsymbol{SOMA = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 DHT9 + \beta_4 DWT9 + e}$ where DWT9 = WT9 - WT2, DHT9 = HT9 - HT2

- R-Code and Output:

```
> lm2.boy=lm(Soma~HT2 + WT2 + I(HT9-HT2) + I(WT9-WT2) , subset=(Sex==0), data=BGS)
> summary(lm2.boy)

Call:
lm(formula = Soma ~ HT2 + WT2 + I(HT9 - HT2) + I(WT9 - WT2),
    data = BGS, subset = (Sex == 0))

Residuals:
    Min      1Q  Median      3Q     Max
-2.2896 -0.8326 -0.0671  0.7614  3.3530

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.75692    5.08000   3.102  0.00291 **
HT2          -0.09570    0.06054  -1.581  0.11913
WT2          -0.24740    0.12279  -2.015  0.04833 *
I(HT9 - HT2) -0.08070    0.04677  -1.725  0.08955 .
I(WT9 - WT2)  0.16642    0.03628   4.587 2.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.252 on 61 degrees of freedom
Multiple R-squared: 0.299,      Adjusted R-squared: 0.253
F-statistic: 6.503 on 4 and 61 DF,  p-value: 0.0001997

> lm2.boy.rsquared=summary(lm2.boy)$r.squared
> lm2.boy.coeff=lm2.boy$coeff
> lm2.boy.sigma=summary(lm2.boy)$sigma
> round(lm2.boy.rsquared,4)
[1] 0.299
> round(lm2.boy.coeff,4)
 (Intercept)          HT2          WT2 I(HT9 - HT2) I(WT9 - WT2)
     15.7569      -0.0957      -0.2474      -0.0807       0.1664
> round(lm2.boy.sigma,4)
[1] 1.2523
```

- Report:

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}$	R^2
15.7569	**-0.0957**	**-0.2474**	-0.0807	0.1664	1.2523	0.299

- $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}, \widehat{\boldsymbol{\sigma}}$ and $\boldsymbol{R^2}$ are the same, while $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$ are different .
- This is because the second model uses the same data set with the same predictors, just added linearly different.
- Note that by expanding the regression equation, the coefficient of HT2 = **-0.0957 – (**-0.0807)= -0.0150, which is equal to the coefficient of HT2 in part a. Similarly, WT2 = **-0.2474 – (**0.1664)=-0.4138, which is equal to the coefficient of WT2 in part a.
- Thus, for variables with repeated terms (WT2 and HT2) the coefficients will be different (since when combined, they have to equal the coefficient of the original equation when they appeared only once).
- For the intercept and variables repeated only once (HT2 and HT9), the coefficient will be the same (since they have to equal the coefficient of the original equation).
- The other values ($\hat{\sigma}$ and R^2) are the same since the equation itself and predicted values have not changed.

- According to the summary in R, the only p-values of the predictor's coefficients smaller than 0.05 correspond to WT2 and DWT9 (WT9 - WT2). Thus, WT2 and DWT9 are the significant variables in this model.These are the same results compared to 4(a), in which only the weights were found to be significant in the corresponding model. Note that since the coefficient ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$) of WT2 has a smaller effect (because of its additional term in DWT9), the p-value is larger (closer to 0.05), meaning less significant compared to part a.

c) For boys: $\boldsymbol{SOMA = \beta_0 + \beta_3 HT9 + \beta_4 WT9 + e}$

```
> lm3.boy=lm(Soma~HT9 + WT9 , subset=(Sex==0), data=BGS)
> summary(lm3.boy)

Call:
lm(formula = Soma ~ HT9 + WT9, data = BGS, subset = (Sex == 0))

Residuals:
     Min       1Q   Median       3Q      Max
-2.26294 -1.12772 -0.01427  0.73258  3.26242

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.17251    4.91188   3.089  0.00299 **
HT9         -0.11637    0.04099  -2.839  0.00608 **
WT9          0.11797    0.03581   3.295  0.00162 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.352 on 63 degrees of freedom
Multiple R-squared: 0.1566,     Adjusted R-squared: 0.1298
F-statistic: 5.848 on 2 and 63 DF,  p-value: 0.004681

> anova(lm3.boy,lm1.boy)
Analysis of Variance Table

Model 1: Soma ~ HT9 + WT9
Model 2: Soma ~ HT2 + WT2 + HT9 + WT9
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     63 115.088
2     61  95.661  2    19.427 6.1941 0.003556 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- According to the F-test, the p-value (0.003556) < 0.05, which means we should reject the null hypothesis (the reduced model is better Ho: $\beta_0 + \beta_3 HT9 + \beta_4 WT9$) and accept the alternative full model (Ha : $\beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9$). Thus, the simultaneous inclusion of the variables (HT2 and WT2) has a significant effect and we should accept the full model proposed in part (a) instead of the reduced model from part (b).

d) For girls $\boldsymbol{SOMA = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9 + e}$

- R-Code and Output:

```
> lm1.girl=lm(Soma~HT2 + WT2 + HT9 + WT9, subset=(Sex==1), data=BGS)
> summary(lm1.girl)

Call:
lm(formula = Soma ~ HT2 + WT2 + HT9 + WT9, data = BGS, subset = (Sex ==
    1))

Residuals:
     Min       1Q   Median       3Q      Max
-2.04024 -0.33675  0.01377  0.37665  1.00031

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.95014    2.29227   4.341 5.07e-05 ***
HT2         -0.06618    0.03473  -1.906   0.0611 .
WT2         -0.06761    0.07419  -0.911   0.3655
HT9         -0.01983    0.02326  -0.852   0.3971
WT9          0.13120    0.02046   6.411 1.88e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5852 on 65 degrees of freedom
Multiple R-squared: 0.5034,     Adjusted R-squared: 0.4728
F-statistic: 16.47 on 4 and 65 DF,  p-value: 2.295e-09

> lm1.girl.rsquared=summary(lm1.girl)$r.squared
> lm1.girl.coeff=lm1.girl$coeff
> lm1.girl.sigma=summary(lm1.girl)$sigma
> round(lm1.girl.rsquared,4)
[1] 0.5034
> round(lm1.girl.coeff,4)
(Intercept)         HT2         WT2         HT9         WT9
     9.9501     -0.0662     -0.0676     -0.0198      0.1312
> round(lm1.girl.sigma,4)
[1] 0.5852
> cor(BGS$WT2[BGS$Sex==1],BGS$HT2[BGS$Sex==1])
[1] 0.6445495
> cor(BGS$WT2[BGS$Sex==1],BGS$HT9[BGS$Sex==1])
[1] 0.6071247
> cor(BGS$WT2[BGS$Sex==1],BGS$WT9[BGS$Sex==1])
[1] 0.692539
> cor(BGS$WT9[BGS$Sex==1],BGS$HT2[BGS$Sex==1])
[1] 0.5229277
> cor(BGS$WT9[BGS$Sex==1],BGS$HT9[BGS$Sex==1])
[1] 0.7276123
```

- Report:

$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\sigma}$	R^2
9.9501	-0.0662	-0.0676	-0.0198	0.1312	0.5852	0.5034

- According to the summary in R, the only p-value of the predictor's coefficients smaller than 0.05 correspond to WT9. Thus, WT9 is the significant variable in this model.

- The overall F-test determines whether any of the predictors have significance in the model.
  - $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  - $H_1: at\ least\ on\ of\ \beta_1, \beta_2,\ or\ \beta_3 is\ \neq 0$
  - From the R summary, the p-value is very small (2.295e-09). Thus, reject the null hypothesis and conclude the response is linearly related to one or more of the predictors (when controlling for the rest).
- The sign of some of $\widehat{\beta}$ seems to be unexpected. For example, heavier girls at age 2 tend to be thinner (have lower SOMA) at age 18. Explain
  - The problem of the unexpected sign in the coefficient can be attributed to the existence of multicollinearity. See answer in part a for detailed response.
  - For girls, WT2 is not significant at the 0.05 level.

e) For girls: $\boldsymbol{SOMA = \beta_0 + \beta_3 HT9 + \beta_4 WT9 + e}$

```
> lm3.girl=lm(Soma~HT9 + WT9, subset=(Sex==1), data=BGS)
> summary(lm3.girl)

Call:
lm(formula = Soma ~ HT9 + WT9, data = BGS, subset = (Sex == 1))

Residuals:
     Min       1Q   Median       3Q      Max
-1.91683 -0.30528  0.03317  0.39234  1.28122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.13136    2.18425   3.723 0.000406 ***
HT9         -0.05363    0.01901  -2.822 0.006278 **
WT9          0.12313    0.01832   6.723 4.73e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6078 on 67 degrees of freedom
Multiple R-squared: 0.4477,     Adjusted R-squared: 0.4312
F-statistic: 27.15 on 2 and 67 DF,  p-value: 2.311e-09

> anova(lm3.girl,lm1.girl)
Analysis of Variance Table

Model 1: Soma ~ HT9 + WT9
Model 2: Soma ~ HT2 + WT2 + HT9 + WT9
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     67 24.755
2     65 22.258  2    2.4964 3.6452 0.03159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- According to the F-test, the p-value (0.003556) < 0.05, which means we should reject the null hypothesis (the reduced model is better $Ho: \beta_0 + \beta_3 HT9 + \beta_4 WT9$) and accept the alternative full model ($Ha:\ \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9$). Thus, the simultaneous inclusion of the variables (HT2 and WT2) has a significant effect and we should accept the full model proposed in part (d) instead of the reduced model from part (e).

f) Compare and contrast results for boys and girls

Sex	$\widehat{\boldsymbol{\beta}}_0$	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\beta}}_3$	$\widehat{\boldsymbol{\beta}}_4$	$\widehat{\boldsymbol{\sigma}}$	$\boldsymbol{R^2}$
Boy	15.7569	-0.0150	-0.4138	-0.0807	0.1664	1.2523	0.299
Girl	9.9501	-0.0662	-0.0676	-0.0198	0.1312	0.5852	0.5034

- The results for the full model (part a. and part d.) are different for boys and girls. For boys, both the weight at age 2 and 9 (WT2 and WT9) were determined to be significant. On the other hand, for girls only the weight at age 9 (WT9) was significant for the proposed model.
- The full model (part a. and part d.) seems to be a better fit to the data for girls rather than boys, as illustrated by the larger R^2 and smaller $\hat{\sigma}$ for girls.
- The corresponding signs of the coefficients for the full models did not change
- The corresponding relative sizes of the coefficients for the full models were about the same except for the WT2, which is much larger for boys.
- The F-test for the respective models in part a. and part d. had the same result: the response is linearly related to one or more of the predictors (when controlling for the rest)
- The F-test for the respective models in part a-c. and part d-e had the same result: accept the full models instead of the reduced models.
- The correlations between predictors were higher for girls than that of boys.