KIM_JAEJIN_2122_Proj2

Jaejin Kim

3/21/2019

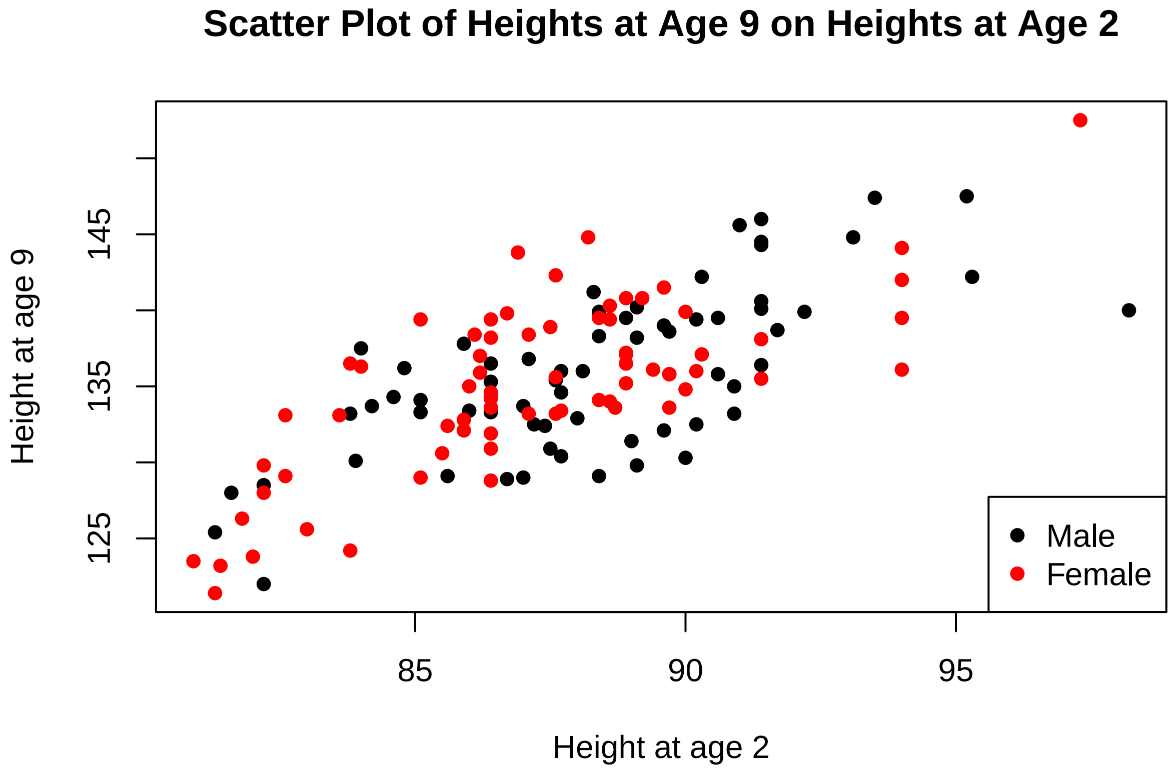
```r
data <- read.csv("BGS.csv")
```

1.

(a)

```r
##Plot a scatter plot of heights at age 9 on heights at age 2
plot(data$HT2, data$HT9, pch=16,
     col = data$Sex + 1,
     xlab = "Height at age 2", ylab = "Height at age 9",
     main = "Scatter Plot of Heights at Age 9 on Heights at Age 2")

legend(x="bottomright", legend = c("Male", "Female"), col = c(1, 2), pch=16)
```



- Looking at the scatter plot, it seems that the patterns for boys and girls are similar. They seem to have similar intercepts, but might have different slopes. We will need to investigate further.

(b)

```
##Fit a simple linear regression
lm1.1 <- lm(HT9 ~ HT2, data = data)
summary(lm1.1)
```

```
##
## Call:
## lm(formula = HT9 ~ HT2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7938 -2.4884 -0.0801  2.9806  9.3631
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.92705    8.59960   3.713    3e-04 ***
## HT2          1.17963    0.09788  12.052   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.822 on 134 degrees of freedom
## Multiple R-squared:  0.5201, Adjusted R-squared:  0.5166
## F-statistic: 145.2 on 1 and 134 DF,  p-value: < 2.2e-16
```

- The intercept is 31.92705 and the estimated regression coefficient of HT2 is 1.17963, and both are significant at $\alpha = 0.05$ significance level. However, the intercept cannot be interpreted here, as it would require that an individual to be 0cm tall at the age of 2, which is impossible. The estimated coefficient suggests that if we were to compare two groups of people who were equivalent on all other variables and one unit (1cm) apart in the height at age 2, then the expected value of height at age 9 of the group that was 1cm taller would be bigger by 1.180 compared to the group that had shorter height at age 2.

```
##Test the hypothesis of H_0: \beta_1 = 0
#Calculate t-statistic
n <- nrow(data)
rss <-  sum((data$HT9 - lm1.1$fitted.values)^2)
se_b <- sqrt((1 / (n - 2))*rss / (var(data$HT2) * (n - 1)))
t_statistic <- lm1.1$coefficients[[2]] / se_b
t_statistic
```

```
## [1] 12.05192
```

```
#Calculate p-value
p_value = 2*pt(t_statistic, df = n-2, lower=FALSE)
p_value
```

```
## [1] 4.085372e-23
```

- We reject the null hypothesis of $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ significance level.

```
##Numerically show that the value of the T-statistic is equal to
##the square root of the F-statistic from the ANOVA
#Calculate F-statistic
ssreg <- sum((lm1.1$fitted.values - mean(data$HT9))^2)
F_statistic <- (ssreg  / 1) / (rss / (n - 2))
F_statistic
```

```
## [1] 145.2487
```

```
all.equal(t_statistic, sqrt(F_statistic))
```

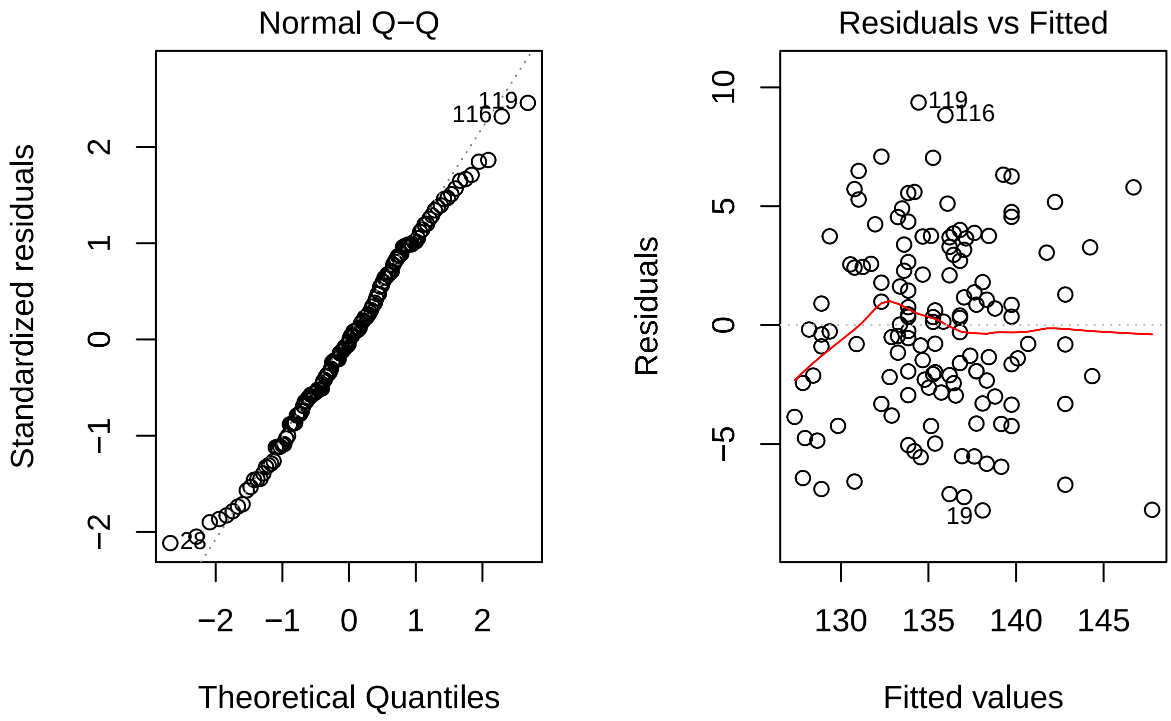
```
## [1] TRUE
```

```
#Alternatively, use regression outputs
t_2 <- summary(lm1.1)$coefficients[2,3]
F_2 <- summary(lm1.1)$fstatistic[[1]]
all.equal(t_2, sqrt(F_2))
```

```
## [1] TRUE
```

- We can see that the value of the T-statistic for the above hypothesis test is equal to the square root of the F-statistic from the ANOVA at the bottom of the regression output.

```
##Check the normality and homoscedasticity assumptions on the residuals.
par(mfrow=c(1, 2))
plot(lm1.1, 2) # normality of residuals
plot(lm1.1, 1) # homoscedasticity of residuals
```



- The first plot shows if the residuals are normally distributed. While the residuals are generally in a straight line suggesting they align with the normal distribution, we can see that the residuals on the extreme ends do not align perfectly. The second plot shows if residuals are spread equally along the ranges of predictors. We can see that while the residuals are generally homoscedastic in the middle and in the high end, it is somewhat heteroscedastic in the lower end.

(c)

```
##Generate a model that allows for separate intercepts for boys and girls.
lm1.2 <- lm(HT9 ~ HT2 + as.factor(Sex), data = data)
summary(lm1.2)
```

```
##
## Call:
## lm(formula = HT9 ~ HT2 + as.factor(Sex), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6223 -2.5692  0.0397  2.9872  9.1012
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     30.39838    8.79454   3.457 0.000735 ***
## HT2              1.19373    0.09938  12.012  < 2e-16 ***
## as.factor(Sex)1  0.56562    0.66571   0.850 0.397051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.826 on 133 degrees of freedom
## Multiple R-squared:  0.5227, Adjusted R-squared:  0.5156
## F-statistic: 72.83 on 2 and 133 DF,  p-value: < 2.2e-16
```

```
AIC(lm1.1, lm1.2)
```

```
##       df      AIC
## lm1.1  3 754.5998
## lm1.2  4 755.8636
```

```
BIC(lm1.1, lm1.2)
```

```
##       df      BIC
## lm1.1  3 763.3378
## lm1.2  4 767.5143
```

The adjusted R-squared of the model that allows for separate intercepts for boys and girls is 0.5156 while that of the base model is 0.5166. Also, both AIC and BIC values are lower for the base model. Therefore, this model is not significantly better than the simple linear regression fit above.

(d)

```
##Generate a model that allows for both
##the separate slope and separate intercepts for boys and for girls.
lm1.3 <- lm(HT9 ~ HT2*as.factor(Sex), data = data)
summary(lm1.3)
```

```
##
## Call:
## lm(formula = HT9 ~ HT2 * as.factor(Sex), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4457 -2.5821 -0.1209  2.9664  9.1191
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          35.1732    12.6724   2.776  0.00631 **
## HT2                   1.1397     0.1433   7.953  7.2e-13 ***
## as.factor(Sex)1      -8.6231    17.5263  -0.492  0.62353
## HT2:as.factor(Sex)1   0.1046     0.1994   0.525  0.60070
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.836 on 132 degrees of freedom
## Multiple R-squared:  0.5237, Adjusted R-squared:  0.5129
## F-statistic: 48.38 on 3 and 132 DF,  p-value: < 2.2e-16
```

```
AIC(lm1.1, lm1.3)
```

```
##       df      AIC
## lm1.1  3 754.5998
## lm1.3  5 757.5803
```

```
BIC(lm1.1, lm1.3)
```

```
##       df      BIC
## lm1.1  3 763.3378
## lm1.3  5 772.1436
```

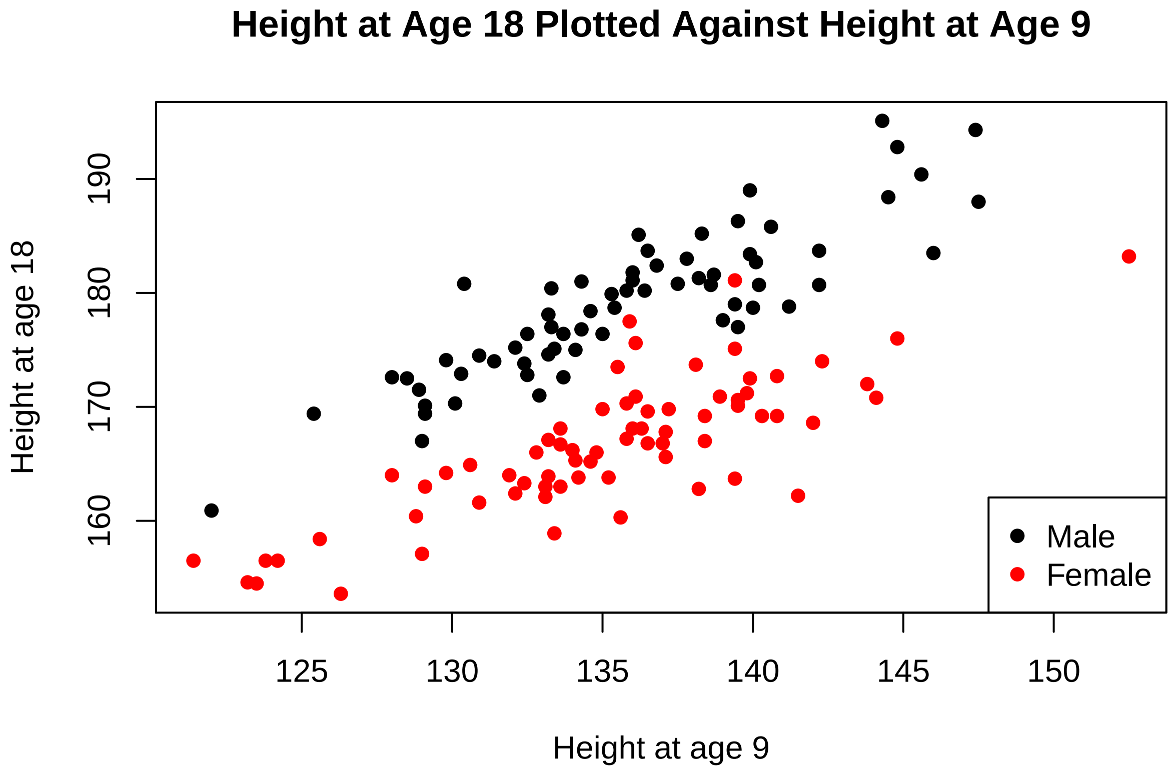
The adjusted R-squared of the model that allows for both the separate slope and separate intercepts for boys and for girls is 0.5129 while that of the original model is 0.5166. Also, both AIC and BIC values are lower for the base model. This model is not better than the simple linear regression fit above.

2.

(a)

```
##Create a scatter plot of heights at age 18 on heights at age 9.
plot(data$HT9, data$HT18,
     col = data$Sex + 1, pch = 16,
     xlab = "Height at age 9", ylab = "Height at age 18",
     main = "Height at Age 18 Plotted Against Height at Age 9")

legend(x = "bottomright",
       legend = c("Male", "Female"),
       col = c(1, 2), pch=16)
```



Height at Age 18 Plotted Against Height at Age 9

Yes, it seems that there is a different pattern for boys than for girls. They seem to have different slopes, and their intercept may also be different.

(b)

```
##Fit a simple linear regression of heights at age 18 on heights at age 9.
lm2.1 <- lm(HT18 ~ HT9, data = data)
summary(lm2.1)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5956  -5.8362   0.2947   5.9733  13.4930
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.3416    14.4329   2.241   0.0267 *
## HT9           1.0350     0.1064   9.724   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.797 on 134 degrees of freedom
## Multiple R-squared:  0.4137, Adjusted R-squared:  0.4094
## F-statistic: 94.56 on 1 and 134 DF,  p-value: < 2.2e-16
```

The intercept is 32.3416 and the estimated regression coefficient of HT9 is 1.035, and both are significant at

$\alpha = 0.05$ significance level. However, the intercept cannot be interpreted here, as it would require that an individual to be 0cm tall at the age of 9, which is impossible. The estimated coefficient suggests that if we were to compare two groups of people who were equivalent on all other variables and one unit (1cm) apart in the height at age 9, then the expected value of height at age 18 of the group that was 1cm taller at age 9 would be bigger by 1.035 compared to the group that was shorter.

(c)

```
##Generae a model that allows for separate intercepts for boys and girls
lm2.2 <- lm(HT18 ~ HT9 + as.factor(Sex), data = data)
summary(lm2.2)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + as.factor(Sex), data = data)
##
## 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
```

```
AIC(lm2.1, lm2.2)
```

```
##       df      AIC
## lm2.1  3 911.2241
## lm2.2  4 726.3621
```

```
BIC(lm2.1, lm2.2)
```

```
##       df      BIC
## lm2.1  3 919.9621
## lm2.2  4 738.0127
```

The adjusted R-squared of the model that allows for separate intercepts for boys and girls is 0.8494 while that of the base model is 0.4094. Also, both AIC and BIC values are lower for the model that allows for separate intercepts for boys and girls than for the base model. Therefore, this model is better than the simple linear regression fit above.

(d)

```
##Generate a model that allows for both
##the separate slope and separate intercepts for boys and for girls
```

```r
lm2.3 <- lm(HT18 ~ HT9*as.factor(Sex), data = data)
summary(lm2.3)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 * as.factor(Sex), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9224 -1.9453 -0.0081  1.7906 10.8136
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           35.07880   10.67406   3.286   0.0013 **
## HT9                    1.05895    0.07849  13.492   <2e-16 ***
## as.factor(Sex)1       13.32748   14.54695   0.916   0.3612
## HT9:as.factor(Sex)1 -0.18463    0.10725  -1.722   0.0875 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.407 on 132 degrees of freedom
## Multiple R-squared:  0.8549, Adjusted R-squared:  0.8516
## F-statistic: 259.2 on 3 and 132 DF,  p-value: < 2.2e-16
```

```r
AIC(lm2.1, lm2.3)
```

```
##       df      AIC
## lm2.1  3 911.2241
## lm2.3  5 725.3423
```

```r
BIC(lm2.1, lm2.3)
```

```
##       df      BIC
## lm2.1  3 919.9621
## lm2.3  5 739.9055
```

The adjusted R-squared of the model that allows for both the separate slope and separate intercepts for boys and for girls is 0.8516 while that of the base model is 0.4094. Also, both AIC and BIC values are lower for the model that allows for separate intercepts for boys and girls than for the base model. Therefore, this model is better than the simple linear regression fit above.

(e)

Judging from the scatterplot and the regression output, it seems that the last model with the interaction term is the best among the three models.

```r
summary(lm2.3)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 * as.factor(Sex), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9224 -1.9453 -0.0081  1.7906 10.8136
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         35.07880   10.67406   3.286   0.0013 **
## HT9                  1.05895    0.07849  13.492   <2e-16 ***
## as.factor(Sex)1     13.32748   14.54695   0.916   0.3612
## HT9:as.factor(Sex)1 -0.18463    0.10725  -1.722   0.0875 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.407 on 132 degrees of freedom
## Multiple R-squared:  0.8549, Adjusted R-squared:  0.8516
## F-statistic: 259.2 on 3 and 132 DF,  p-value: < 2.2e-16
```

The intercept for male is 35.079 and the intercept for female is $35.079 + 13.327 = 48.406$. The slope for men is 1.059 and the slope for female is $1.059 - 0.185 = 0.874$. Again, the intercepts cannot be interpreted here, because they would require that an individual is 0cm tall at the age of 9, which is impossible.\ The slope for male suggests that when comparing two groups of boys that are one unit (1cm) apart in the height at age 9, all other covariates held fixed, the expected value of height at age 18 of the group that was 1cm taller at age 9 would be bigger by 1.059cm compared to the group that was shorter. The slope for female on the other hand, suggests that when comparing two groups of girls that are one unit (1cm) apart in the height at age 9, all other covariates held fixed, the expected value of height at age 18 of the group that was 1cm taller at age 9 would be bigger by 0.874cm compared to the group that was shorter. The F-statistic is statistically significant at $\alpha = 0.05$ level.

3.

(a)

```
##Create a new dataset that includes only the boys in the sample
data2 <- data[data$Sex == 0,]
```

```
##Fit two linear regression models, M1 and M2
m1.3 <- lm(WT18 ~ WT9, data = data2)
m2.3 <- lm(WT18 ~ WT9 + LG9, data = data2)
summary(m1.3)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.024  -3.607   0.024   2.858  29.592
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.1124     4.9686   7.469 2.78e-10 ***
## WT9           1.0481     0.1542   6.796 4.23e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 64 degrees of freedom
## Multiple R-squared:  0.4192, Adjusted R-squared:  0.4101
## F-statistic: 46.19 on 1 and 64 DF,  p-value: 4.235e-09
```

```
summary(m2.3)
```

```
## 
## Call:
## lm(formula = WT18 ~ WT9 + LG9, data = data2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20.5617  -3.2447  -0.3437   3.1478  29.1951 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.9585    18.2766   1.092    0.279
## WT9           0.6299     0.4557   1.382    0.172
## LG9           1.1046     1.1326   0.975    0.333
## 
## Residual standard error: 7.667 on 63 degrees of freedom
## Multiple R-squared:  0.4278, Adjusted R-squared:  0.4096 
## F-statistic: 23.55 on 2 and 63 DF,  p-value: 2.304e-08
```

```
#Correlation coefficient between WT9 and LG9
cor(data2$WT9, data2$LG9)
```

```
## [1] 0.9409453
```

WT9 is significant in the first model (M1.3) but it is not in the second model (M2.3). This is because WT9 and LG9 are highly correlated. When two covariates are highly correlated, the marginal contribution of one covariate in reducing the residual sum of squares varies depending on the other variable. Thus, a variable can be made statistically insignificant by introducing another covariate that is highly correlated with it. It also makes sense that the two covariates are highly correlated because the thicker one's limbs, chances are that he/she weighs more than he/she would have with thin limbs. We can see that the corrleation coefficient of WT9 and LG9 is 0.941, suggesting a high degree of correlation.

(b)

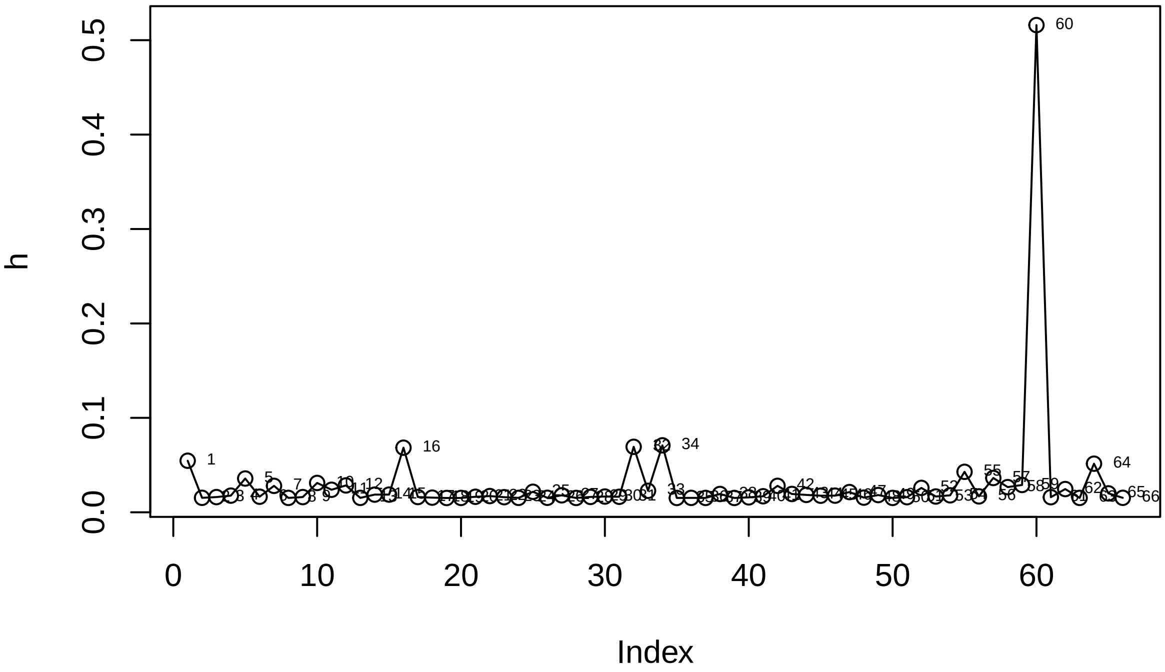
- The hat matrix is the projection matrix that expresses the values of y in terms of the linear combinations of the column vectors of the design matrix X. X contains a column of 1s to estimate the intercept, and the values of explanatory variables we are regressing on which are WT9 and LG9 in this case. In regression, we estimate each observations in y with linear combinations of the intercept and explanatory variables which are the column vectors of the design matrix, but the actual values of y and the estimated values using the linear combinations are not always the same. In other words, the observations in y usually do not lie on the column space of X and there will be difference between y and the estimated y, $\hat{y}$. The hat matrix, H, projects y onto the column space of X which results in $\hat{y}$. Thus, $Hy = \hat{y}$, hence the name 'hat matrix' because H puts the hat on y and results in $\hat{y}$.

```
##Calculate H
#generate the design matrix
X <- cbind(1, data2$WT9)

#calculate the hat matrix
H <- X %*% solve(t(X) %*% X) %*% t(X)
h <- diag(H)

#plot the leverage of each point
```

```
plot(h, type = "o")
text(h, labels=1:length(h), pos=4, cex=0.5)
```



```
match(max(h), h) #the 60th data point has the highest leverage
```

```
## [1] 60
```

- From the diagonal values of the hat matrix, which determine the leverage that each point has in the fit of the regression model, we can see that the 60th data point has much higher leverage than any of the other points.

```
##Fit a simple linear regression model without using the high-leverage point
m3.3 <- lm(WT18 ~ WT9, data = data2[- 60,])
summary(m1.3) #model using all of the boys
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.024  -3.607   0.024   2.858  29.592
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.1124     4.9686   7.469 2.78e-10 ***
## WT9           1.0481     0.1542   6.796 4.23e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 64 degrees of freedom
## Multiple R-squared:  0.4192, Adjusted R-squared:  0.4101
## F-statistic: 46.19 on 1 and 64 DF,  p-value: 4.235e-09
```

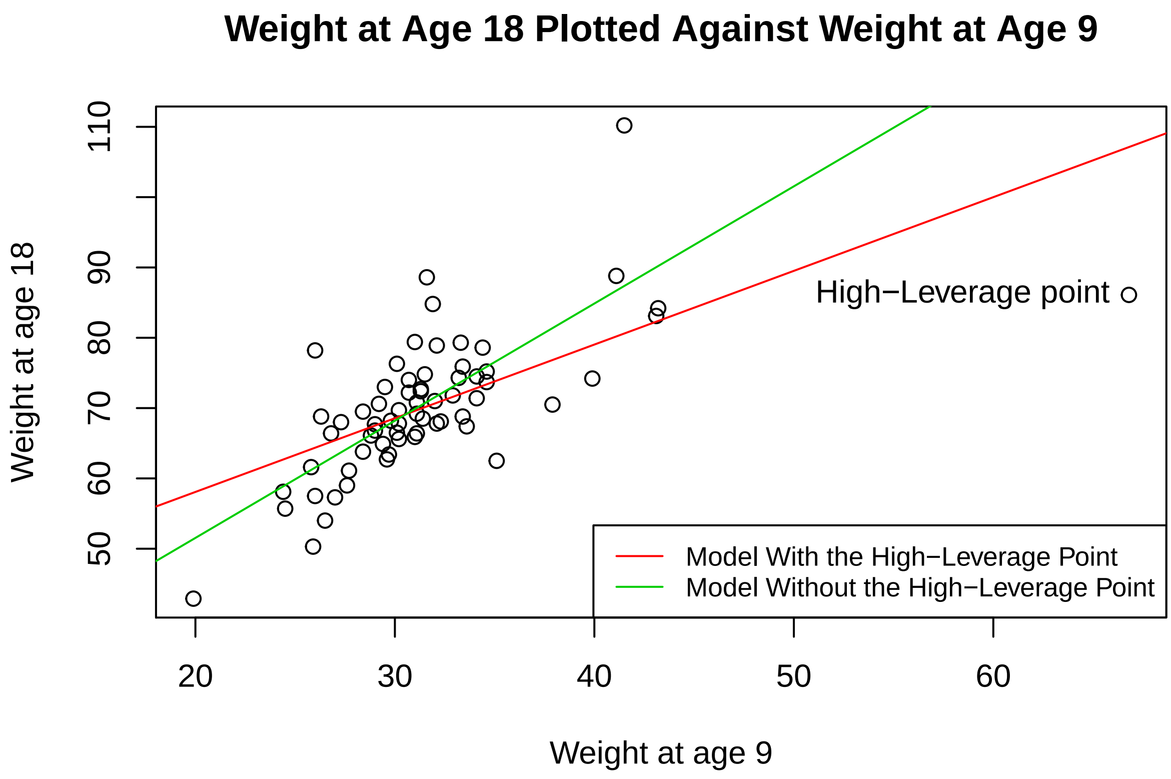
```
summary(m3.3) #model without the high-leverage point
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = data2[-60, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2037  -3.9370  -0.6703   3.0630  22.8295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.2029     6.0556   3.006   0.0038 **
## WT9           1.6667     0.1929   8.639 2.73e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.721 on 63 degrees of freedom
## Multiple R-squared:  0.5423, Adjusted R-squared:  0.535
## F-statistic: 74.64 on 1 and 63 DF,  p-value: 2.734e-12
```

- The intercept is 37.112 for the model using all of the boys and is 18.203 for the model without the high-leverage point, and both are statistically significant at $\alpha = 0.05$ level. The coefficients for weight at age 9 are also both statistically significant at $\alpha = 0.05$ significance level in both models. However, it is higher for the model that removed the high-leverage point at 1.667 compared to that of the model with all boys at 1.048.

```
##Create a scatter plot of weight at age 18 on weight at age 9 with regression lines

HLpoint <- data2[60,]
plot(data2$WT9, data2$WT18,
     xlab = "Weight at age 9", ylab = "Weight at age 18",
     main = "Weight at Age 18 Plotted Against Weight at Age 9")
text(HLpoint$WT9, HLpoint$WT18, labels = "High-Leverage point", pos = 2)
abline(m1.3, col = 2)
abline(m3.3, col = 3)
legend(x = "bottomright",
       legend = c("Model With the High-Leverage Point",
                  "Model Without the High-Leverage Point"),
       col = c(2, 3), lty=1,
       cex = 0.8)
```



- As we can see above, the high-leverage point pulls down the fitted regression line in red, causing it to be a less effective model to explain our data. Also, it is hard to believe that a single child recorded over 60kg at the age of 9, when other children recorded at most around 45kg at the same age. This may be a measurement error or mistake in recording the data. Therefore, I believe it is better to remove the outlier.

```
summary(m3.3)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = data2[-60, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2037  -3.9370  -0.6703   3.0630  22.8295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.2029     6.0556   3.006   0.0038 **
## WT9           1.6667     0.1929   8.639 2.73e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.721 on 63 degrees of freedom
## Multiple R-squared:  0.5423, Adjusted R-squared:  0.535
## F-statistic: 74.64 on 1 and 63 DF,  p-value: 2.734e-12
```

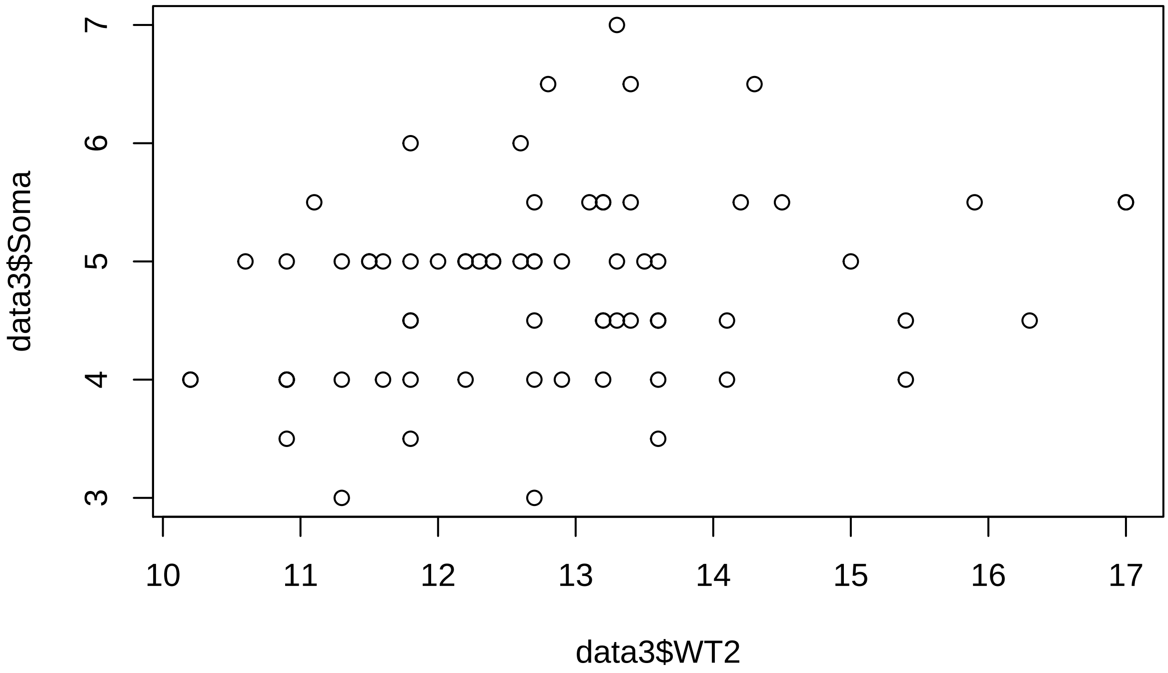
- From the model that removed the high-leverage point, both the intercept and coefficient for WT9 are statistically significant at $\alpha = 0.05$ significance level. Here, we cannot interpret the intercept, as it is the expected weight at age 18 when a child is 0kg at the age of 9, which is impossible. The estimated coefficient for WT9 is suggests that if we were to compare two groups of people who were equivalent on all other variables and one unit (1kg) apart in the weight at age 9, then the expected value of weight at age 18 of the group that was 1kg heavier would be bigger by 1.6667 compared to the group that had lighter weight at age 9.

4.

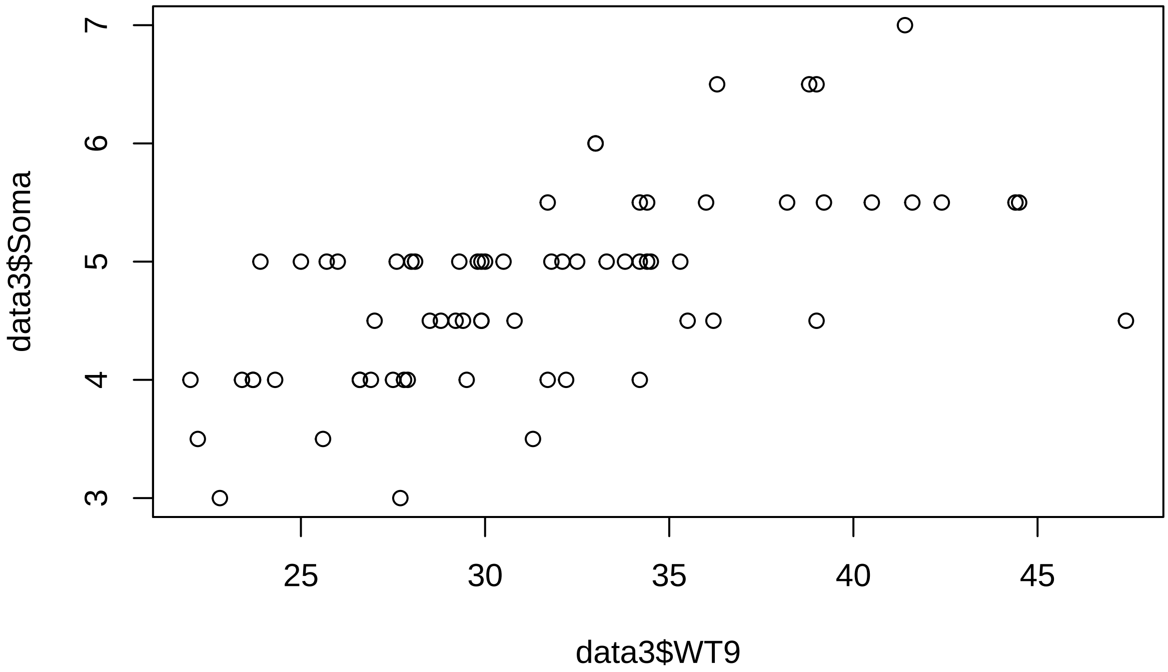
(a)

```
##Create a new dataset that includes only the girls
data3 <- data[data$Sex == 1,]
```

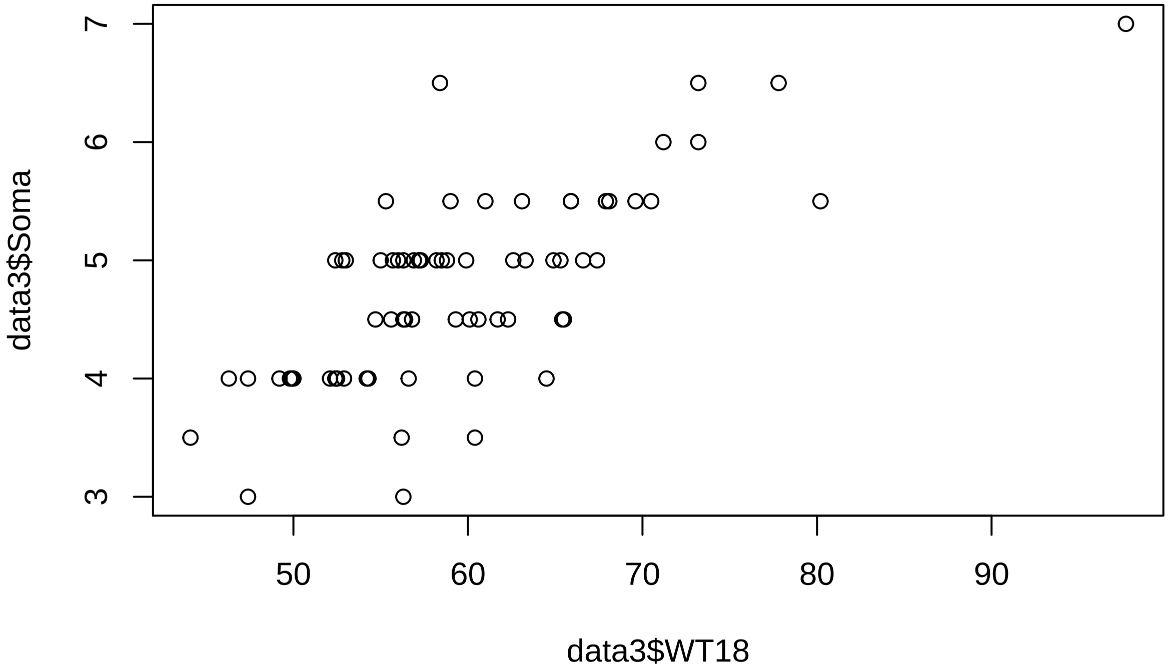
```
##Plot somatotype against weight at each of the three time points
plot(data3$WT2, data3$Soma)
```



```
plot(data3$WT9, data3$Soma)
```



```r
plot(data3$WT18, data3$Soma)
```



At the age of 2, there does not seem to be strong linear relationship beween weight and somatotype. However, at the age of 9 there seems to be a positive relationship between weight and somatotype and at age 18 the relationship seems to be stronger. The relationship maybe linear.

(b)

```r
##Create new variables
data3$DW9 <- data3$WT9 - data3$WT2
```

```
data3$DW18 <- data3$WT18 - data3$WT9
data3$AVE <- (data3$WT2 + data3$WT9 + data3$WT18)/3
data3$LIN <- data3$WT18 - data3$WT2
data3$QUAD = data3$WT2 - 2 * data3$WT9 + data3$WT18
```

(c)

```
##Fit three models
m1.4 <- lm(Soma ~ WT2 + WT9 + WT18, data = data3)
m2.4 <- lm(Soma ~ WT2 + DW9 + DW18, data = data3)
m3.4 <- lm(Soma ~ AVE + LIN + QUAD, data = data3)
summary(m1.4)
```

```
##
## Call:
## lm(formula = Soma ~ WT2 + WT9 + WT18, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40296 -0.26081 -0.03177  0.38015  1.44088
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59210    0.67425   2.361  0.02117 *
## WT2         -0.11564    0.06169  -1.874  0.06530 .
## WT9          0.05625    0.02011   2.797  0.00675 **
## WT18         0.04834    0.01060   4.559 2.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.543 on 66 degrees of freedom
## Multiple R-squared:  0.5658, Adjusted R-squared:  0.5461
## F-statistic: 28.67 on 3 and 66 DF,  p-value: 5.497e-12
```

```
summary(m2.4)
```

```
##
## Call:
## lm(formula = Soma ~ WT2 + DW9 + DW18, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40296 -0.26081 -0.03177  0.38015  1.44088
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59210    0.67425   2.361   0.0212 *
## WT2         -0.01106    0.05194  -0.213   0.8321
## DW9          0.10459    0.01570   6.659 6.50e-09 ***
## DW18         0.04834    0.01060   4.559 2.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.543 on 66 degrees of freedom
## Multiple R-squared:  0.5658, Adjusted R-squared:  0.5461
## F-statistic: 28.67 on 3 and 66 DF,  p-value: 5.497e-12
```

```
summary(m3.4)
```

```
##
## Call:
## lm(formula = Soma ~ AVE + LIN + QUAD, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40296 -0.26081 -0.03177  0.38015  1.44088
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59210    0.67425   2.361   0.0212 *
## AVE         -0.01106    0.05194  -0.213   0.8321
## LIN          0.08199    0.03041   2.696   0.0089 **
## QUAD        -0.02997    0.01620  -1.850   0.0688 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.543 on 66 degrees of freedom
## Multiple R-squared:  0.5658, Adjusted R-squared:  0.5461
## F-statistic: 28.67 on 3 and 66 DF,  p-value: 5.497e-12
```

- The three models are similar in that they all have the same estimates and standard error of the intercept, the residual standard error, the R^2 and adjusted-R^2, and the F-statistic. This is because the three models all have WT2, WT9, WT18 either explicitly or embedded in the newly created variables which are linear combinations of parts or all of the three weight variables, and the design matrices for all four matrices span the same space or would reduce down to the same thing if the collinearity were removed.

```
summary(m1.4)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  1.59210082 0.67424702  2.361302 2.117198e-02
## WT2         -0.11564269 0.06169479 -1.874432 6.529750e-02
## WT9          0.05624766 0.02011137  2.796809 6.754965e-03
## WT18         0.04833850 0.01060206  4.559350 2.281963e-05
```

```
summary(m2.4)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  1.59210082 0.67424702  2.3613020 2.117198e-02
## WT2         -0.01105652 0.05194166 -0.2128642 8.320890e-01
## DW9          0.10458616 0.01570483  6.6594896 6.501853e-09
## DW18         0.04833850 0.01060206  4.5593499 2.281963e-05
```

- The coefficient for DW18 in model2 equal the coefficient for WT18 in model 1, but the coefficient for DW9 in model 2 does not equal the coefficient for WT9 in model 1, because WT18 is only accounted for in DW18 while WT9 is accounted for in both DW9 and DW18.

- When all other covariates are held fixed, with one unit increase in DW18 in the second model we expect to observe 0.048 higher somatotype value, which is the same amount of difference when one unit incrase in WT18 in the first model, because as we hold DW9 fixed, we also hold WT9 fixed which is also in DW18. Therefore the change in DW18 can only affect somatotype by the amount WT18 does. However, for DW9, even if we hold WT2 and DW18, as DW9 increase by one unit it will affect DW18 to change. By increasing

DW9, WT2 is held fixed and it will increase the value of WT9 in DW9, which will also affect the value of DW18 via WT9 in it. Therefore, the coefficient of DW9 must account for this. We can also derive this algebraically.

$$
\begin{aligned}
E(Soma|X) &= \beta_0 + \beta_1 WT2 + \beta_2 DW9 + \beta_3 DW18 && (1)\\
&= \beta_0 + \beta_1 WT2 + \beta_2 (WT9 - WT2) + \beta_3 (WT18 - WT9) && (2)\\
&= \beta_0 + (\beta_1 - \beta_2) WT2 + (\beta_2 - \beta_3) WT9 + \beta_3 WT18 && (3)\\
& && (4)
\end{aligned}
$$

- As shown below, M1 and M3 are equivalent. The coefficients in M3 can be algebraically manipulated to get the coefficients in M1, and the opposite is also true if the process is done backwards.

$$
\begin{aligned}
E(Soma|X) &= \beta_0 + \beta_1 AVE + \beta_2 LIN + \beta_3 QUAD && (5)\\
&= \beta_0 + \beta_1(\frac{1}{3}(WT2 + WT9 + WT18)) + \beta_2(WT18 - WT2) + \beta_3(WT2 - 2\dot{W}T9 + WT18) && \\
& && (6)\\
&= \beta_0 + (\frac{1}{3}\beta_1 - \beta_2 + \beta_3) WT2 + (\frac{1}{3}\beta_1 - 2\beta_3) WT9 + (\frac{1}{3}\beta_1 + \beta_2 + \beta_3) WT18 && (7)\\
&= \alpha_0 + \alpha_1 WT2 + \alpha_2 WT9 + \alpha_3 WT18 && (8)
\end{aligned}
$$

(d)

```
m4.4 <- lm(Soma ~ WT2 + WT9 + WT18 + DW9, data = data3)
summary(m4.4)
```

```
##
## Call:
## lm(formula = Soma ~ WT2 + WT9 + WT18 + DW9, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40296 -0.26081 -0.03177  0.38015  1.44088
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59210    0.67425   2.361  0.02117 *
## WT2         -0.11564    0.06169  -1.874  0.06530 .
## WT9          0.05625    0.02011   2.797  0.00675 **
## WT18         0.04834    0.01060   4.559 2.28e-05 ***
## DW9               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.543 on 66 degrees of freedom
## Multiple R-squared:  0.5658, Adjusted R-squared:  0.5461
## F-statistic: 28.67 on 3 and 66 DF,  p-value: 5.497e-12
```

DW9 is not estimated. There is perfect multicollinearity between DW9, WT2, and WT9 as DW9 is a linear combination of WT2 and WT9. Since WT2 and WT9 are already included in the model, adding DW9 does not add any new information, and is not estimated by R.