Regression Project (Part 1)

Ramya Dhatri Vunikili (rdv253)

April 01, 2017

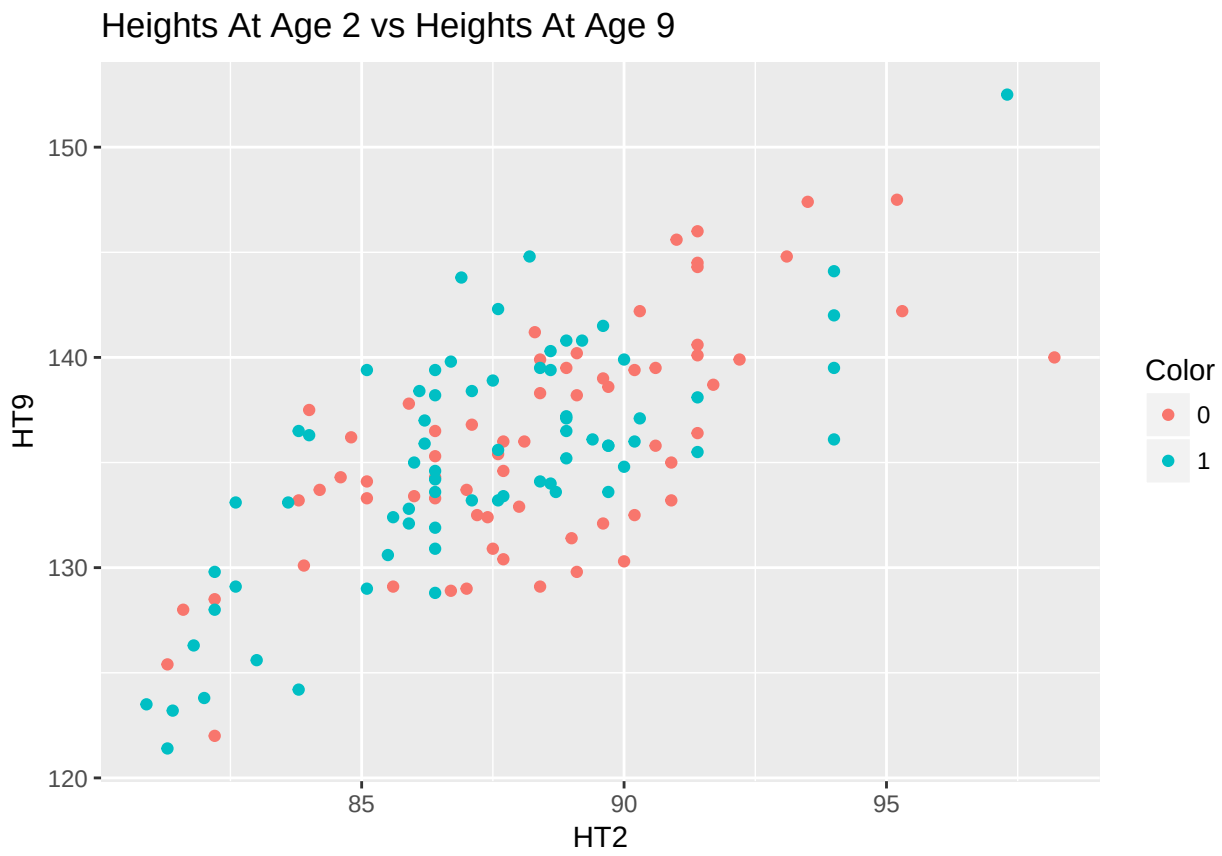
Answer 1a

The scatter plot has been color coded to represent boys (0) in red and girls (1) in blue colors. As we can see in the plot, there appears to be no different pattern for boys and girls.

```
berkeley_data <- read.csv("BGS.csv")
n <- nrow(berkeley_data)

################################### 1a)  Scatter Plot #######################################

##Color coding for boys (red) and girls (green)
berkeley_data$Color <- as.factor(berkeley_data$Sex)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
ggplot(berkeley_data,aes(x=HT2, y=HT9))+
  geom_point(aes(colour=Color))+
  ggtitle("Heights At Age 2 vs Heights At Age 9")
```



Answer 1b

```
########################## 1b) Linear Regression Of HT9 On HT2  ############################
lm.ht9v2=lm(HT9 ~ HT2, data=berkeley_data)

## Beta Hat Matrix Calculation
## Creating X and Y matrices for the regression lm(HT9 ~ HT2)
X <- as.matrix(cbind(1,berkeley_data$HT2))
Y <- as.matrix(berkeley_data$HT9)

## Calculating matrix of estimated coefficients:
beta.matrix <- round(solve(t(X)%*%X)%*%t(X)%*%Y, digits=5)

## Labeling and organizing results into a data frame
beta.hat <- as.data.frame(cbind(c("Intercept","HT2"),beta.matrix))
names(beta.hat) <- c("Coefficient","Value")

## Calculating vector of residuals
res <- as.matrix(berkeley_data$HT9-beta.matrix[1]-beta.matrix[2]*berkeley_data$HT2)
k <- ncol(X)

## Calculating Variance-Covariance Matrix
CV = 1/(n-k) * as.numeric(t(res)%*%res) * solve(t(X)%*%X)

## Standard errors of the estimated coefficient HT2
StdErr = sqrt(diag(CV))

## Calculating p-value for a t-test of coefficient significance
P.Value = rbind(2*pt(abs(beta.matrix[1]/StdErr[1]), df=n-k,lower.tail= FALSE),
                2*pt(abs(beta.matrix[2]/StdErr[2]), df=n-k,lower.tail= FALSE))


## Concatenating into a single data.frame
beta.hat = cbind(beta.hat,StdErr,P.Value)
beta.hat
```

```
##   Coefficient    Value     StdErr      P.Value
## 1   Intercept 31.92705 8.59959532 2.998481e-04
## 2         HT2  1.17963 0.09787906 4.085436e-23
```

```
## Ho: b1 = 0; Ha: b1 != 0;
b1.pval = P.Value[2]
b1.pval
```

```
## [1] 4.085436e-23
```

```
## At alpha = 0.05, b1.pval < alpha and hence we reject the null hypothesis b1 = 0.


## T-Statistic for b1
b1.tstat <- (beta.matrix[2]/StdErr[2])
b1.tstat
```

```
## [1] 12.05191
```

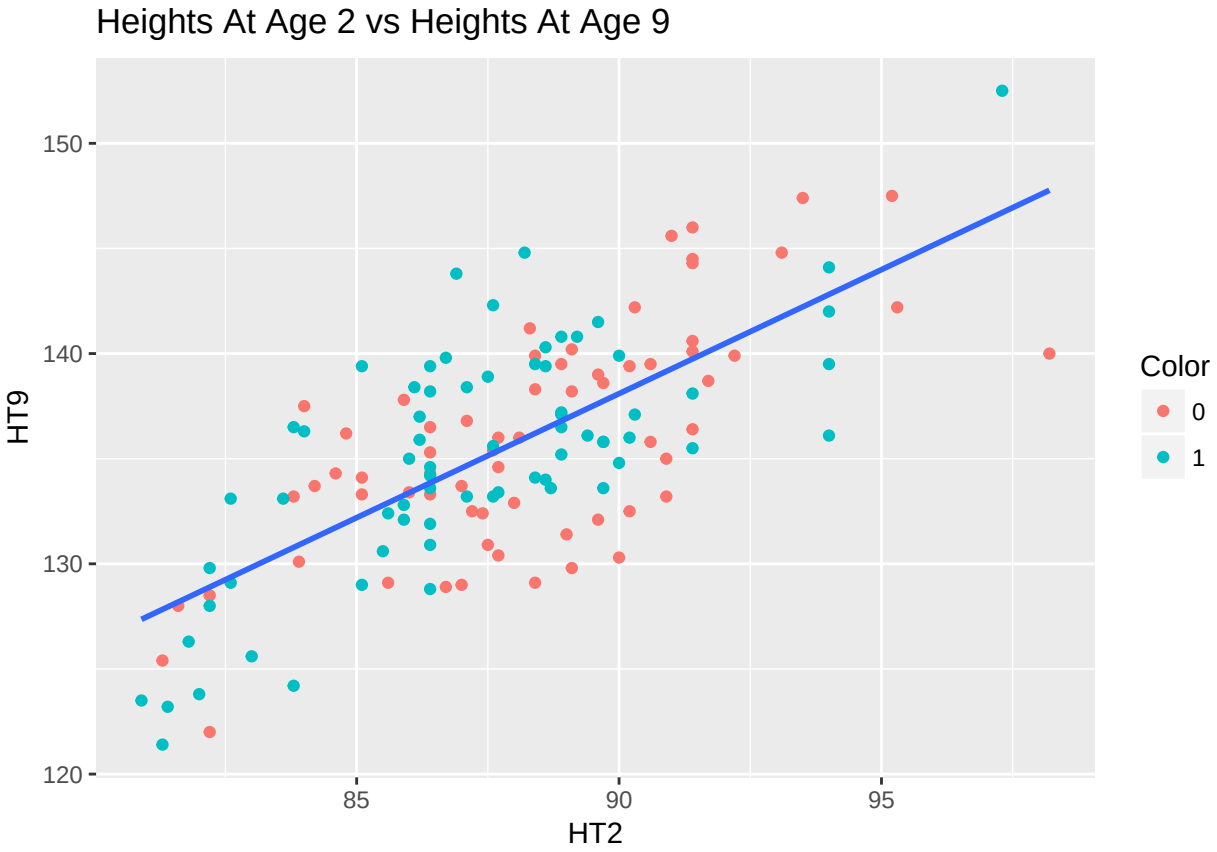
```
## F-statistic: 145.2 on 1 and 134 DF (from summary of the regression)
summary(lm.ht9v2)
```

```
##
## Call:
## lm(formula = HT9 ~ HT2, data = berkeley_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
```

```
fstat <- 145.2
sqrt(fstat) ## is 12.0499
```

```
## [1] 12.0499
```

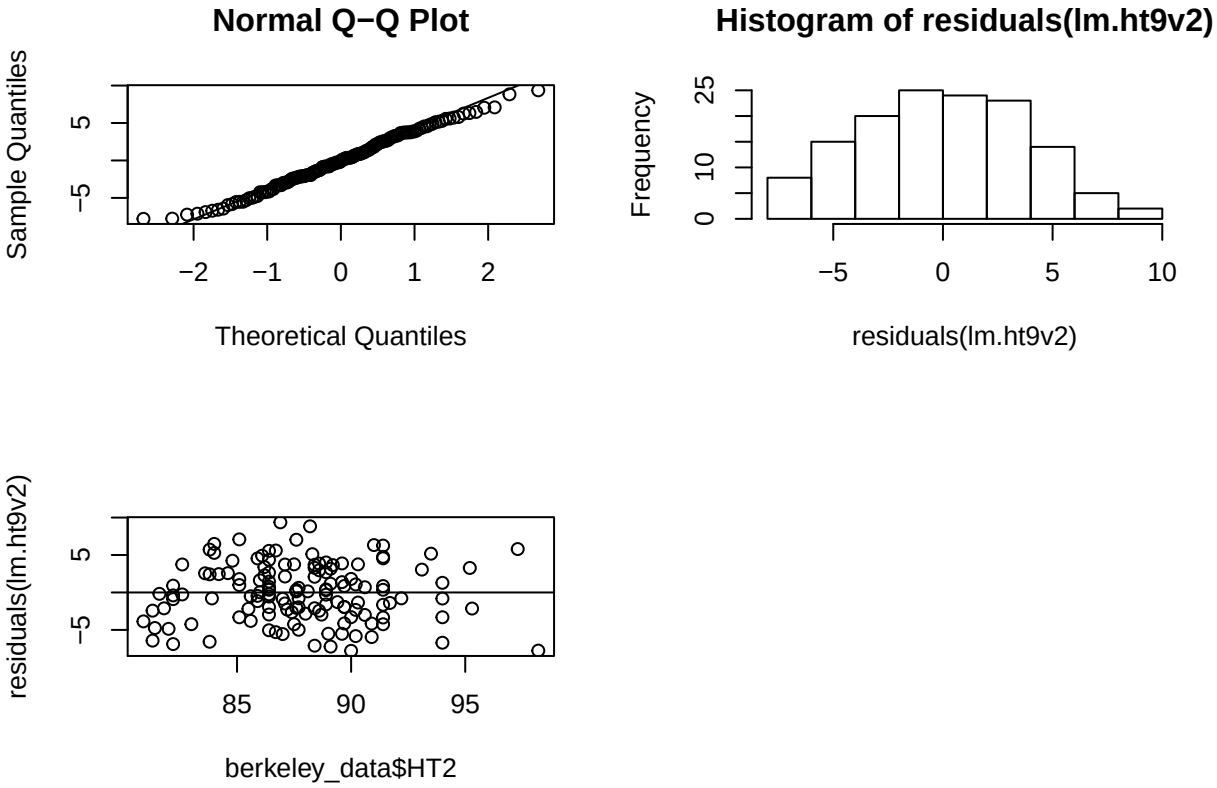
```
## Calculated value of b1.tstat = 12.051 which is approx. equal to sqrt(fstat)
ggplot(berkeley_data,aes(x=HT2, y=HT9))+
  geom_point(aes(colour=Color))+
  geom_smooth(method=lm, se=FALSE)+
  ggtitle("Heights At Age 2 vs Heights At Age 9")
```



At alpha = 0.05, b1.pval (= 4.085436e-23) < alpha and hence we reject the null hypothesis that $\beta_1 = 0$

Calculated value of b1.tstat = 12.051 which is approximately equal to sqrt(fstat) i.e., 12.0499

Normality & Homoscedasticity



Looking at the qqplot and histogram of residuals it can be concluded that the normality assumption has been satisfied for the linear regression. Also, as the residual plot shows no funnel shaped pattern it can be said that the regression satisfies the homoscedaticity assumption too.

Answer 1c (Seperate Slopes)

Table 1: Fitting linear model: HT9 ~ HT2 + Sex

	Estimate	Std. Error	t value	Pr(>|t|)
HT2	1.19373	0.0993754	12.0123	5.73595e-23
Sex	0.565616	0.66571	0.849643	0.397051
(Intercept)	30.3984	8.79454	3.45651	0.000735112

Table 2: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
133	1946.47	NA	NA	NA	NA
134	1957.04	-1	-10.565	0.721893	0.397051

As th p value for ANOVA between original model and the model with seperate slopes is 0.397051 ($\not< 0.05$), we do not find any significant difference between the two models.

Answer 1d (Seperate Slopes & Intercepts)

Table 3: Fitting linear model: HT9 ~ HT2 + Sex + Sex * HT2

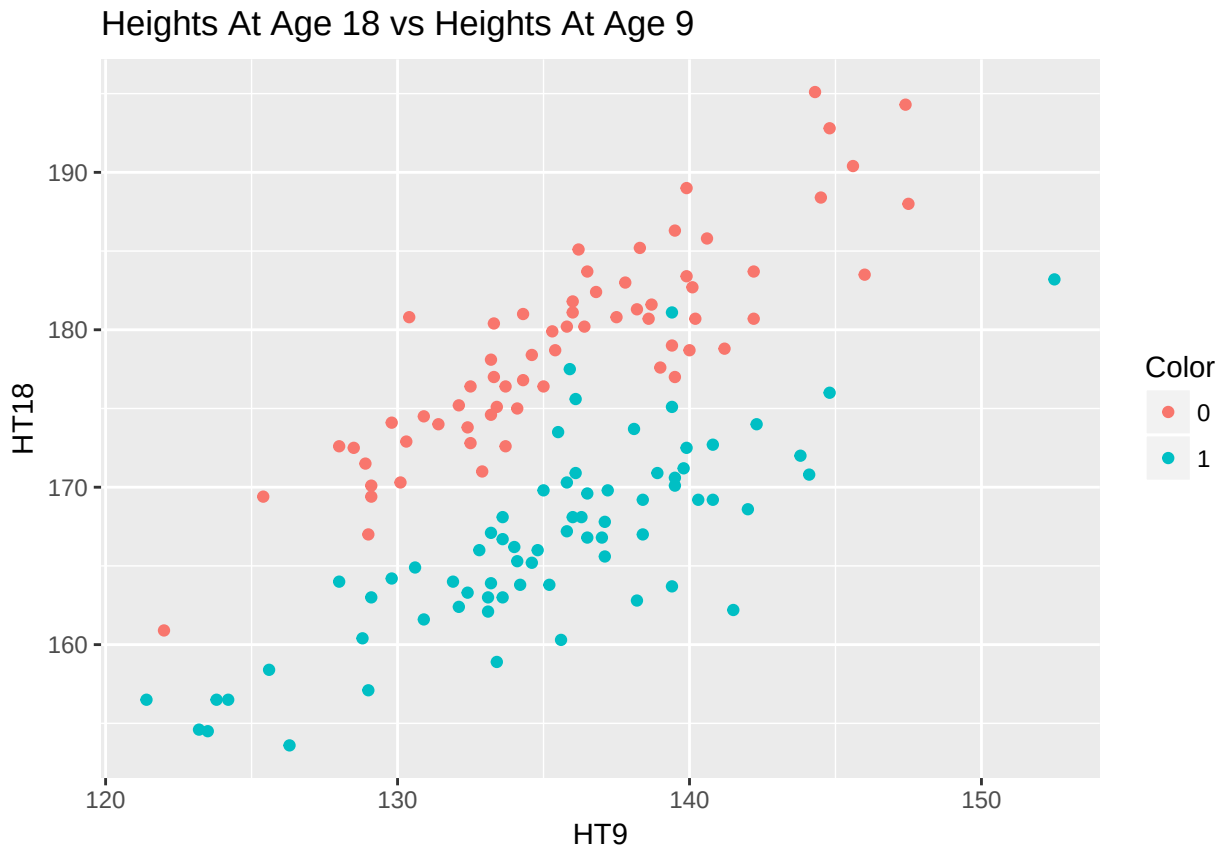
	Estimate	Std. Error	t value	Pr(>|t|)
HT2	1.13969	0.143301	7.95317	7.2028e-13
Sex	-8.62313	17.5263	-0.49201	0.623529
HT2:Sex	0.104619	0.199403	0.524663	0.600697
(Intercept)	35.1732	12.6724	2.77556	0.00631144

Table 4: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
132	1942.42	NA	NA	NA	NA
134	1957.04	-2	-14.6157	0.496615	0.60972

As th p value for ANOVA between original model and the model with seperate slopes and intercepts is 0.60972 ($\not< 0.05$), we do not find any significant difference between the two models.

Answer 2a



Yes,there appears to be a different pattern of heights for boys and girls. At the age 18, boys seem to be taller than the girls at the same age. ————————————-

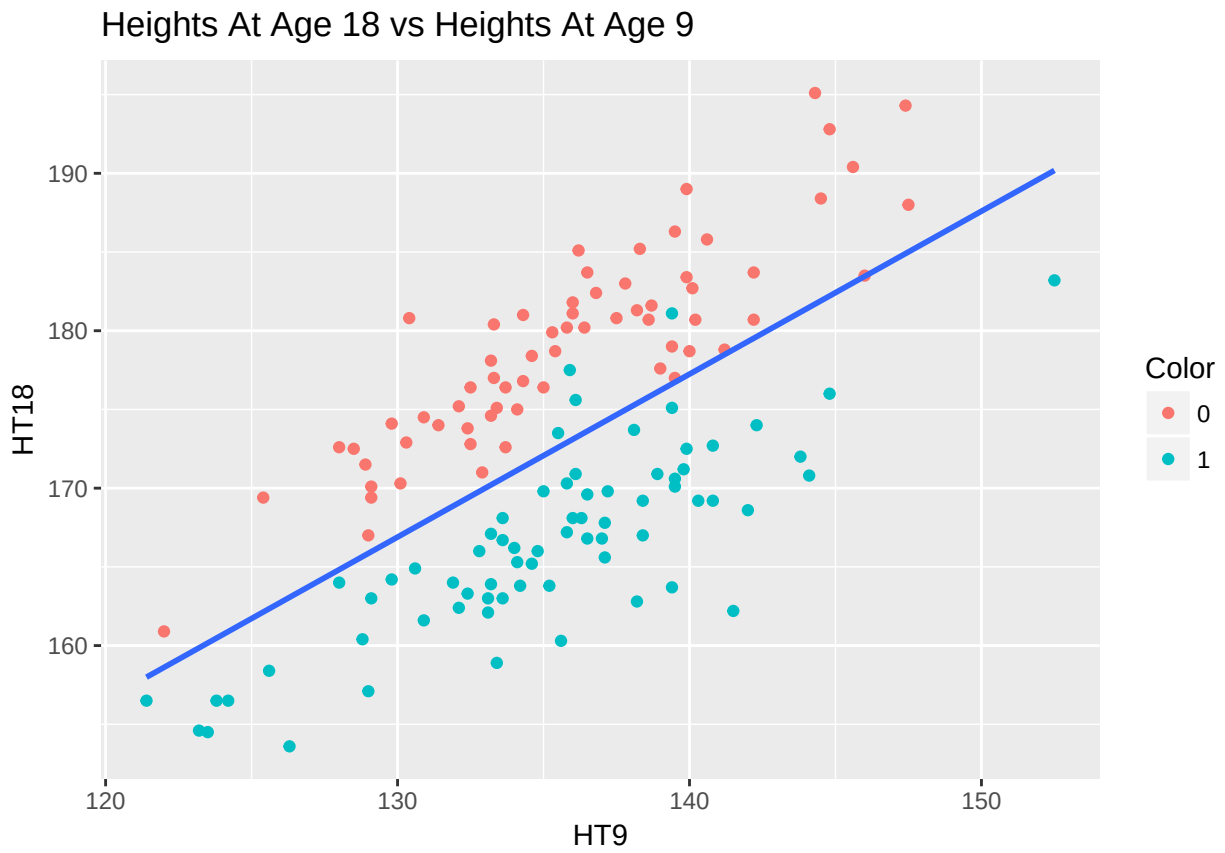
Answer 2b

```
############################## 2b) Linear Regression Of HT18 On HT9 ##########################
lm.ht18v9 = lm(HT18 ~ HT9, data = berkeley_data)


## Beta Hat Matrix Calculation
## Creating X and Y matrices for the regression lm(HT18 ~ HT9)
X <- as.matrix(cbind(1,berkeley_data$HT9))
Y <- as.matrix(berkeley_data$HT18)

## Calculating matrix of estimated coefficients:
beta.matrix <- round(solve(t(X)%*%X)%*%t(X)%*%Y, digits=5)

## Labeling and organizing results into a data frame
beta.hat <- as.data.frame(cbind(c("Intercept","HT9"),beta.matrix))
names(beta.hat) <- c("Coefficient","Value")

## Calculating vector of residuals
res <- as.matrix(berkeley_data$HT18-beta.matrix[1]-beta.matrix[2]*berkeley_data$HT9)
```

```r
k <- ncol(X)

## Calculating Variance-Covariance Matrix
CV = 1/(n-k) * as.numeric(t(res)%*%res) * solve(t(X)%*%X)

## Standard errors of the estimated coefficient HT9
StdErr = sqrt(diag(CV))

## Calculating p-value for a t-test of coefficient significance
P.Value = rbind(2*pt(abs(beta.matrix[1]/StdErr[1]), df=n-k,lower.tail= FALSE),
                2*pt(abs(beta.matrix[2]/StdErr[2]), df=n-k,lower.tail= FALSE))


## Concatenating into a single data.frame
beta.hat = cbind(beta.hat,StdErr,P.Value)
beta.hat
```

```
##   Coefficient    Value     StdErr      P.Value
## 1   Intercept 32.34156 14.4329266 2.668357e-02
## 2         HT9  1.03501  0.1064344 3.071258e-17
```

```r
ggplot(berkeley_data,aes(x=HT9, y=HT18))+
  geom_point(aes(colour=Color))+
  geom_smooth(method=lm, se=FALSE)+
  ggtitle("Heights At Age 18 vs Heights At Age 9")
```



Answer 2c (Seperate Slopes)

Table 5: Fitting linear model: HT18 ~ HT9 + Sex

	Estimate	Std. Error	t value	Pr(>|t|)
HT9	0.960056	0.0538796	17.8185	4.77802e-37
Sex	-11.6958	0.590359	-19.8114	1.65951e-41
(Intercept)	48.5173	7.33385	6.61553	8.26588e-10

Table 6: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
133	1566.9	NA	NA	NA	NA
134	6190.91	-1	-4624.02	392.492	1.65951e-41

As the p value < 0.05, we can conclcude that the two models lm.ht18v9 and T1 differ significantly and hence T1 is a better fit to the data.

Answer 2d (Seperate Intercepts and Slopes)

Table 7: Fitting linear model: HT18 ~ HT9 + Sex + HT2:Sex

	Estimate	Std. Error	t value	Pr(>|t|)
HT9	0.95498	0.0642709	14.8587	5.79794e-30
Sex	-13.5871	12.9533	-1.04893	0.296127
Sex:HT2	0.0216305	0.147995	0.146157	0.884021
(Intercept)	49.2071	8.74402	5.62752	1.04954e-07

Table 8: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
132	1566.64	NA	NA	NA	NA
134	6190.91	-2	-4624.27	194.813	4.09465e-40

As the p value < 0.05, we can conclcude that the two models lm.ht18v9 and T2 differ significantly and hence T2 is a better fit to the data.

Answer 2e

```
a3 <- anova(T2, T1)
library(pander)
panderOptions("digits", 6)
panderOptions('knitr.auto.asis', FALSE)
pander(a3)
```

Table 9: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
132	1566.64	NA	NA	NA	NA
133	1566.9	-1	-0.253533	0.0213618	0.884021

```
############################ Parameter Estimates For The Best Model ########################
pander(T1)
```

Table 10: Fitting linear model: HT18 ~ HT9 + Sex

	Estimate	Std. Error	t value	Pr(>|t|)
HT9	0.960056	0.0538796	17.8185	4.77802e-37
Sex	-11.6958	0.590359	-19.8114	1.65951e-41
(Intercept)	48.5173	7.33385	6.61553	8.26588e-10

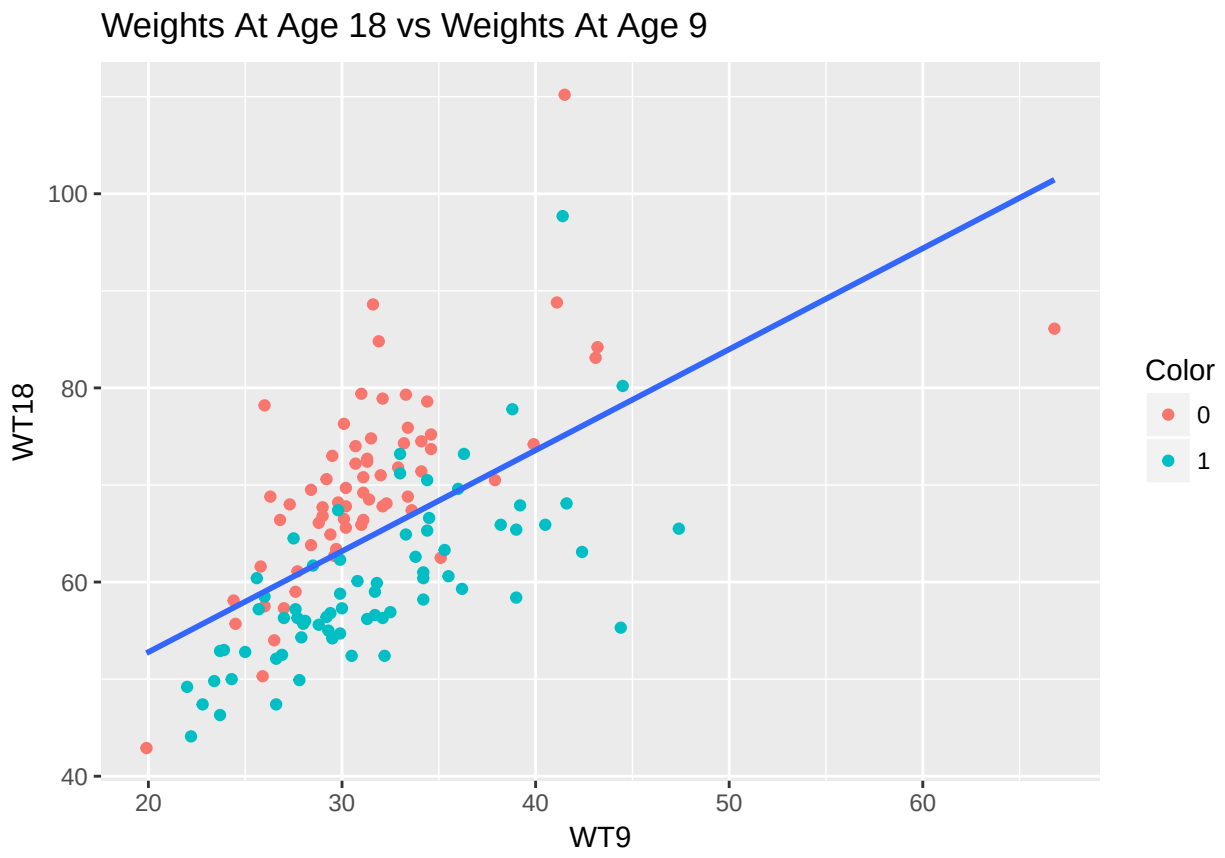
As th p value for anova(T2,T1) is 0.884 ($\not< 0.05$), we do not find any significant difference between the two models with different slopes and different slopes and intercepts. Moreover, the RSS values of the two models are nearly equal.

Answer 3a

```
################################### 3a) M1: WT18 ON WT9 ######################################
M1 <- lm(WT18 ~ WT9, data = berkeley_data)
summary(M1)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = berkeley_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.853  -6.378  -0.292   5.572  35.063
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.9847     4.0446   7.908 8.61e-13 ***
## WT9           1.0398     0.1257   8.273 1.15e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.717 on 134 degrees of freedom
## Multiple R-squared:  0.3381, Adjusted R-squared:  0.3332
## F-statistic: 68.45 on 1 and 134 DF,  p-value: 1.149e-13
```

```
ggplot(berkeley_data,aes(x=WT9, y=WT18))+
  geom_point(aes(colour=Color))+
  geom_smooth(method=lm, se=FALSE)+
  ggtitle("Weights At Age 18 vs Weights At Age 9")
```

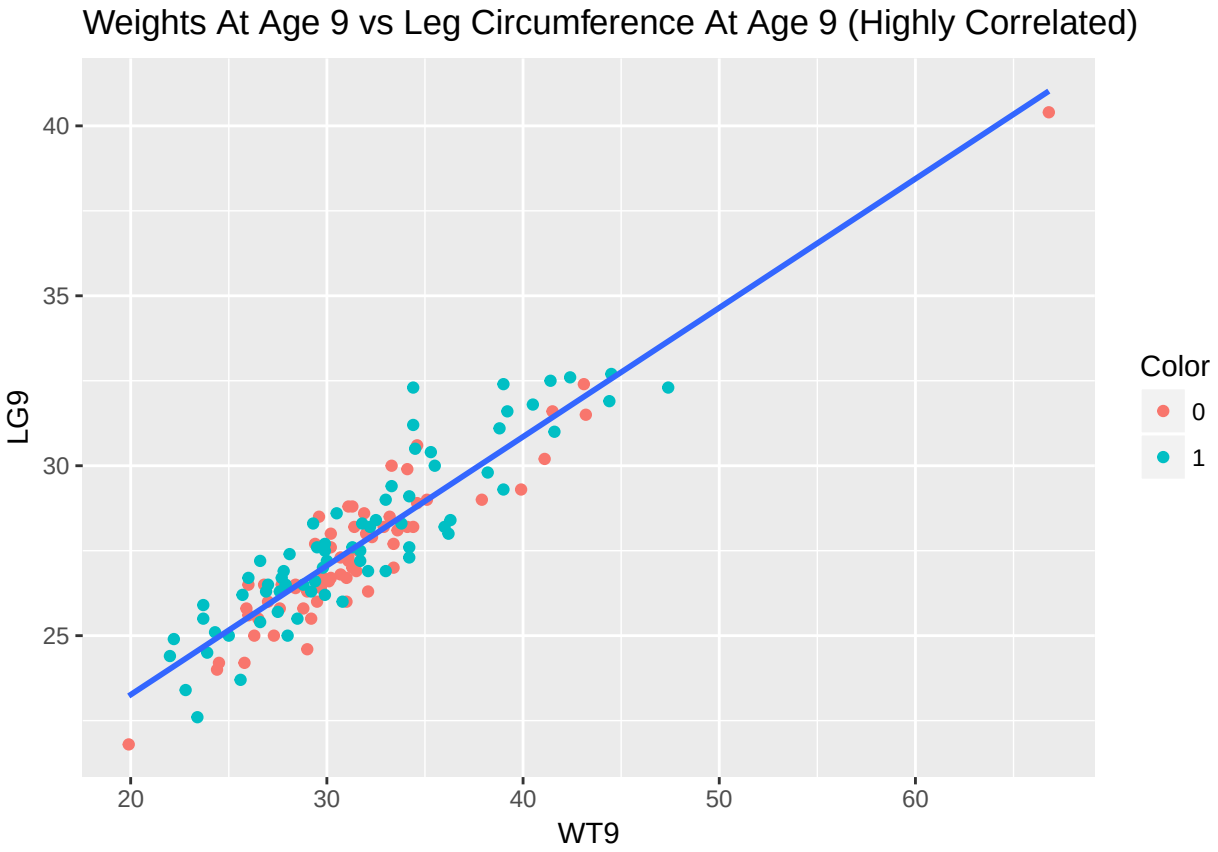



```
############################## 3a) M1: WT18 ON WT9 AND LG9 ####################################
M2 <- lm(WT18 ~ WT9 + LG9, data = berkeley_data)
summary(M2)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9 + LG9, data = berkeley_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.890  -6.406  -0.215   5.536  35.073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.9298    12.9180   2.549  0.01193 *
## WT9           1.0627     0.3226   3.294  0.00127 **
## LG9          -0.0603     0.7825  -0.077  0.93869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.749 on 133 degrees of freedom
## Multiple R-squared:  0.3381, Adjusted R-squared:  0.3282
## F-statistic: 33.97 on 2 and 133 DF,  p-value: 1.207e-12
```

Correlation

```
## [1] 0.9203837
```



The correlation coefficient between LG9 and WT9 is found to be 0.9204. This implies that WT9 and LG9 exhibit a strong positive linear relationship. So, we can drop one of the predictors while regressing on HT18. This is because a model with two highly correlated predictors imparts nearly the same information to the regression model. But, by including both we are actually weakening the model. We are not adding incremental information. Instead, we are infusing your model with noise. Hence, considering WT9 would be enough in the regression model.

Answer 3b

This matrix is called as the hat matrix because it puts a hat on the column vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^T$. This means that it transforms vector of observed responses ($\mathbf{Y}$) into a vector of fitted responses ($\hat{\mathbf{Y}}$) using $\hat{\mathbf{Y}} = \mathbf{HY}$.

Answer 3b ii

For Regression M1,

```
############################## 3b ii) Hat Matrix For M1  ######################################
X.M1 <- as.matrix(cbind(1,berkeley_data$WT9))
H.M1 <- round(X.M1 %*% solve(t(X.M1) %*% X.M1) %*% t(X.M1), digits=5)
sum(diag(H.M1))
```

```
## [1] 2
```

```
max(diag(H.M1))
```

```
## [1] 0.26456
```

```
which.max(diag(H.M1))
```

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

```
############################## 3b ii) Hat Matrix For M2  ###################################
X.M2 <- as.matrix(cbind(1,berkeley_data$WT9,berkeley_data$LG9))
H.M2 <- round(X.M2 %*% solve(t(X.M2) %*% X.M2) %*% t(X.M2), digits=5)
sum(diag(H.M2))
```

```
## [1] 3.00003
```

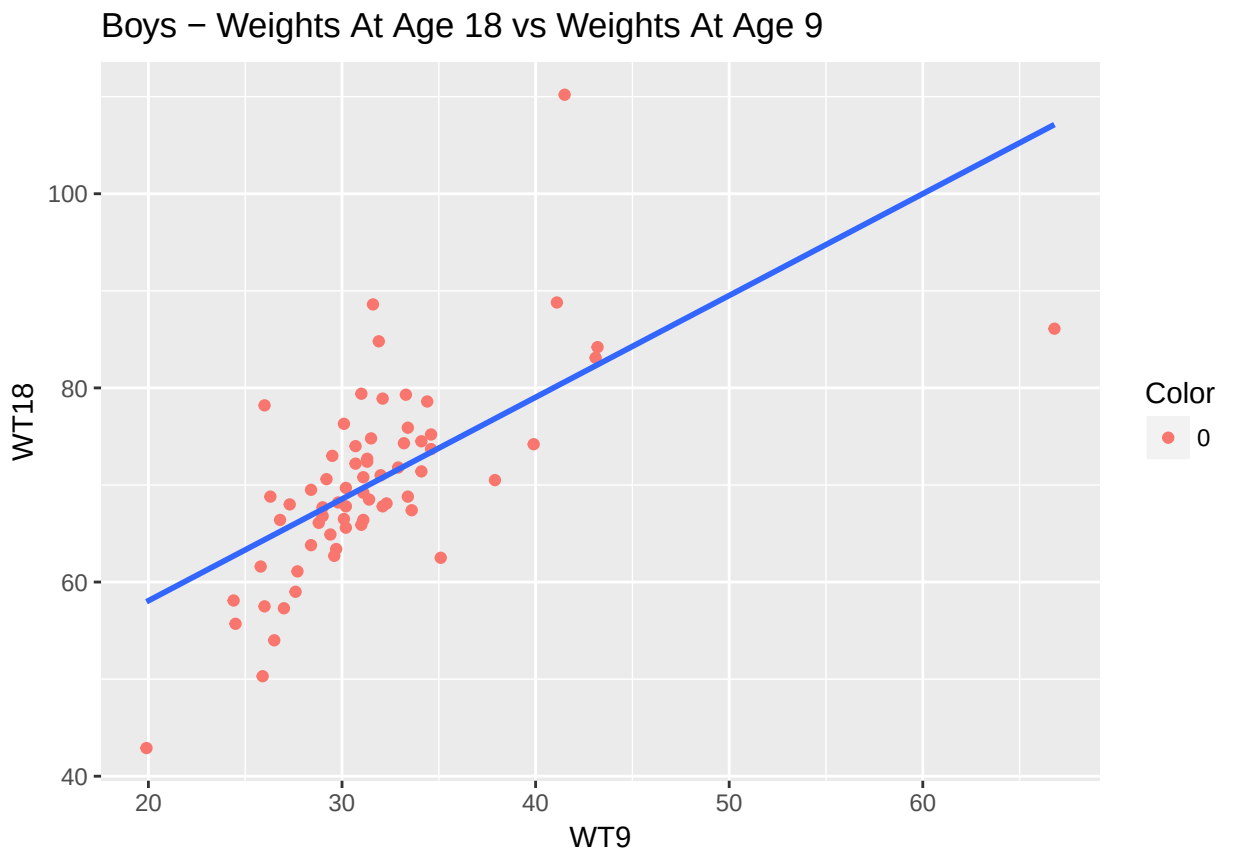
```
max(diag(H.M2))
```

```
## [1] 0.26767
```

The Hat matrix (H1) has be calculated for boys data and one of the elements of H1 had a high leverage of 0.26 while the other elements of the order 10_{-3}.

Answer 3b iii

```
######################### 3b iii) Boys Data With High Leverage Point ########################
boys_data <- subset(berkeley_data, berkeley_data$Sex=="0")
B1 <- lm(WT18 ~ WT9, data = boys_data)
ggplot(boys_data,aes(x=WT9, y=WT18))+
  geom_point(aes(colour=Color))+
  geom_smooth(method=lm, se=FALSE)+
  ggtitle("Boys - Weights At Age 18 vs Weights At Age 9")
```




```
summary(B1)
```

Call: lm(formula = WT18 ~ WT9, data = boys_data)

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

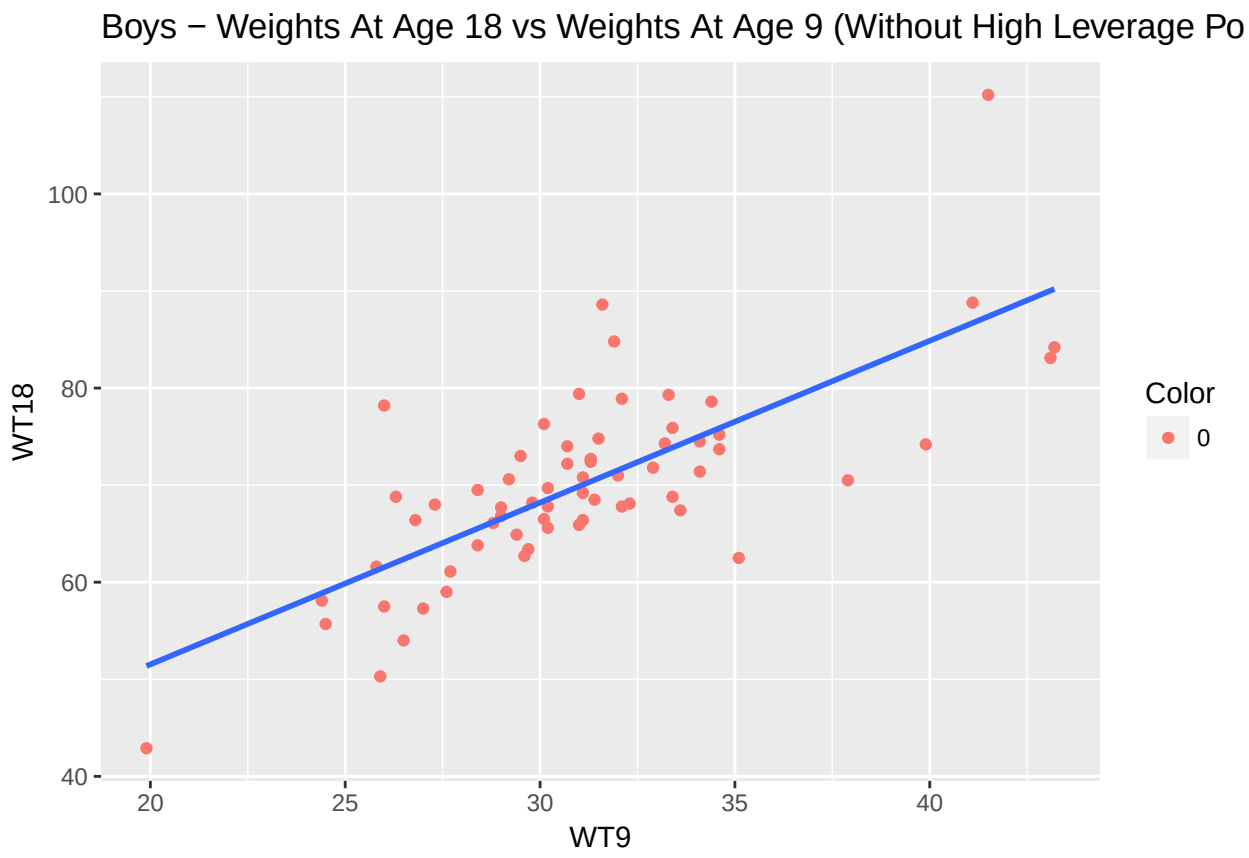
```
library(pander)
panderOptions("digits", 6)
panderOptions('knitr.auto.asis', FALSE)
pander(B1)
```

Table 11: Fitting linear model: WT18 ~ WT9

	Estimate	Std. Error	t value	Pr(>|t|)
WT9	1.04808	0.154218	6.79611	4.23486e-09
(Intercept)	37.1124	4.96857	7.46942	2.77884e-10

```
####################### 3b iii) Boys Data Without High Leverage Point ######################
boys_without_high_lvg <- boys_data[-(which.max(diag(H.M1))),]
B2 <- lm(WT18 ~ WT9, data = boys_without_high_lvg)
```

```
ggplot(boys_without_high_lvg,aes(x=WT9, y=WT18))+
  geom_point(aes(colour=Color))+
  geom_smooth(method=lm, se=FALSE)+
  ggtitle("Boys - Weights At Age 18 vs Weights At Age 9 (Without High Leverage Point)")
```




```
summary(B2)
```

Call: lm(formula = WT18 ~ WT9, data = boys_without_high_lvg)

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

```
pander(B2)
```

Table 12: Fitting linear model: WT18 ~ WT9

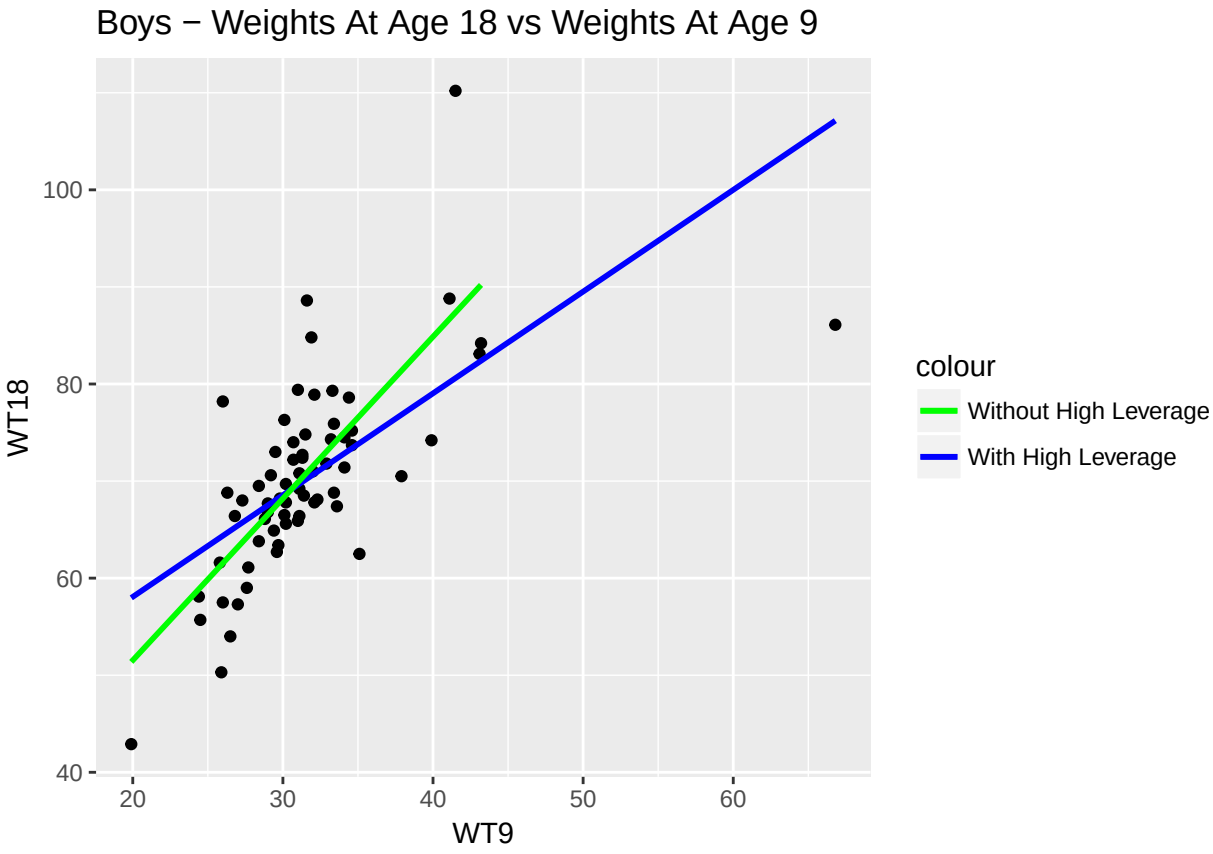
	Estimate	Std. Error	t value	Pr(>|t|)
WT9	1.66669	0.192918	8.63939	2.73414e-12
(Intercept)	18.2029	6.05561	3.00595	0.00379843

The regression coefficient of WT9 in B1 is 1.04808 with a t value of 6.79611. Whereas, the same in B2 is

1.66669 with a t value of 8.63939. Hence, it is found to be statistically signifcant in B2.

Answer 3b iv

```
################### 3b iv) Regression With & Without High Leverage Point ###################
ggplot(boys_data,aes(x=WT9, y=WT18))+
  geom_point()+
  geom_smooth(data=boys_data, aes(x=WT9, y=WT18, colour="green"), method=lm, se=FALSE)+
  geom_smooth(data=boys_without_high_lvg, aes(x=WT9, y=WT18, colour="blue"), method=lm, se=FALSE)+
  scale_colour_manual(values = c("green", "blue"), labels = c("Without High Leverage", "With High Lever
  ggtitle("Boys - Weights At Age 18 vs Weights At Age 9")
```



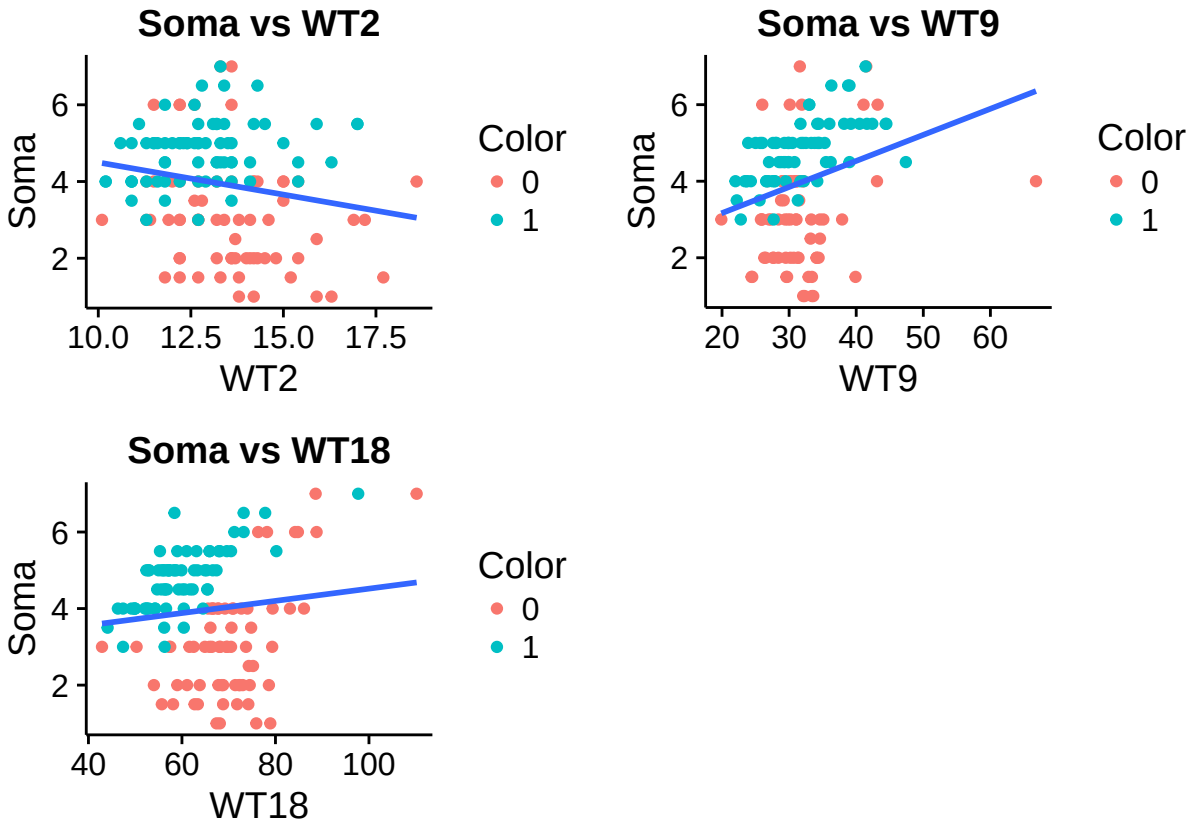
Answer 3b v

The R^2 and adjusted R^2 values for the model without the high leverage point(B2) have been found to be higher than those of the model B1. Hence, B2 can be considered as a better fit to the data.

Table 13: Fitting linear model: WT18 ~ WT9

	Estimate	Std. Error	t value	Pr(>|t|)
WT9	1.66669	0.192918	8.63939	2.73414e-12
(Intercept)	18.2029	6.05561	3.00595	0.00379843

Answer 4a



The response variable Somatotype exhibits a negative relationship against weight at the age of 2. However, it is seen to be increasing with weights at age 9 and 18 with the regression line being steeper against WT9. The relationship between somatotype and WT9 would have been stronger if the high leverage point in WT9 was not considered.

Answer 4b

```
berkeley_data$DW9 <- (berkeley_data$WT9 - berkeley_data$WT2)
berkeley_data$DW18 <- (berkeley_data$WT18 - berkeley_data$WT9)
berkeley_data$AVE <- (berkeley_data$WT2 + berkeley_data$WT9 + berkeley_data$WT18)/3
berkeley_data$LIN <- (berkeley_data$WT18 - berkeley_data$WT2)
berkeley_data$QUAD <- (berkeley_data$WT2 - 2*berkeley_data$WT9 + berkeley_data$WT18)
```

Answer 4c i

```
M1 <- lm(Soma ~ WT2 +WT9 +WT18, data = berkeley_data)
M2 <- lm(Soma ~ WT2 + DW9 + DW18, data = berkeley_data)
M3 <- lm(Soma ~ AVE + LIN + QUAD, data = berkeley_data)
summary(M1)
```

Call: lm(formula = Soma ~ WT2 + WT9 + WT18, data = berkeley_data)

Residuals: Min 1Q Median 3Q Max -2.90048 -0.79326 0.05584 0.90721 3.15680

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.053198 0.929930 6.509 1.44e-09 ***WT2 -0.524132 0.083462 -6.280 4.54e-09*** WT9 0.158270 0.025118 6.301 4.08e-09 *** WT18 -0.002713 0.012129 -0.224 0.823
— Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.219 on 132 degrees of freedom Multiple R-squared: 0.2943, Adjusted R-squared: 0.2783 F-statistic: 18.35 on 3 and 132 DF, p-value: 5.189e-10

```
summary(M2)
```

Call: lm(formula = Soma ~ WT2 + DW9 + DW18, data = berkeley_data)

Residuals: Min 1Q Median 3Q Max -2.90048 -0.79326 0.05584 0.90721 3.15680

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.053198 0.929930 6.509 1.44e-09 ***WT2 -0.368575 0.071785 -5.134 9.91e-07*** DW9 0.155557 0.022394 6.947 1.54e-10 *** DW18 -0.002713 0.012129 -0.224 0.823
— Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.219 on 132 degrees of freedom Multiple R-squared: 0.2943, Adjusted R-squared: 0.2783 F-statistic: 18.35 on 3 and 132 DF, p-value: 5.189e-10

```
summary(M3)
```

Call: lm(formula = Soma ~ AVE + LIN + QUAD, data = berkeley_data)

Residuals: Min 1Q Median 3Q Max -2.90048 -0.79326 0.05584 0.90721 3.15680

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.05320 0.92993 6.509 1.44e-09 ***AVE -0.36858 0.07178 -5.134 9.91e-07*** LIN 0.26071 0.04269 6.107 1.06e-08 ***QUAD -0.14056 0.01992 -7.055 8.74e-11*** — Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.219 on 132 degrees of freedom Multiple R-squared: 0.2943, Adjusted R-squared: 0.2783 F-statistic: 18.35 on 3 and 132 DF, p-value: 5.189e-10

```
#install.packages("stargazer")
#library(stargazer)
#stargazer(M1, M2, M3, title="Results", align=TRUE, type = "latex")

#install.packages("pander")
#install.packages("memisc")
library(pander)
panderOptions("digits", 6)
panderOptions('knitr.auto.asis', FALSE)
pander(M1)
```

Table 14: Fitting linear model: Soma ~ WT2 + WT9 + WT18

	Estimate	Std. Error	t value	Pr(>|t|)
WT2	-0.524132	0.0834619	-6.2799	4.53564e-09
WT9	0.15827	0.0251175	6.30117	4.08183e-09
WT18	-0.00271291	0.0121286	-0.223679	0.823353
(Intercept)	6.0532	0.92993	6.50931	1.44219e-09

```
pander(M2)
```

Table 15: Fitting linear model: Soma ~ WT2 + DW9 + DW18

	Estimate	Std. Error	t value	Pr(>|t|)
WT2	-0.368575	0.0717848	-5.13445	9.90894e-07
DW9	0.155557	0.0223935	6.94652	1.54171e-10
DW18	-0.00271291	0.0121286	-0.223679	0.823353
(Intercept)	6.0532	0.92993	6.50931	1.44219e-09

```
pander(M3)
```

Table 16: Fitting linear model: Soma ~ AVE + LIN + QUAD

	Estimate	Std. Error	t value	Pr(>|t|)
AVE	-0.368575	0.0717848	-5.13445	9.90894e-07
LIN	0.26071	0.0426891	6.10718	1.06032e-08
QUAD	-0.140564	0.0199227	-7.05548	8.74454e-11
(Intercept)	6.0532	0.92993	6.50931	1.44219e-09

```
#library(memisc)
#panderOptions('table.alignment.rownames', 'left')
#pander(mtable(M1,M2,M3))

#install.packages("texreg")
#install.packages("xtable")
#library(xtable)
#library(texreg)
#print(xtable(M1), type = "html")
```

Same Regression Coefficients:
Intercept in M1, M2 and M3
WT18 in M1 and **DW18** in M2
WT2 in M2 and **AVE** in M3

Answer 4c ii

In Model 1, the estimate for WT18 is the effect on Somatotype of changing WT18 by one unit, with all other terms held fixed. In Model 2, the estimate for DW18 is the change in Somatotype when DW18 changes by one unit, when all other terms are held fixed. But the only way $DW18 = WT18 - WT9$ can be changed by one unit with the other variables including $WT9 = DW9 - WT2$ held fixed is by changing WT18 by one unit. Consequently, the terms WT18 in Model 1 and DW18 in Model 2 are same.

In Model 1, the estimate for WT9 is the effect on Somatotype of changing WT9 by one unit, with WT2 and WT18 fixed. In Model 2, the estimate for DW9 is the change in Somatotype when DW9 changes by one unit, when DW18 and WT2 are fixed. As $DW9 = WT9 - WT2$, the only way to change DW9 is by changing WT9 as WT2 is held constant. But in order to keep $DW18 = WT18 - WT9$ constant, we need to change WT18 with a corresponding change in WT2. Hence, the difference in estimates for WT9 in M1 and DW9 in M2 is accounted to the variation in WT18 in M2 while WT18 and WT2 are both held constant in M1.

Answer 4c iii

Let the regression coefficients of model M2 be $\gamma_0, \gamma_1, \gamma_2 and \gamma_3$. Thus, the model M2 can be represented by the below equation:

$$Somatotype = \gamma_0 + \gamma_1 AVE + \gamma_2 LIN + \gamma_3 QUAD$$
$$Somatotype = \gamma_0 + \frac{\gamma_1}{3}(WT2 + WT9 + WT18) + \gamma_2(WT18 - WT2) + \gamma_3(WT2 - 2 * WT9 + WT18)$$
$$Somatotype = \gamma_0 + WT2(\frac{\gamma_1}{3} - \gamma_2 + \gamma_3) + WT9(\frac{\gamma_1}{3} - 2\gamma_3) + WT18(\frac{\gamma_1}{3} + \gamma_2 + \gamma_3)$$

Let the regression coefficients of model M2 be $\beta_0, \beta_1, \beta_2 and \beta_3$. Then M1 ca be represented as

$$Somatotype = \beta_0 + \beta_1 WT2 + \beta_2 WT9 + \beta_3 WT18$$

Hence by comparing the two models we can say that

$$\beta_0 = \gamma_0$$
$$\beta_1 = (\frac{\gamma_1}{3} - \gamma_2 + \gamma_3)$$
$$\beta_2 = (\frac{\gamma_1}{3} - 2\gamma_3)$$
$$\beta_3 = (\frac{\gamma_1}{3} + \gamma_2 + \gamma_3)$$

Answer 4d

```
M4 <- lm(Soma ~ WT2 +WT9 +WT18 +DW9, data = berkeley_data)
summary(M4)
```

```
##
## Call:
## lm(formula = Soma ~ WT2 + WT9 + WT18 + DW9, data = berkeley_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90048 -0.79326  0.05584  0.90721  3.15680
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.053198   0.929930   6.509 1.44e-09 ***
## WT2         -0.524132   0.083462  -6.280 4.54e-09 ***
## WT9          0.158270   0.025118   6.301 4.08e-09 ***
## WT18        -0.002713   0.012129  -0.224    0.823
## DW9               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.219 on 132 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.2783
## F-statistic: 18.35 on 3 and 132 DF,  p-value: 5.189e-10
```

```
#panderOptions("digits", 6)
#panderOptions('knitr.auto.asis', FALSE)
#pander(M4)
```

As the number of linearly independent quantitites is three, we cannot use more than three linear combinations of the predictors. Since DW9 can be written as an exact linear combination of the other predictors, DW9 = WT9 - WT2, the residuals from this regression are all exactly zero. A slope coefficient for DW9 is not defined after adjusting for the other three terms. Hence, the four terms WT2, WT9, WT18, and DW9 are linearly dependent, since one can be determined exactly from the others.
Thus, the regression coefficient of DW9 is set to "NA" to indicate that the predictor has been aliased and hence not estimated.