Assignment 2

Ri Liu (rl4508)

3/21/2022

Part I: Exploring Various Models

The Berkeley Guidance Study, under the direction of Jean Macfarlane, started with a sample of infants who were born in Berkeley, California in 1928-1929. Most of the children were Caucasian and Protestant, and two-thirds came from middle-class families. The basic cohort includes 136 of these children who participated in the study through the 1930s and up to the end of World War II. Annual data collection ended in 1946. In this project, you are asked to prepare a short data analysis using these data. The dataset contains a short list of variables pertaining to the child at three time points: age 2, age 9 and age 18.

The variables collected in this study include: Sex(0:male, 1:female), Height (cm) and Weight (kg) at ages 2, 9 and 18, leg circumference (cm) and strength (kg) at ages 9 and 18, and Somatotype (a 1 = thinner to 7 = heavier scale of body type).

1. **Model height growth from age 2 to age 9 by answering the following questions:**

(a) **Create a scatter plot of heights at age 9 on heights at age 2, with the points colored based on the gender of the child. Does there appear to be a different pattern for boys than for girls?**

```
library(dplyr)
```

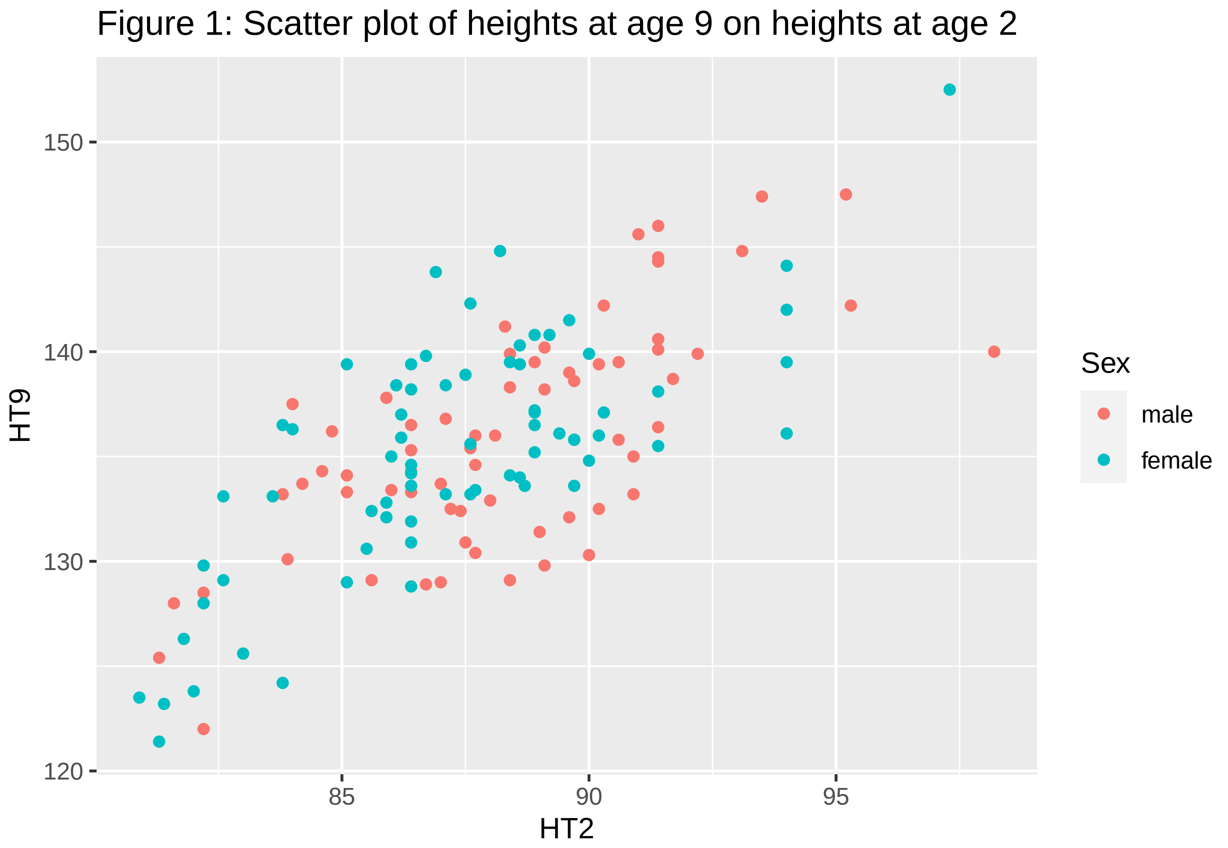
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
BGS <- read.csv('BGS.csv')
head(BGS, 3)
```

```
##   X Sex  WT2  HT2  WT9   HT9  LG9 ST9  WT18  HT18 LG18 ST18 Soma
## 1 1   0 13.6 90.2 41.5 139.4 31.6  74 110.2 179.0 44.1  226    7
## 2 2   0 12.7 91.4 31.0 144.3 26.0  73  79.4 195.1 36.1  252    4
## 3 3   0 12.6 86.4 30.1 136.5 26.6  64  76.3 183.7 36.9  216    6
```

```
# Change `sex` to factor
BGS$Sex <- factor(BGS$Sex,levels = c(0, 1), label=c('male', 'female'))
```

```
library(ggplot2)
ggplot(data = BGS)+
  geom_point(aes(x=HT2, y=HT9, color=Sex ))+
  ggtitle('Figure 1: Scatter plot of heights at age 9 on heights at age 2')
```



There is no constantly obvious patterns that are different between boys and girls. It is very close between boys and girls, though in the middle part girls have some higher points on HT9.

(b) **Fit a simple linear regression of heights at age 9 on heights at age 2.**

```
model_1 <- lm(HT9 ~ HT2, data = BGS)
summary(model_1)
```

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

- **Report and interpret the estimated regression coefficients.**

The estimated coefficient of `HT2` is 1.18. A one cm increase in height at age 2 is associated with an increase of 1.18 in the expected height at age 9.

- **Test the hypothesis of H0 : beta1 = 0 against the two-sided alternative.**

When the null hypothesis is that beta_1 is 0, the t statistic is 12.052, and the two-sided p-value is less than 4e-16. So we can reject the null hypothesis at 0.05 level.

- **Show numerically 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.**

```r
sqrt(summary(model_1)$fstatistic[1])
```

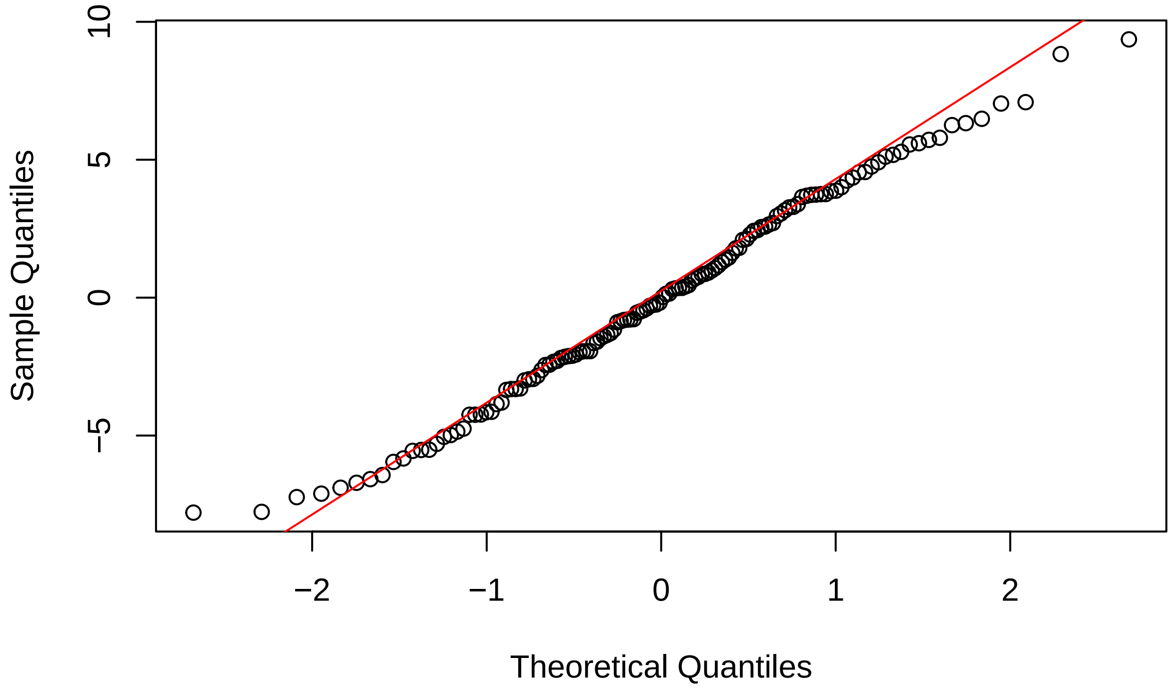
```
##    value 
## 12.05192
```

We can see that the square root of the F-statistic 12.052 equals the t-statistic for the only predictor 12.052.

- **Check the normality and homoscedasticity assumptions on the residuals. Include any plots you consult.**

```r
qqnorm(model_1$residuals, main = 'Figure 2: Quantile- Quantile plot of the residuals')
qqline(residuals(model_1), col = 'red')
```

Figure 2: Quantile− Quantile plot of the residuals



```
plot(model_1$residuals ~ model_1$fitted.values, main = 'Figure 3: Residuals ~ Fitted Values')
```

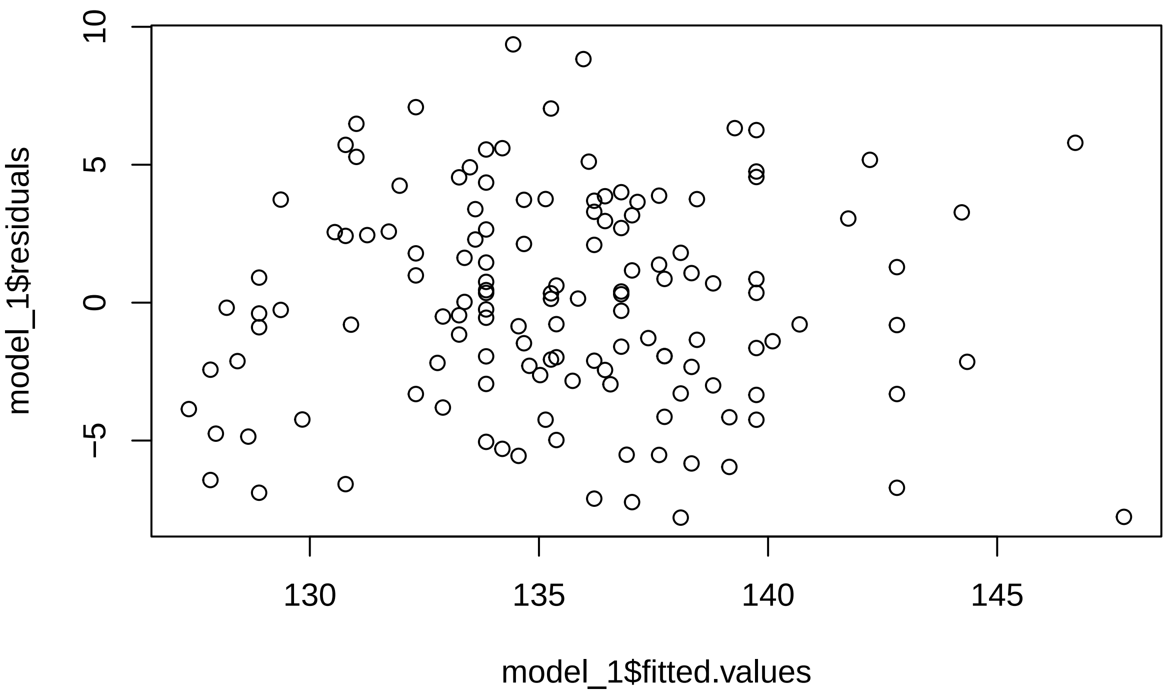


Figure 3: Residuals ~ Fitted Values

According to the Quantile-Quantile plot, the normality assumption on the residuals is not perfectly satisfied, because the lower-left and upper-right of the points are not close to the line. According to the residuals vs. fitted values plot, the homoscedasticity assumption on the residuals is satisfied.

(c) **Considering a model that allows for separate intercepts for boys and girls, is this model better than the simple linear regression fit above?**

```
model_1_c <- lm(HT9 ~ HT2+Sex, data = BGS)
summary(model_1_c)
```

```
##
## Call:
## lm(formula = HT9 ~ HT2 + Sex, data = BGS)
##
## 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 ***
## Sexfemale    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(model_1, model_1_c)
```

```
##           df      AIC
## model_1    3 754.5998
## model_1_c  4 755.8636
```

According to figure 1, for boys and girls who have similar height at age 2, they have close height at age 9. And according to the adjusted R-squared and AIC, the model having separate intercepts for boys and girls does not have a better fit.

(d) **Considering a model that allows for both the separate slope and separate intercepts for boys and for girls, is this model better than the simple linear regression fit above?**

```
model_1_d <- lm(HT9 ~ HT2+Sex+HT2*Sex, data = BGS)
summary(model_1_d)
```

```
##
## Call:
## lm(formula = HT9 ~ HT2 + Sex + HT2 * Sex, data = BGS)
##
## 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 ***
## Sexfemale      -8.6231    17.5263  -0.492  0.62353
## HT2:Sexfemale   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(model_1, model_1_d)
```

```
##           df      AIC
## model_1    3 754.5998
## model_1_d  5 757.5803
```

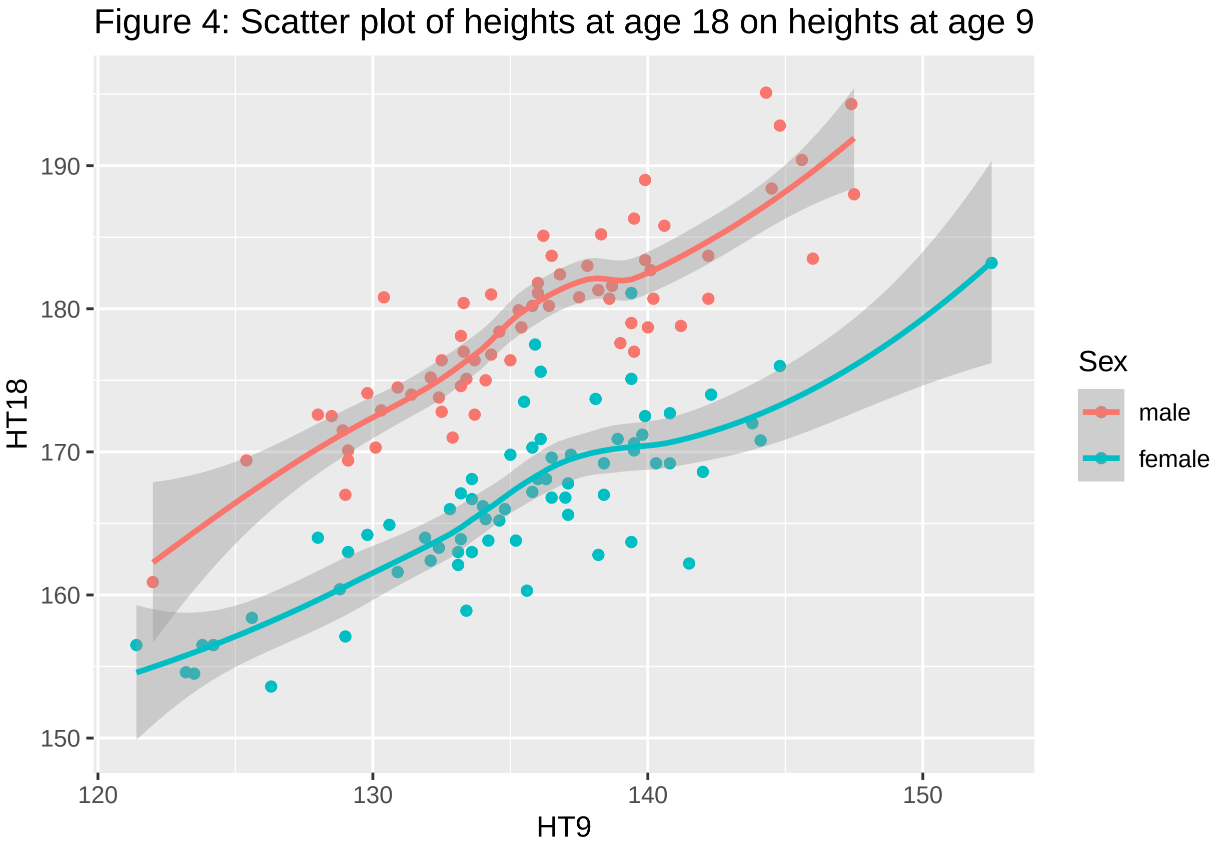
According to the adjusted R-squared and AIC, the model having separate intercepts and slope for boys and girls does not have a better fit.

2. **Model height growth from age 9 to age 18 by answering the following questions:**

(a) **Create a scatter plot of heights at age 18 on heights at age 9, with the points colored based on the gender of the child. Does there appear to be a different pattern for boys than for girls?**

```
ggplot(data = BGS)+
  geom_point(aes(x=HT9, y=HT18, color=Sex ))+
  geom_smooth(aes(x=HT9, y=HT18, color=Sex ))+
  ggtitle('Figure 4: Scatter plot of heights at age 18 on heights at age 9')
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



For boys and girls who have the same height at age 9, boys tend to be have obvious higher heights at age 18, and the difference is larger as the height at age 9 increasing.

(b) **Fit a simple linear regression of heights at age 18 on heights at age 9. Report the estimated regression coefficients.**

```
model_2 <- lm(HT18 ~ HT9, data = BGS)
summary(model_2)
```

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

```
## lm(formula = HT18 ~ HT9, data = BGS)
##
## 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 estimated coefficient of `HT9` is 1.035. A one cm increase in height at age 9 is associated with an increase of 1.035 in the expected height at age 18.

(c) **Considering a model that allows for separate intercepts for boys and girls, is this model better than the simple linear regression fit above?**

```
model_2_c <- lm(HT18 ~ HT9+Sex, data = BGS)
summary(model_2_c)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + Sex, data = BGS)
##
## 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 ***
## Sexfemale   -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(model_2, model_2_c)
```

```
##           df      AIC
## model_2    3 911.2241
## model_2_c  4 726.3621
```

It is obvious in the plot that boys and girls have different `HT18` on the far left. And according to adjusted R-squared and AIC, a model having separate intercepts for boys and girls will be better than the simple linear regression.

(d) **Considering a model that allows for both the separate slope and separate intercepts for boys and for girls, is this model better than the simple linear regression fit above?**

```
model_2_d <- lm(HT18 ~ HT9+Sex+HT9*Sex, data = BGS)
summary(model_2_d)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + Sex + HT9 * Sex, data = BGS)
##
## 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 ***
## Sexfemale      13.32748   14.54695   0.916   0.3612
## HT9:Sexfemale  -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
```

```
AIC(model_2, model_2_c, model_2_d)
```

```
##           df      AIC
## model_2    3 911.2241
## model_2_c  4 726.3621
## model_2_d  5 725.3423
```

According to adjusted R-squared and AIC, a model having separate slope and separate intercepts for boys and girls is better than the simple linear regression. But the improvement compared to `model_2_c` is not too much.

(e) **Choose which of the above 3 models you think best describes the data and interpret the parameter estimates for this model.**

The model having different slope and intercept for boys and girls is the best.

3. **Create a new dataset that includes only the boys in the sample. Use this new dataset to model the change in weight from age 9 to age 18.**

(a) **Fit two linear regression models: (M1) Weight at age 18 on weight at age 9 and (M2) Weight at age 18 on weight at age 9 and leg circumference at age 9. Explain why weight at age 9 is significant in one model but not the other. Justify your answer by calculating the appropriate correlation coefficient.**

```
boys <- BGS[BGS$Sex == 'male', ]
```

```
M1 <- lm(WT18 ~ WT9, data = boys)
summary(M1)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = boys)
##
## 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
```

```
M2 <- lm(WT18 ~ WT9+LG9, data = boys)
summary(M2)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9 + LG9, data = boys)
##
## 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 between WT9 and LG9
cor(boys$WT9, boys$LG9)
```

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

In M2, `WT9` and `LG9` is highly correlated. In the model those two are competing to interpret `WT18`.

(b) **The hat matrix can be calculated as $H = X(X^TX)^{-1}X^T$, where X is the design matrix. The diagonal values of the hat matrix determine the leverage that each point has in the fit of the regression model.**

- **Explain why this matrix is known as the hat matrix.**

We know the estimated β is $(X^TX)^{-1}X^Ty$,

and fitted values $\hat{y} = X(X^TX)^{-1}X^Ty$,

then $\hat{y} = Hy$.

So the hat matrix is a transformation on y, through which we can get the estimated $\hat{y}$.

- **Calculate this matrix in R using the design matrix corresponding to this set of questions. Show that the leverage of one of the points is much higher than any of the other points.**

```
# For M1: WT18 ~ WT9
# Design matrix X:
X <- model.matrix(M1)
# Hat matrix:
H <- X %*% solve(t(X) %*% X) %*% t(X)
# Check the distribution of the diagnal values:
hist(diag(H),breaks = 10, main = 'Figure 5: Distribution of diagonal values of M1', xlab = 'Diagonal val
```

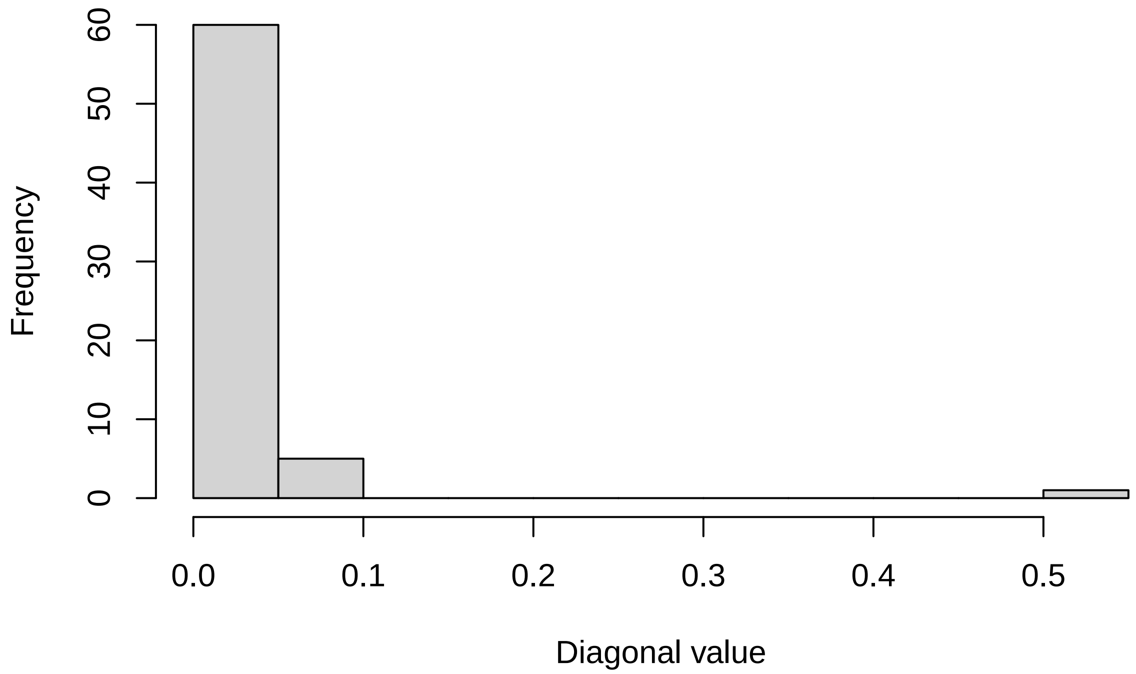


Figure 5: Distribution of diagonal values of M1

We can find that there is a diagonal value that is far from other diagonal values.

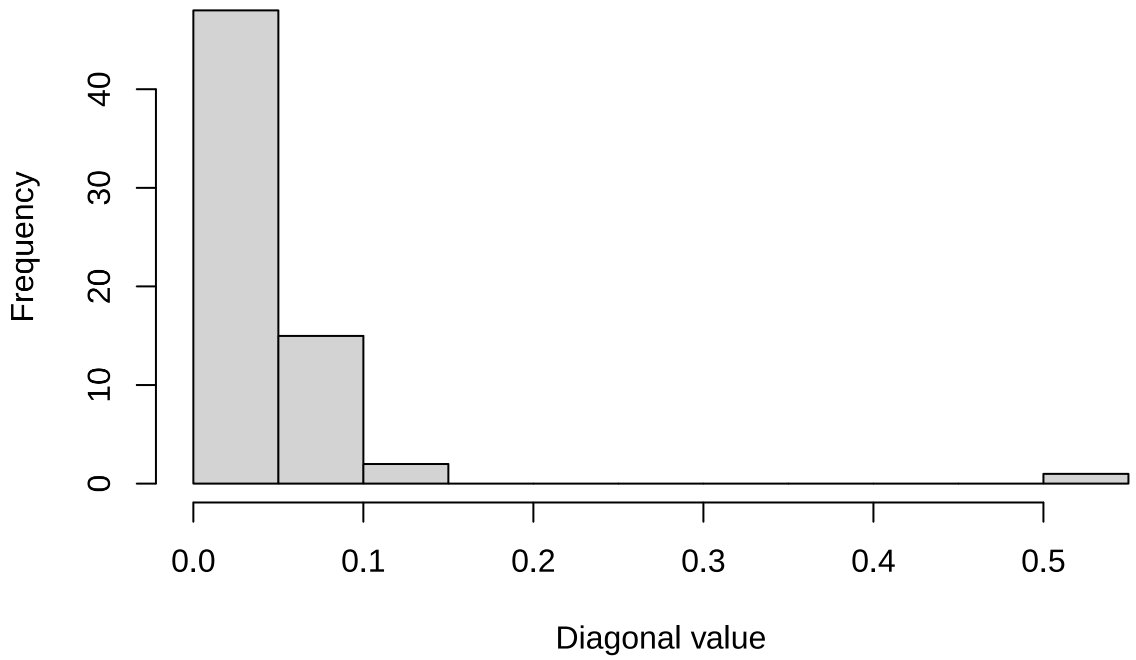
```
which(diag(H) > 0.3)
```

```
## 60
## 60
```

And we can know it is the 60th subject in the boy data set.

```
# For M2: WT18 ~ WT9+LG9
# Design matrix X:
X <- model.matrix(M2)
# Hat matrix:
H <- X %*% solve(t(X) %*% X) %*% t(X)
# Check the distribution of the diagonal values:
hist(diag(H),breaks = 10, main = 'Figure 6: Distribution of diagonal values of M2', xlab = 'Diagonal val
```

Figure 6: Distribution of diagonal values of M2



We have the same result for M2.

- **Fit two simple linear regression models, both regressing weight at age 18 on weight at age 9. One model should use all of the boys in the dataset, and the other should remove the high-leverage point. Compare the coefficients for weight at age 9 obtained from both models.**

```
# model 1: all of the boys in the dataset
M_all <- lm(WT18 ~ WT9, data = boys)
summary(M_all)
```

```
## 
## Call:
## lm(formula = WT18 ~ WT9, data = boys)
## 
## 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
```

```
boys_removed <- boys[-60, ]
M_removed <- lm(WT18 ~ WT9, data = boys_removed)
summary(M_removed)
```

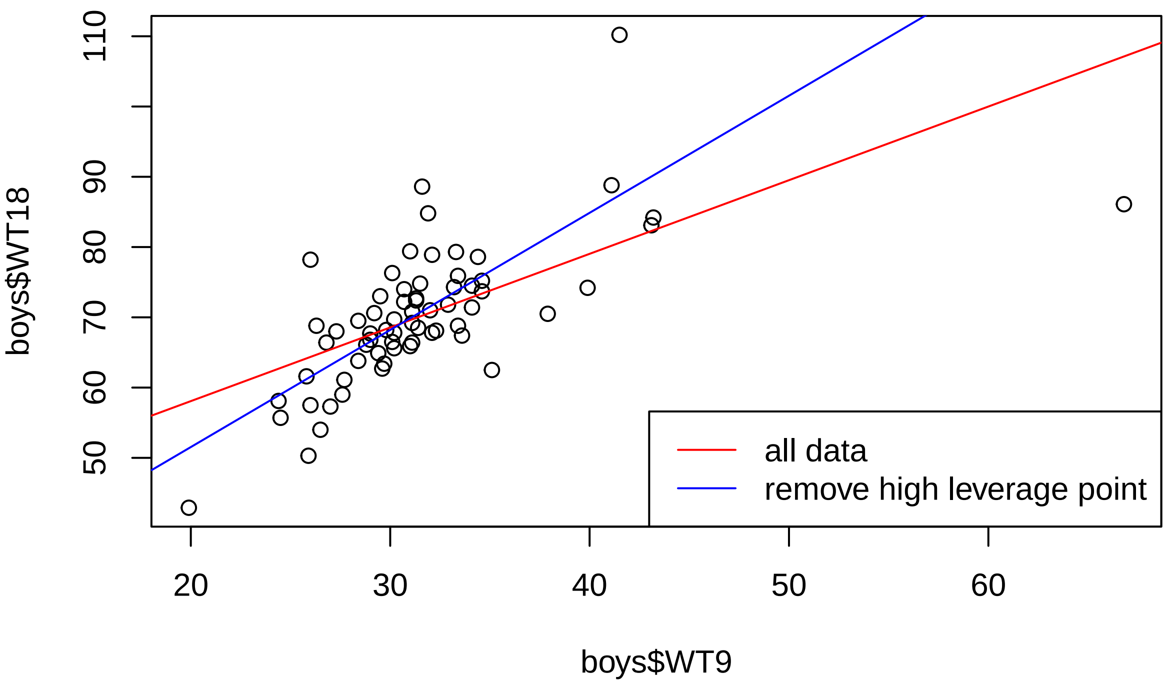
```
##
## Call:
## lm(formula = WT18 ~ WT9, data = boys_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2037  -3.9370  -0.6703   3.0630  22.8295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.2029     6.0556   3.006   0.0038 **
## 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
```

After removing the high-leverage point, the coefficient of `WT9` increased from 1.0481 to 1.6667, the p-value for it also improved from 1e-09 to 1e-12 level.

- **Create a scatter plot of weight at age 18 on weight at age 9. Plot both regression lines fit in the previous part on the plot in different colors.**

```
plot(boys$WT9, boys$WT18, main = 'Figure 7: Scatter plot of weight at age 18 on weight at age 9')
abline(M_all, col='red')
abline(M_removed, col='blue')
legend('bottomright', legend = c('all data', 'remove high leverage point'), col=c('red', 'blue'), lty=1
```

Figure 7: Scatter plot of weight at age 18 on weight at age 9



- **Based on the above parts, which regression line you think better fits the data? Report and interpret the estimated regression parameters for the model you choose.**

The model without the high-leveraged data is better. As we know earlier, it has a high leverage value that is far from other leverage values. So it is potential to be an influential point. According to the scatter plot, it pulled the red regression line toward itself, while away from the trend of the majority.

4. **Create a new dataset that includes only the girls in the sample. Use this new dataset to model Somatotype in the following ways.**

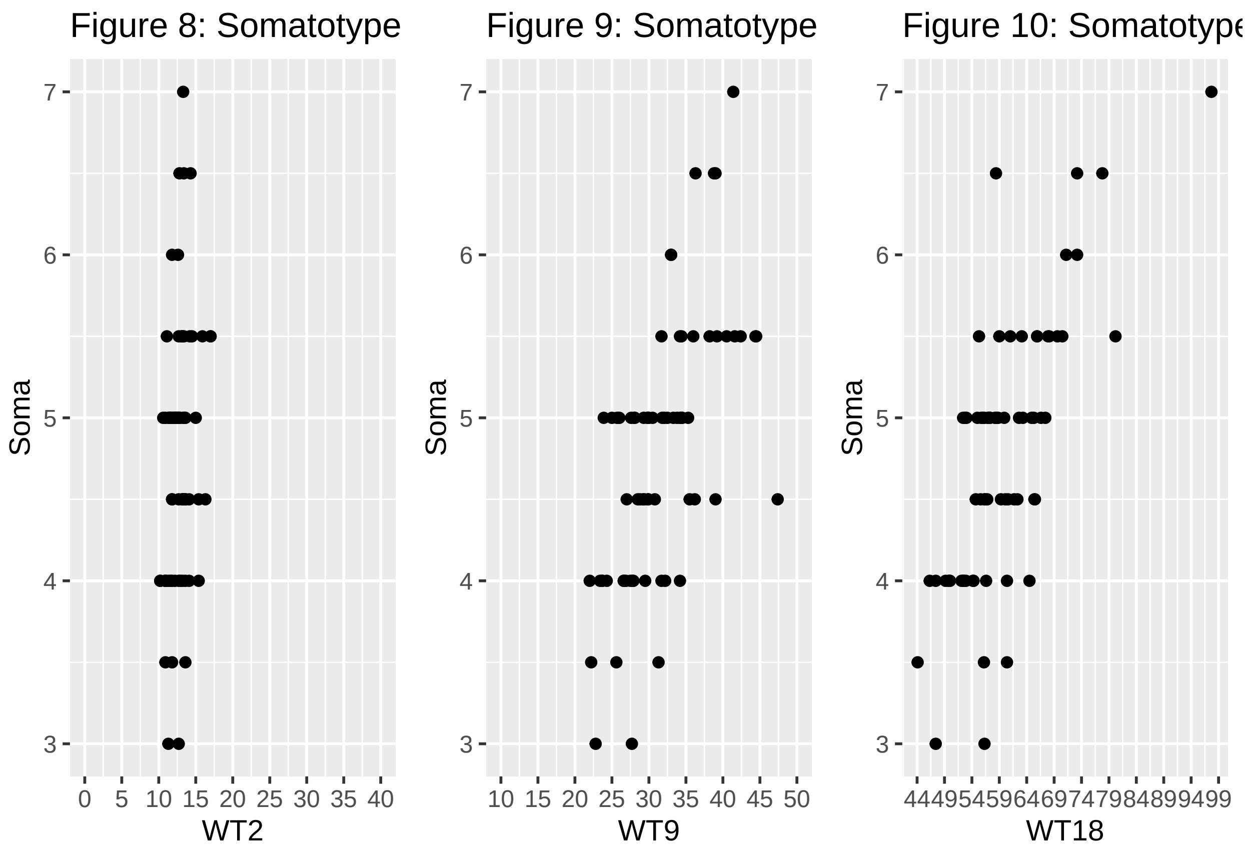
(a) **Plot somatotype against weight at each of the three time points. Comment on how the relationship between weight and somatotype changes over time.**

```
girls <- BGS[BGS$Sex == 'female', ]
```

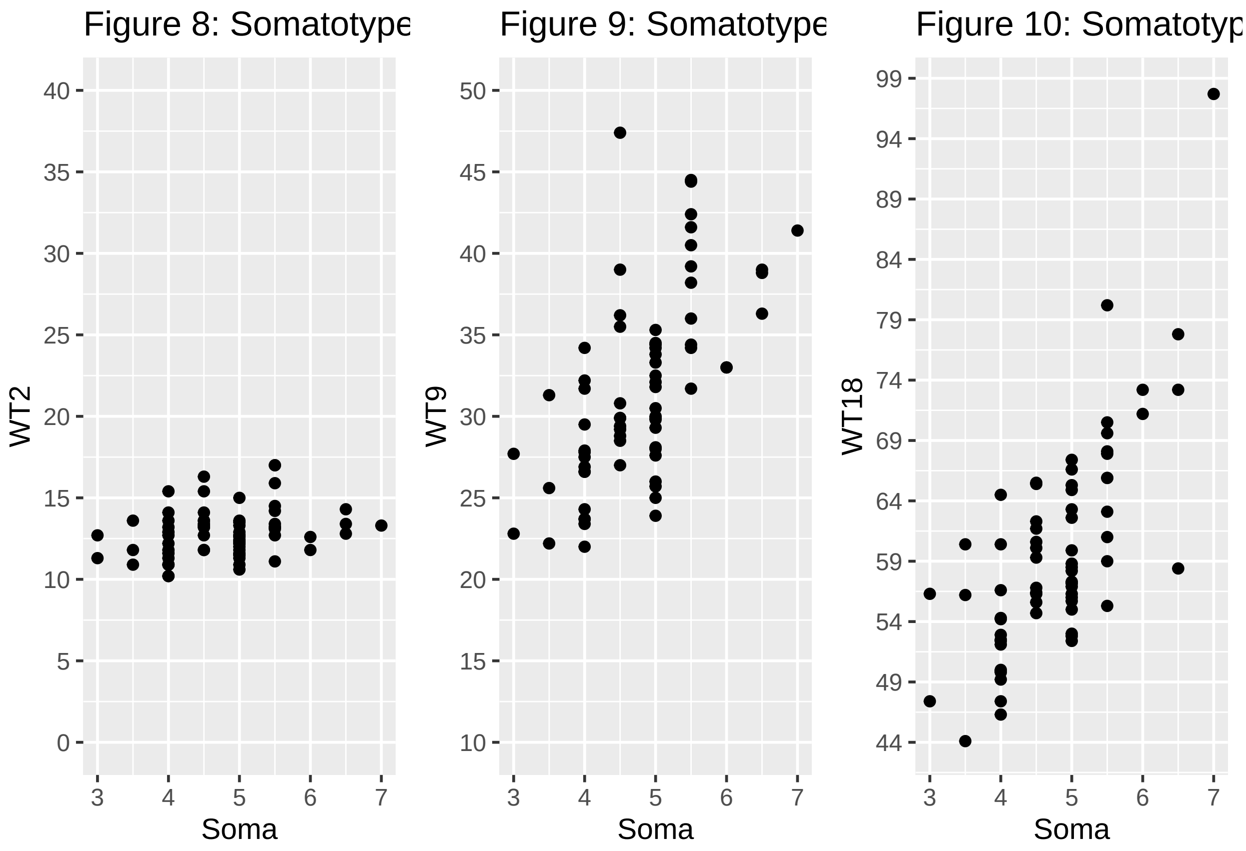
```
library(ggplot2)
library(dplyr)
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
a <- girls %>%
  ggplot(aes(x=WT2, y=Soma)) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 0.0, to = 40, by = 5))+
  coord_cartesian(xlim = c(0.0, 40.0)) +
  ggtitle('Figure 8: Somatotype against weight at age 2')
b <- girls %>%
  ggplot(aes(x=WT9, y=Soma)) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 10.0, to = 50, by = 5))+
  coord_cartesian(xlim = c(10.0, 50.0)) +
  ggtitle('Figure 9: Somatotype against weight at age 9')
c <- girls %>%
  ggplot(aes(x=WT18, y=Soma)) +
  geom_point() +
  scale_x_continuous(breaks = seq(from = 44, to = 99, by = 5))+
  coord_cartesian(xlim = c(44, 98)) +
  ggtitle('Figure 10: Somatotype against weight at age 18')
grid.arrange(a, b, c, nrow=1)
```



```r
a <- girls %>%
  ggplot(aes(y=WT2, x=Soma)) +
  geom_point() +
  scale_y_continuous(breaks = seq(from = 0.0, to = 40, by = 5))+
  coord_cartesian(ylim = c(0.0, 40.0)) +
  ggtitle('Figure 8: Somatotype against weight at age 2')
b <- girls %>%
  ggplot(aes(y=WT9, x=Soma)) +
  geom_point() +
  scale_y_continuous(breaks = seq(from = 10.0, to = 50, by = 5))+
  coord_cartesian(ylim = c(10.0, 50.0)) +
  ggtitle('Figure 9: Somatotype against weight at age 9')
c <- girls %>%
  ggplot(aes(y=WT18, x=Soma)) +
  geom_point() +
  scale_y_continuous(breaks = seq(from = 44, to = 99, by = 5))+
  coord_cartesian(ylim = c(44, 98)) +
  ggtitle('Figure 10: Somatotype against weight at age 18')
grid.arrange(a, b, c, nrow = 1)
```



Along with the age growing, we can see that the weights at certain age are more dispersed for each level of somatotype (the x-axis and y-axis are zoomed into same range) . Weight at 18 and weight at 9 have more potential outliers than weight at age 2. Though all of the 3 weights looks linear correlated with Somatotype, the more dispersed the more difficult to fit a line on them.

(b) **Create new variables:**

$DW9 = WT9 - WT2$
$DW18 = WT18 - WT9$
AVE = 1/3 (WT2+WT9+WT18) LIN =WT18−WT2
$QUAD = WT2 - 2 \cdot WT9 + WT18$

DW9 and DW18 measure the change in weight between consecutive timepoints. AVE, LIN, and QUAD measure the average, linear and quadratic trends over time (since the timepoints are roughly evenly spaced).

```
girls$DW9 <- girls$WT9 - girls$WT2
girls$DW18 <- girls$WT18 - girls$WT9
girls$AVE <- 1/3 * (girls$WT2 + girls$WT9 + girls$WT18)
girls$LIN <- girls$WT18 - girls$WT2
girls$QUAD <- girls$WT2 - 2 * girls$WT9 + girls$WT18
```

(c) **Fit the following three models: M1 : Somatotype WT2 + WT9 + WT18 M2 : Somatotype WT2 + DW9 + DW18 M3 : Somatotype AVE + LIN + QUAD**

```
M1 <- lm(Soma ~ WT2+WT9+WT18, data = girls)
summary(M1)
```

```
##
## Call:
## lm(formula = Soma ~ WT2 + WT9 + WT18, data = girls)
##
## 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
```

```
M2 <- lm(Soma ~ WT2+DW9+DW18, data = girls)
summary(M2)
```

```
##
## Call:
## lm(formula = Soma ~ WT2 + DW9 + DW18, data = girls)
##
## 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
```

```r
M3 <- lm(Soma ~ AVE+LIN+QUAD, data = girls)
summary(M3)
```

```
## 
## Call:
## lm(formula = Soma ~ AVE + LIN + QUAD, data = girls)
## 
## 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
```

Compare and contrast these models by answering the following questions:

- **What attributes of the models are the same across all three models? What attributes of the models are different?**

The same: intercept, residuals, residual standard error, R-squared, adjusted R-squared, F-statistic and p-value of F-test, degree of freedom of the model.

Different: estimated coefficients.

- **Why does the coefficient for DW18 in model 2 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?**

For model 1:

$$Somatotype = \alpha_0 + \alpha_1 * WT2 + \alpha_2 * WT9 + \alpha_3 * WT18$$

For model 2:

$$\begin{aligned}
Somatotype &= \beta_0 + \beta_1 * WT2 + \beta_2 * DW9 + \beta_3 * DW18 \\
&= \beta_0 + \beta_1 * WT2 + \beta_2 * (WT9 - WT2) + \beta_3 * (WT18 - WT9) \\
&= \beta_0 + (\beta_1 - \beta_2) * WT2 + (\beta_2 - \beta_3) * WT9 + \beta_3 * WT18
\end{aligned}$$

As `DW9` is derived from `WT9` and `WT2`, and `DW18` is derived from `WT18` and `WT8`, model 2 can be transformed towards `DW9` and `DW18`. We can see model 2 is a model with `WT2`, `WT9` and `WT18` as covariates in nature, which is the same as model 1. `WT18` is only contained in derivate predictor `DW18`. However, `WT9` is contained in both `DW9` and `DW18`. So the coefficient of `DW18` equals to the coefficient of `WT18`, but the coefficient of `WT9` has to be calculated from the coefficients of `DW9` and `DW18`.

- **Show algebraically (not numerically) why M1 and M3 are equivalent by showing how the coefficients in M3 can be obtained by algebraically manipulating the coefficients in M1.**

For model 1:

$$Somatotype = \alpha_0 + \alpha_1 * WT2 + \alpha_2 * WT9 + \alpha_3 * WT18$$

For model 3:

$$\begin{aligned}
Somatotype &= \beta_0 + \beta_1 * AVE + \beta_2 * LIN + \beta_3 * QUAD \\
&= \beta_0 + \beta_1 * \frac{1}{3}(WT2 + WT9 + WT18) + \beta_2 * (WT18 - WT2) + \beta_3 * (WT2 - 2WT9 + WT18) \\
&= \beta_0 + (\frac{1}{3}\beta_1 - \beta_2 + \beta_3) * WT2 + (\frac{1}{3}\beta_1 - 2\beta_3) * WT9 + (\frac{1}{3}\beta_1 + \beta_2 + \beta_3) * WT18
\end{aligned}$$

$$\begin{aligned}
\alpha_1 &= \frac{1}{3}\beta_1 - \beta_2 + \beta_3 \\
\alpha_2 &= \frac{1}{3}\beta_1 - 2\beta_3 \\
\alpha_3 &= \frac{1}{3}\beta_1 + \beta_2 + \beta_3
\end{aligned}$$

$$\begin{aligned}
\beta_1 &= \alpha_1 + \alpha_2 + \alpha_3 \\
\beta_2 &= \frac{1}{2}(\alpha_3 - \alpha_1) \\
\beta_3 &= \frac{1}{6}\alpha_1 - \frac{1}{3}\alpha_2 + \frac{1}{6}\alpha_3
\end{aligned}$$

(d) **M4 : Somatotype WT2 + WT9 + WT18 + DW9**

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

```
##
## Call:
## lm(formula = Soma ~ WT2 + WT9 + WT18 + DW9, data = girls)
##
## 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
```

It indicates us that there is strong collinearity in the variables that `DW9` can be derived from `WT9` and `WT2` linearly. From the perspective of linear algebra, it cannot have a solution for one of the linear correlated variables.

Part 2: Reproduce Output

Data were collected on 97 men before radical prostatectomony and we take as response the log of prostate specific antigen (PSA) which was being proposed as a preoperative marker to predict the clinical stage of cancer. Eight other covariates were available for modeling log PSA: log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason), and percentage Gleason scores 4 or 5 (pgg45). Let Yi represent log PSA and xi = (xi1,…,xi8) denote the eight covariates for individual i, i = 1,…,n = 97.

(a) **Give interpretations for each of the parameters of the model.**

```
Prostate <- read.csv('Prostate.csv')
```

```
lmod <- lm(lpsa ~ ., data = Prostate)
summary(lmod)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = Prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73316 -0.37133 -0.01702  0.41414  1.63811
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669399   1.296381   0.516  0.60690
## lcavol       0.587023   0.087920   6.677 2.11e-09 ***
## lweight      0.454461   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766156   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045136   0.157464   0.287  0.77506
## pgg45        0.004525   0.004421   1.024  0.30885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

- A one unit increase in `lcavol` is associated with an increase of 0.587 in the expected value of `lpsa`, if all other covariates are held fixed.
- A one unit increase in `lweight` is associated with an increase of 0.454 in the expected value of `lpsa`, if all other covariates are held fixed.
- A one unit increase in `age` is associated with a decrease of 0.020 in the expected value of `lpsa`, if all other covariates are held fixed.
- A one unit increase in `lbph` is associated with an increase of 0.107 in the expected value of `lpsa`, if all other covariates are held fixed.
- A one unit increase in `svi` is associated with an increase of 0.766 in the expected value of `lpsa`, if all other covariates are held fixed.
- A one unit increase in `lcp` is associated with a decrease of 0.105 in the expected value of `lpsa`, if all other covariates are held fixed.
- A one unit increase in `gleason` is associated with an increase of 0.045 in the expected value of `lpsa`, if all other covariates are held fixed.
- A one unit increase in `pgg45` is associated with an increase of 0.005 in the expected value of `lpsa`, if all other covariates are held fixed.
- If all covariates are fixed as 0, the expected value of `lpsa` will be 0.669.

(b) **Using R, reproduce every number in the output using matrix and arithmetic operations. Look back through the lecture slides (all the formulas are in there!). You may not use lm to do any of this.**

```r
# Y
Y <- as.matrix(Prostate[, ncol(Prostate)])
# Design matrix
X <- as.matrix(Prostate[, -ncol(Prostate)])
X <- cbind(intercept = 1, X)
dim(X)
```

```
## [1] 97  9
```

```r
# estimated betas
Beta <- solve(t(X) %*% X) %*% t(X) %*% Y
```

```r
# fitted Y
fitted_values <- X %*% Beta
# residuals between Y and fitted Y
residuals <- Y - fitted_values
```

```
# Sum of squared residuals
SS_residual <- sum(residuals^2)
MSE <- SS_residual / (nrow(X) - ncol(X))
RSE <- sqrt(MSE)
```

```
# standard error of the coefficients
std_error <- t(sqrt(t(residuals) %*% residuals %*% diag(solve(t(X) %*% X)) / (nrow(X) - ncol(X))))
```

```
# t statistics of the coefficients
t_value <- c(Beta) / c(std_error)
```

```
# p-value of the coefficients' t-test
p_value <- 2*(1 - pt(abs(t_value), nrow(X) - ncol(X), lower.tail = T))
```

```
df_coefficients <- data.frame('Estimate' = round(Beta, 6), 'Std. Error' = round(std_error, 6), 't value
```

```
# Sum of squared total difference
SS_total <- sum((Y - mean(Y))^2)
# R squared
R_squared <- 1- SS_residual / SS_total
```

```
# adjusted R squared
adjusted_R <- 1 - (1 - R_squared) * (nrow(X) - 1) / (nrow(X) - ncol(X))
```

```
# Sum of squared regression deviation
SS_regression <- sum((fitted_values - mean(Y))^2)
#f_stat <- ((SS_total - SS_residual) / 1) / MSE
F_stat <- (SS_regression / (ncol(X)-1)) / (SS_residual / (nrow(X) - ncol(X)))
```

```
# p-value for the F-test
F_pvalue <- pf(F_stat, df1 = ncol(X)-1, df2 = nrow(X) - ncol(X), lower.tail = F)
```

```
paste('Residuals:', c(summary(residuals)))
```

```
## [1] "Residuals: Min.   :-1.73316  " "Residuals: 1st Qu.:-0.37133  "
## [3] "Residuals: Median :-0.01702  " "Residuals: Mean   : 0.00000  "
## [5] "Residuals: 3rd Qu.: 0.41414  " "Residuals: Max.   : 1.63811  "
```

```
print(df_coefficients)
```

```
##             Estimate Std..Error t.value           Pr
## intercept   0.669399   1.296381   0.516 6.068984e-01
## lcavol      0.587023   0.087920   6.677 2.110634e-09
## lweight     0.454461   0.170012   2.673 8.956206e-03
## age        -0.019637   0.011173  -1.758 8.229321e-02
## lbph        0.107054   0.058449   1.832 7.039819e-02
## svi         0.766156   0.244309   3.136 2.328823e-03
## lcp        -0.105474   0.091013  -1.159 2.496408e-01
## gleason     0.045136   0.157464   0.287 7.750601e-01
## pgg45       0.004525   0.004421   1.024 3.088513e-01
```

```r
paste('Residual standard error: ', round(RSE, 4), 'on', nrow(X) - ncol(X), 'degrees of freedom')
```

```
## [1] "Residual standard error:  0.7084 on 88 degrees of freedom"
```

```r
paste('Multiple R-squared', round(R_squared, 4))
```

```
## [1] "Multiple R-squared 0.6548"
```

```r
paste('Adjusted R-squared', round(adjusted_R, 4))
```

```
## [1] "Adjusted R-squared 0.6234"
```

```r
paste('F-statistic:', round(F_stat, 2), 'on', ncol(X)-1, 'and', nrow(X) - ncol(X), 'DF')
```

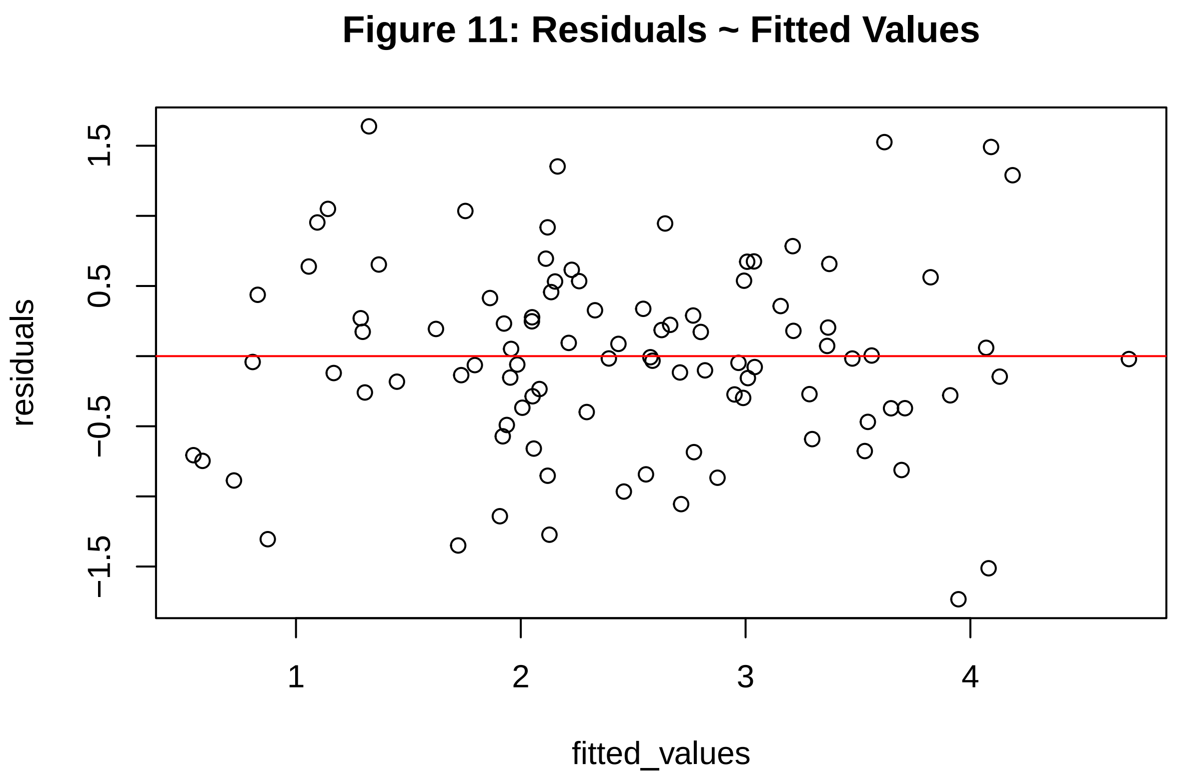
```
## [1] "F-statistic: 20.86 on 8 and 88 DF"
```

```r
paste('p-value:', F_pvalue)
```

```
## [1] "p-value: 2.24484836792811e-17"
```

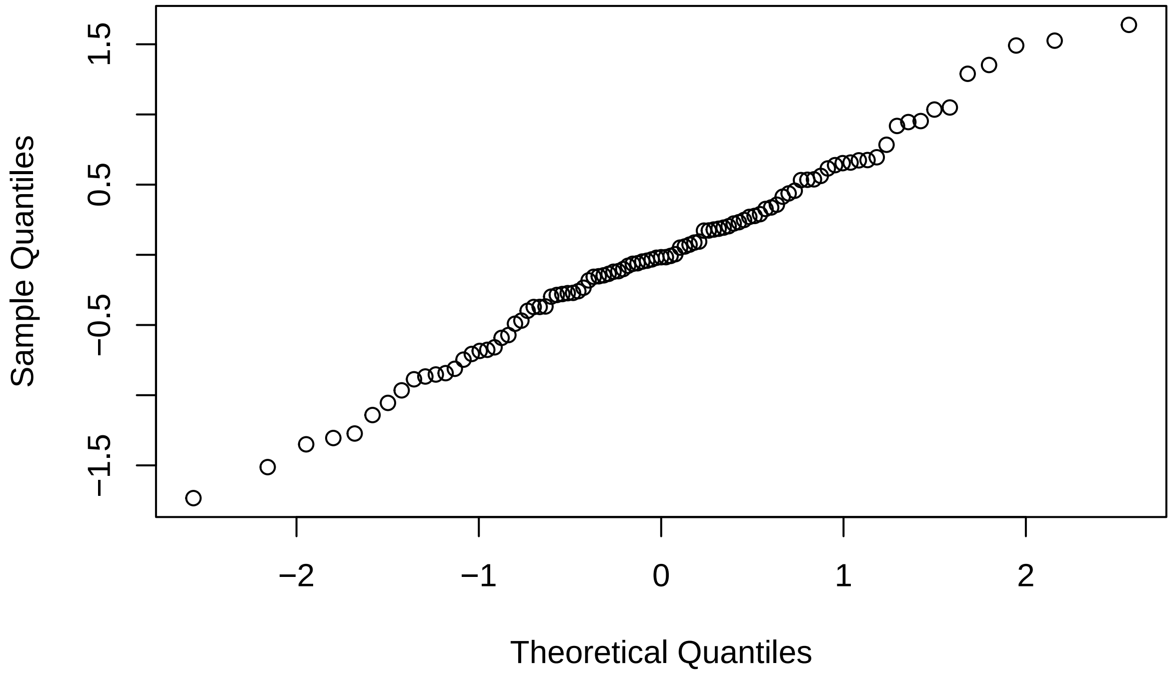
(c) **Create a plot the residuals from the full model against the fitted values and a Quantile-Quantile plot of the residuals. Use these two plots to comment on the plausibility of the modelling assumptions.**

```r
plot(residuals ~ fitted_values, main = 'Figure 11: Residuals ~ Fitted Values')
abline(h=0, col='red')
```



```
qqnorm(residuals, main = 'Figure 12: Quantile- Quantile plot of the residuals')
```

Figure 12: Quantile− Quantile plot of the residuals



From the first plot, we can see that the residuals have constant deviation along the increasing fitted values. So the residuals are likely to have the same variance and have a mean of 0.

From the second plot, we can tell that the residuals are aligned with the theoretical quantiles, but not perfectly. So the residuals are basically normally distributed.

Part 3: Model Selection Simulation Study

Stepwise model selection is a commonly used practice to attempt to select which predictors, out of a set of candidate predictors, should be included in a model. The stepwise algorithm considers the full model and removing subsequent terms from the model (and/or adding them back in) using AIC as the criteria for whether a single variable should be included in the model. The stepAIC function the MASS package runs a stepwise model selection procedure in this way. The goal of this section is to reproduce the following plot.

Proceed as follows:

- **Generate variables: X1 and X2 N (0, 1)**
- **Generate variable X3 which is a Normal(0,1) random variable, but is correlated with X1 at rho1 = 0.5.**
- **Generate variable X4 which is a Normal(0,1) random variable, but is correlated with X2 at rho2 = 0.7.**

```
generate_X <- function(n){
  X1 <- rnorm(n = n, 0, 1)
  X2 <- rnorm(n = n, 0, 1)
  X3 <- 0.5 * X1 / sd(X1) + sqrt(1-0.5^2) * rnorm(n = n)
  X4 <- 0.7 * X2 / sd(X2) + sqrt(1-0.7^2) * rnorm(n = n)
  return (list('X1'=X1, 'X2'=X2, 'X3'=X3, 'X4'=X4))
}
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
data <- list(data.frame(generate_X(100)), data.frame(generate_X(500)), data.frame(generate_X(1000)))
sigmas <- seq(0.1, 1, 0.1)
```

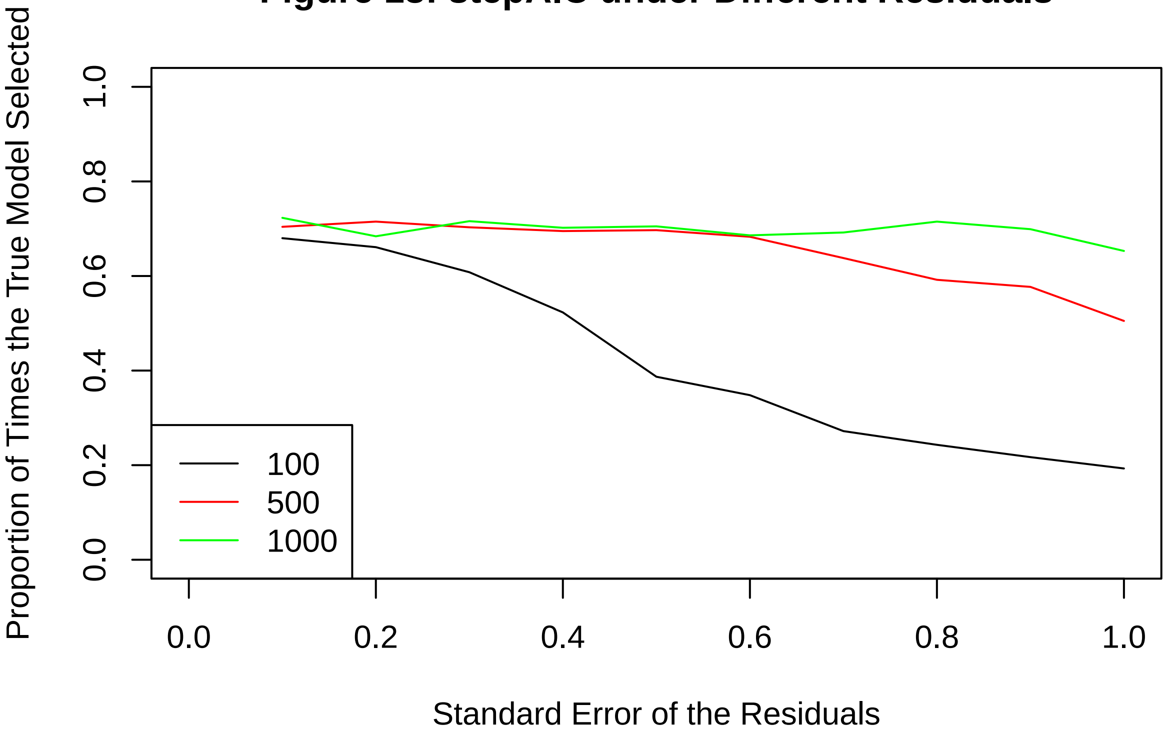
- **For each of 1,000 iterations, generate data from the true model Y = 4+3X1 −0.1X2 + where N(0, e2) and e takes on values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.**

```
model_selection <- function(iter = 1000, data){
  result <- data.frame()
  for (i in 1:iter){
    for (j in 1:length(sigmas)){
      Y = 4 + 3*data$X1 - 0.1*data$X2 + rnorm(nrow(data), 0, sigmas[j])
      lmod <- lm(Y ~ ., data = data)
      step <- stepAIC(lmod, trace = F)
      result[i, j] <- ifelse(step$call$formula == formula(Y~X1+X2), 1, 0)
    }
  }
  return (result)
}
```

```
proportion_100 <- colMeans(model_selection(data = data[[1]]))
proportion_500 <- colMeans(model_selection(data = data[[2]]))
proportion_1000 <- colMeans(model_selection(data = data[[3]]))
```

```
matplot(sigmas,
        cbind(proportion_100, proportion_500, proportion_1000),
        type = 'l',
        lty = 1,
        col = c("black", "red", "green"),
        xlim = c(0, 1),
        ylim = c(0, 1),
        xlab = "Standard Error of the Residuals",
        ylab = "Proportion of Times the True Model Selected",
        main = 'Figure 13: stepAIC under Different Residuals'
        )
legend("bottomleft", legend = c(100,500,1000), col=c("black", "red", "green"), lty=1)
```

Figure 13: stepAIC under Different Residuals



We start from the full model and search the model backwardly. However, the log likelihood term in AIC tends to be larger when the predictors are more. And the error of the model related to sample size. But the AIC formula does not adjust based on sample size. So in the simulation, we can see that the AIC tend to have higher proportion of times selecting the true model as the sample size increase.