Regression Project

Frequentist Inference

Liuyang Xu

03/20/2022

Part 1: Exploring Various Models

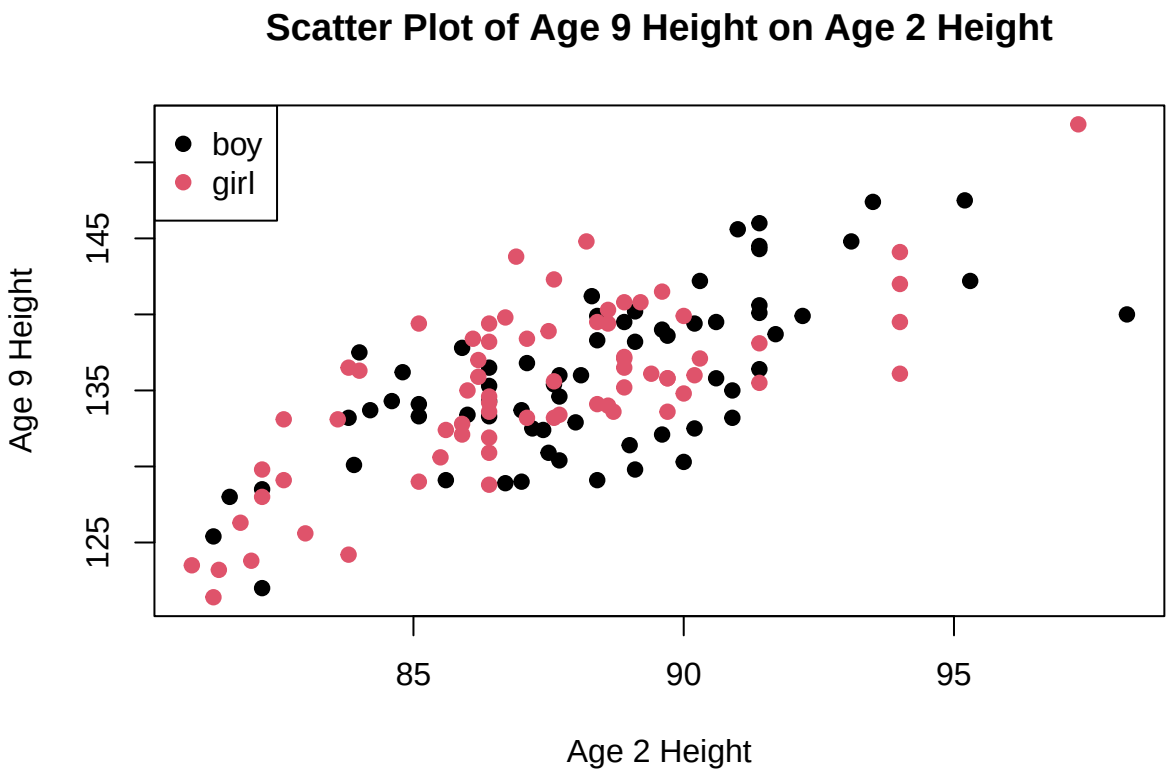
The Berkeley Guidance Study, under the direction of Jean Macfarlane, started with a sample of infants who were born in Berkeley, California in 1928-1929. Most of the children were Caucasian and Protestant, and two-thirds came from middle-class families. The basic cohort includes 136 of these children 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, 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).

```
library(tidyverse)
BGS <- read_csv("BGS.csv")
colnames(BGS)[1] <- "Index"
```

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?

```
# Scatter plot
plot(BGS$HT2, BGS$HT9,
     main = "Scatter Plot of Age 9 Height on Age 2 Height",
     xlab = "Age 2 Height", ylab = "Age 9 Height",
     pch = 19, col = factor(BGS$Sex))

# Legend
legend("topleft",
       legend = c("boy", "girl"),
       pch = 19,
       col = factor(levels(factor(BGS$Sex))))
```



Answer:

There doesn't appear to be a different pattern for boys than for girls.

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

```
fit_2_9 <- lm(HT9 ~ HT2, data = BGS)
summary.fit_2_9 <- summary(fit_2_9)
summary.fit_2_9
```

```
##
## 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.**

Answer:

The intercept is 31.92705, which means that the estimated average age 9 height is 31.92705cm for the children whose age 2 height is 0cm.

The coefficient on HT2 is 1.17963, which means that if we compare two groups of children whose age 2 height differs by 1cm, on average, the expected value of age 9 height over the two groups would differ by 1.17963cm.

- **Test the hypothesis of $H_0 : \beta_1 = 0$ against the two-sided alternative.**

Answer:

$H_0 : \beta_1 = 0$ is the null hypothesis. From the summary table above, if we set $\alpha = 0.05$, we can see that the p-value in the HT2 line is smaller than 0.05. This means that we can reject the hypothesis.

- **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.**

```
print(paste0("The square root of the F-statistic at the bottom of the regression output is ", round(sqr
```

```
## [1] "The square root of the F-statistic at the bottom of the regression output is 12.052."
```

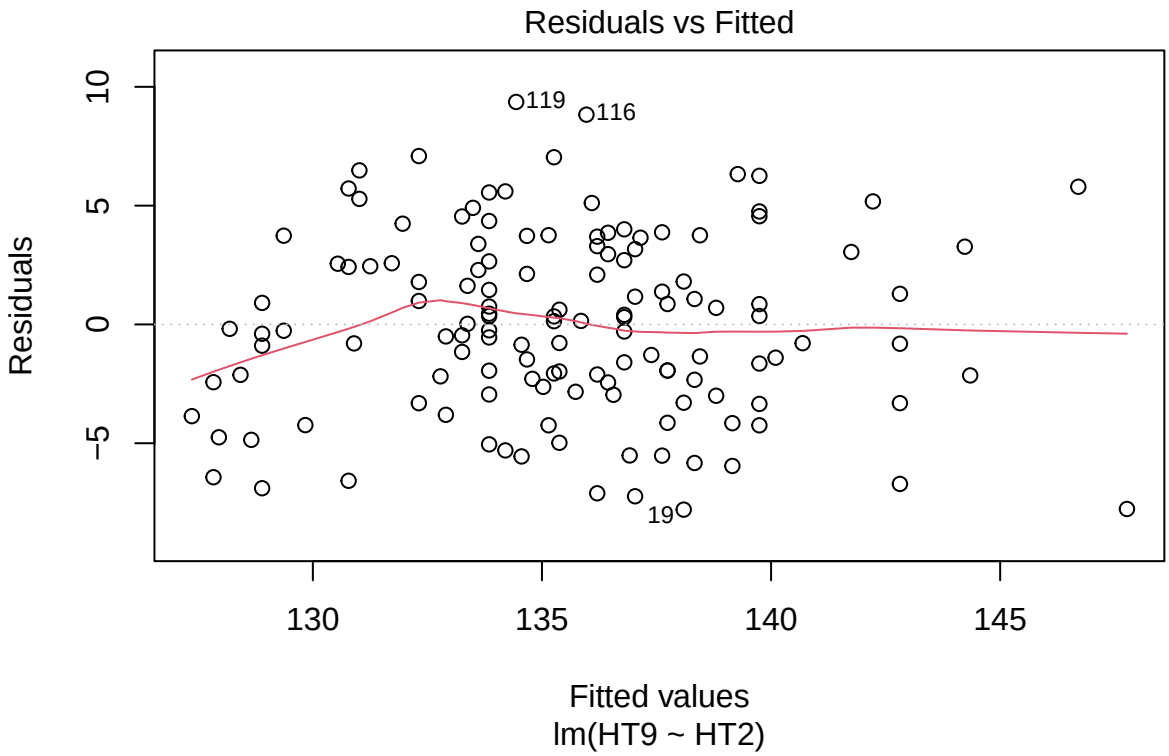
From the summary table above we can see that the T-statistic is also 12.052. So the value of the T-statistic for the above hypothesis test is equal to the square root of the F-statistic from the ANOVA at the bottom of the regression output.

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

Answer:

First, let's draw the residual plot.

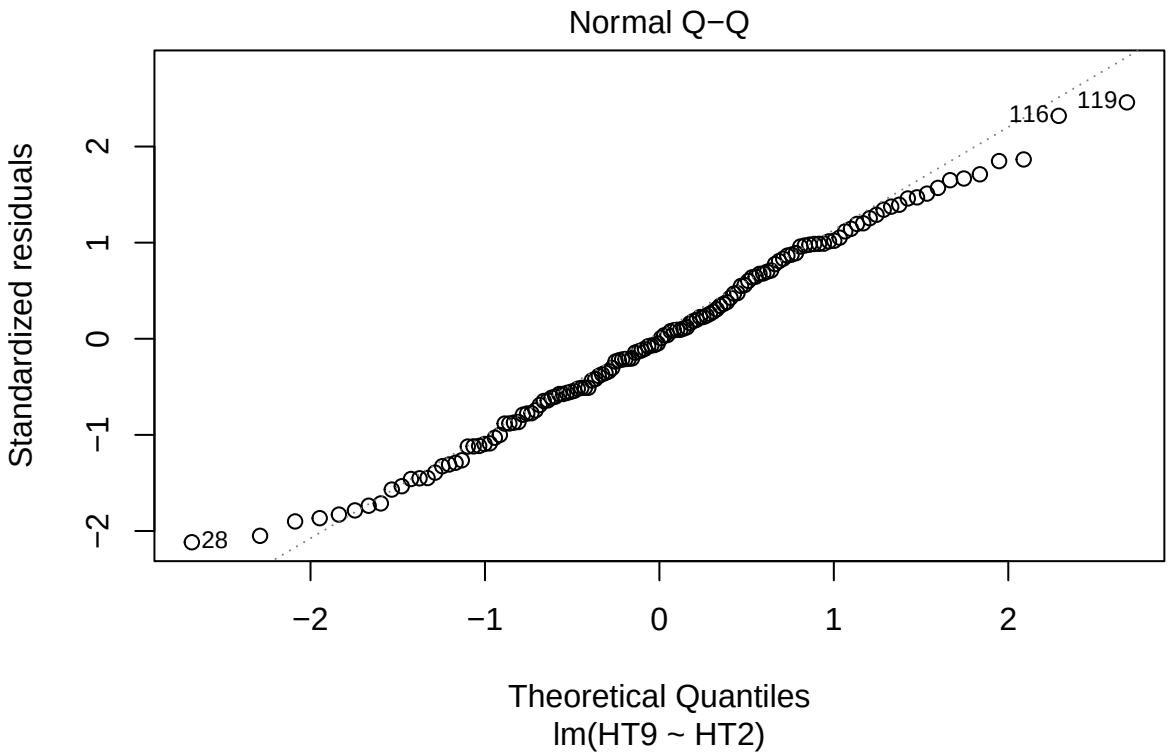
```
plot(fit_2_9, which = 1) # residual plot
```



From this plot, we can see that the residuals distribute around 0. Even though there are some differences when the fitted value is small, the residual points are roughly have the same variance. The red line in the plot also proves this conclusion. Thus the homoscedasticity assumption holds.

The plot also show that for any fixed fitted value, the distribution of the residuals seems to be normal. In order to verify this, let's draw the Q-Q plot.

```
plot(fit_2_9, which = 2) # Q-Q plot
```



From the plot we can tell that most of the points are on the $y = x$ line which indicates that the normality assumption also holds.

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

```
fit_2_9.si <- lm(HT9 ~ HT2 + as.factor(Sex), data = BGS)
summary.fit_2_9.si <- summary(fit_2_9.si)
summary.fit_2_9.si
```

```
##
## Call:
## lm(formula = HT9 ~ HT2 + as.factor(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 ***
## 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(fit_2_9, fit_2_9.si)
```

```
##            df      AIC
## fit_2_9     3 754.5998
## fit_2_9.si  4 755.8636
```

Answer:

We can see that the adjusted R^2 in this model(0.5156) is smaller than the adjusted R^2 in the simple linear regression model(0.5166).

The AIC for this model is 755.8636, which is larger than the AIC(754.5998) in the simple linear regression model.

Thus this model isn't better than the simple linear regression fit above.

(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?

```
fit_2_9.ss_si <- lm(HT9 ~ HT2*as.factor(Sex), data = BGS)
summary.fit_2_9.ss_si <- summary(fit_2_9.ss_si)
summary.fit_2_9.ss_si
```

```
##
## Call:
## lm(formula = HT9 ~ HT2 * as.factor(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 ***
## 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(fit_2_9, fit_2_9.ss_si)
```

```
##               df      AIC
## fit_2_9        3 754.5998
## fit_2_9.ss_si  5 757.5803
```

Answer:

We can see that the adjusted R^2 in this model(0.5129) is smaller than the adjusted R^2 in the simple linear regression model(0.5166).

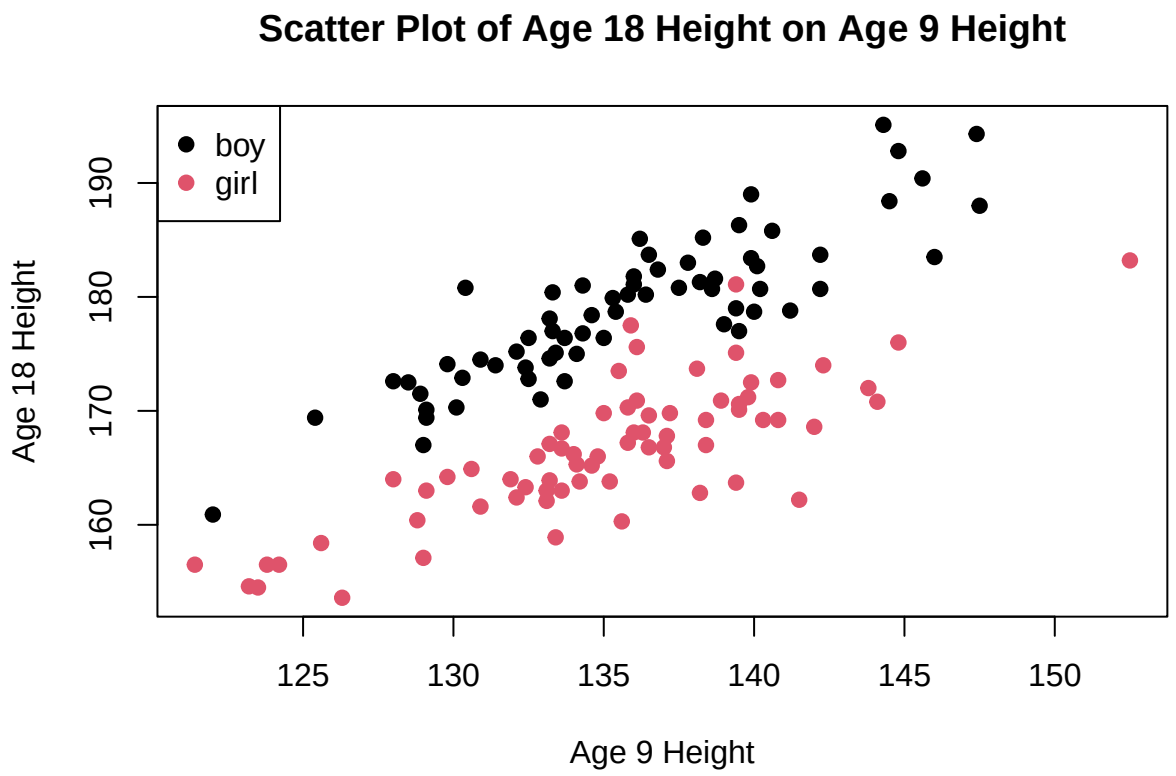
The AIC for this model is 757.5803, which is larger than the AIC(754.5998) in the simple linear regression model.

Thus this model isn't better than the simple linear regression fit above.

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?

```
# Scatter plot
plot(BGS$HT9, BGS$HT18,
     main = "Scatter Plot of Age 18 Height on Age 9 Height",
     xlab = "Age 9 Height", ylab = "Age 18 Height",
     pch = 19, col = factor(BGS$Sex))

# Legend
legend("topleft",
       legend = c("boy", "girl"),
       pch = 19,
       col = factor(levels(factor(BGS$Sex))))
```



Answer:

There does appear to be a different pattern for boys than for girls. Under the same age 9 height, the age 18 height of boys are generally larger than the age 18 height of girls.

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

```
fit_9_18 <- lm(HT18 ~ HT9, data = BGS)
summary.fit_9_18 <- summary(fit_9_18)
summary.fit_9_18
```

```
##
## 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
```

Answer:

The estimated intercept is 32.3416.

The estimated coefficient on HT2 is 1.0350.

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

```
fit_9_18.si <- lm(HT18 ~ HT9 + as.factor(Sex), data = BGS)
summary.fit_9_18.si <- summary(fit_9_18.si)
summary.fit_9_18.si
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + as.factor(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 ***
## 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(fit_9_18, fit_9_18.si)
```

```
##             df      AIC
## fit_9_18     3 911.2241
```

```
## fit_9_18.si  4 726.3621
```

Answer:

We can see that the adjusted R^2 in this model(0.8494) is larger than the adjusted R^2 in the simple linear regression model(0.4094).

The AIC for this model is 726.3621, which is smaller than the AIC(911.2241) in the simple linear regression model.

Thus this model is better than the simple linear regression fit above.

(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?

```
fit_9_18.ss_si <- lm(HT18 ~ HT9 * as.factor(Sex), data = BGS)
summary.fit_9_18.ss_si <- summary(fit_9_18.ss_si)
summary.fit_9_18.ss_si
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 * as.factor(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 ***
## 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
```

```
AIC(fit_9_18, fit_9_18.ss_si)
```

```
##                df      AIC
## fit_9_18        3 911.2241
## fit_9_18.ss_si  5 725.3423
```

Answer:

We can see that the adjusted R^2 in this model(0.8516) is larger than the adjusted R^2 in the simple linear regression model(0.4094).

The AIC for this model is 725.3423, which is smaller than the AIC(911.2241) in the simple linear regression model.

Thus this model is better than the simple linear regression fit above.

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

Answer:

We can see that the adjusted R^2 in the third model(0.8516) is larger than the adjusted R^2 in the rest two models. The AIC (725.3423) is also the smallest. Thus the model that allows for both the separate slope and separate intercepts for boys and for girls best describes the data.

The intercept is 35.07880.

The coefficient on HT9 is 1.05895, the coefficient on Sex is 13.32748, and the coefficient on HT9*Sex is -0.18463.

This means that we would predict an average Age 18 Height of 35.07880cm for the boys whose Age 9 Heights are 0cm, and we would predict an average Age 18 Height of $35.07880 + 13.32748 = 48.40628$cm for the girls whose Age 9 Heights are 0cm.

This also means that if we compare groups of boys whose Age 9 Heights differ in 1cm, the Age 18 Height of the boys who have higher Age 9 Height would be 1.05895cm higher, on average, than the Age 18 Height of the boys who have lower Age 9 Height.

If we compare groups of girls whose Age 9 Heights differ in 1cm, the Age 18 Height of the girls who have higher Age 9 Height would be $1.05895 + (-0.18463) = 0.87432$cm higher, on average, than the Age 18 Height of the girls who have lower Age 9 Height.

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.

```
BGS.boy <- BGS %>%
  filter(Sex == 0)

M1 <- lm(WT18 ~ WT9, data = BGS.boy)
M2 <- lm(WT18 ~ WT9 + LG9, data = BGS.boy)

summary(M1)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = BGS.boy)
##
## 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)
```

```
##
```

```
## Call:
## lm(formula = WT18 ~ WT9 + LG9, data = BGS.boy)
##
## 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
```

```
# check the correlation between the two predictor variables
cor(BGS.boy$WT9, BGS.boy$LG9)
```

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

The correlation between WT9 and LG9 is 0.9409453, which is close to 1. This indicates that these two variables have strong correlations with each other.

We can see from the data that weight at age 18 is related to weight at 9, so the weight at age 9 in the model M1 is significant. However, the leg circumference at age 9 is highly correlated with weight at age 9 in model M2. So when fitting M2, the weight at age 9 would not be significant.

(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. (You may need to do some research to answer this question).**

We know that $\hat{Y} = X\beta$ and $\beta = (X^TX)^{-1}X^TY$. So we can have $\hat{Y} = X(X^TX)^{-1}X^TY$. If we replace the $X(X^TX)^{-1}X^T$ as H, we can have $\hat{Y} = HY$. So this matrix is known as the hat matrix because it transforms Y to $\hat{Y}$. (https://www.sciencedirect.com/topics/mathematics/hat-matrix)

- **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.**

```
X.M1 <- BGS.boy %>%
  select(WT9)
X.M2 <- BGS.boy %>%
  select(WT9, LG9)

X.M1 <- cbind(1, X.M1)
X.M2 <- cbind(1, X.M2)

X.M1 <- as.matrix(X.M1)
X.M2 <- as.matrix(X.M2)

H.M1 <- X.M1 %*% solve(t(X.M1) %*% X.M1) %*% t(X.M1)
H.M2 <- X.M2 %*% solve(t(X.M2) %*% X.M2) %*% t(X.M2)

leverage.M1 <- diag(H.M1)
leverage.M2 <- diag(H.M2)
```

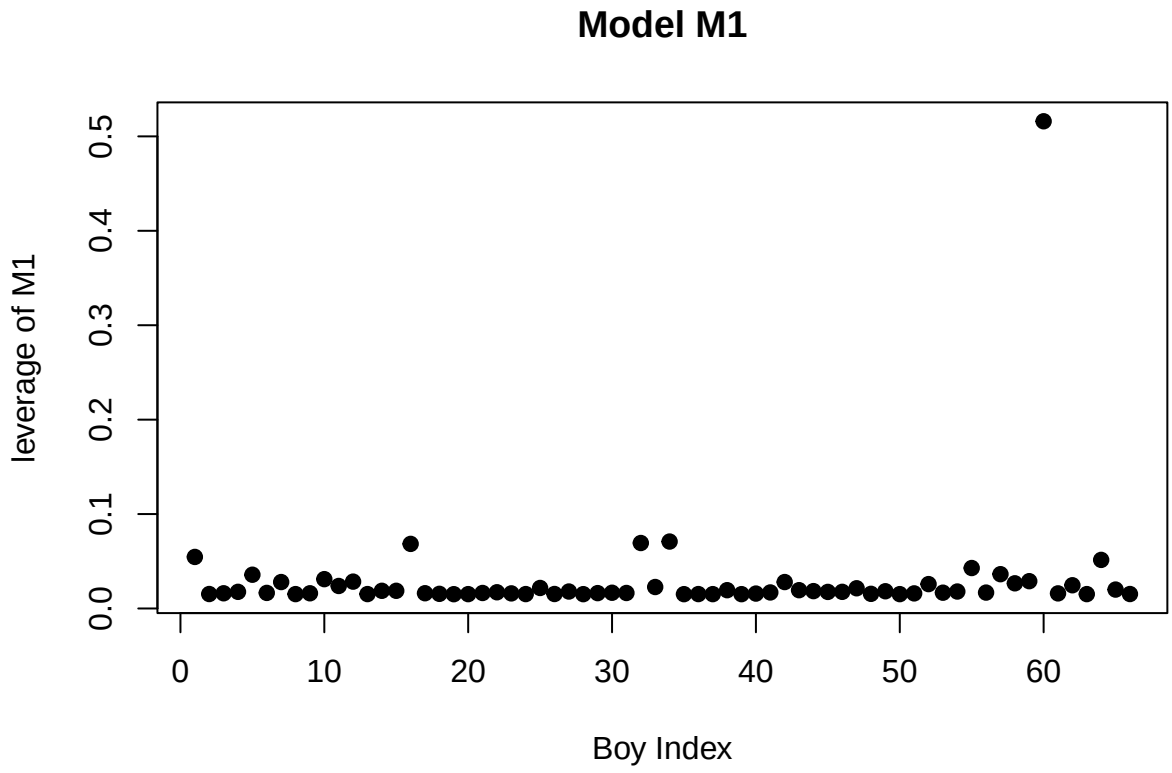
```
which.max(leverage.M1)
```

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

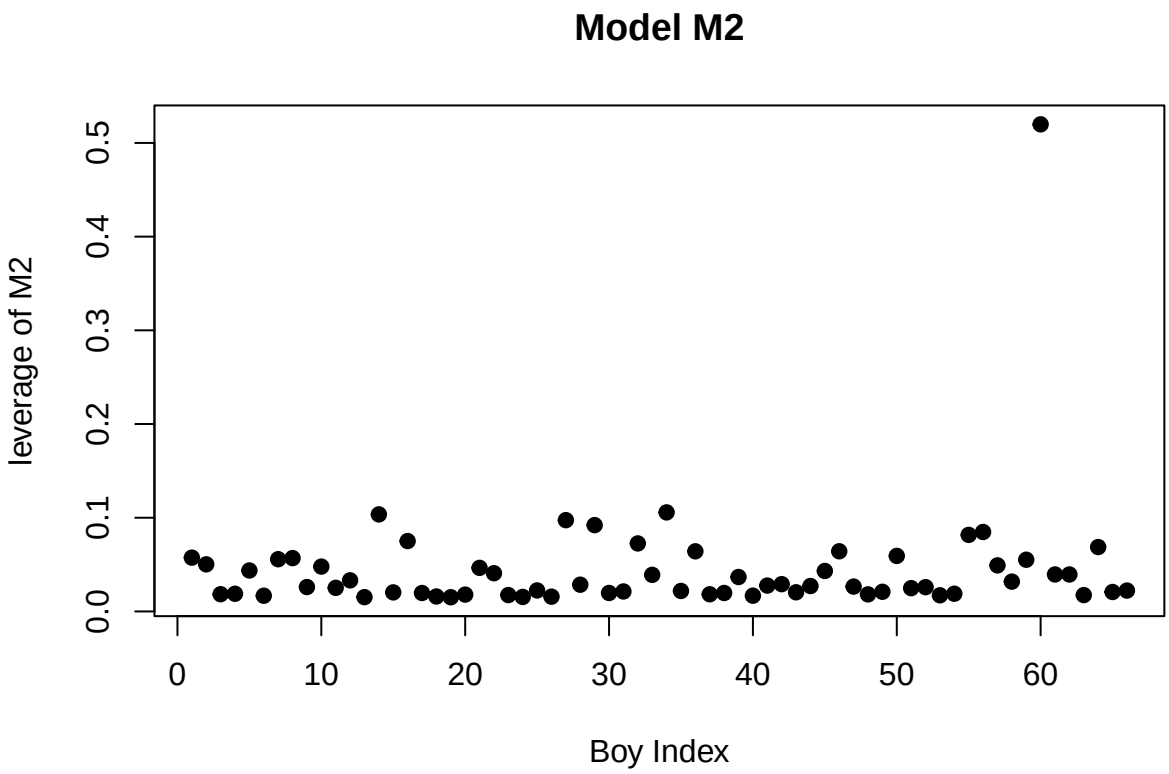
```
which.max(leverage.M2)
```

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

```
plot(leverage.M1, pch = 19, main = "Model M1",
     xlab = "Boy Index", ylab = "leverage of M1")
```




```
plot(leverage.M2, pch = 19, main = "Model M2",
     xlab = "Boy Index", ylab = "leverage of M2")
```



Answer:

From the plot we can observed that the leverage of the number 60 point in both plot is much higher than any of the other points.

- **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.**

```
fit1 <- lm(WT18 ~ WT9, data = BGS.boy)
summary(fit1)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = BGS.boy)
##
## 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
```

```
BGS.boy.no_hleverage <- BGS.boy %>%
  filter(Index != which.max(leverage.M1))
fit2 <- lm(WT18 ~ WT9, data = BGS.boy.no_hleverage)
summary(fit2)
```

```
##
## Call:
## lm(formula = WT18 ~ WT9, data = BGS.boy.no_hleverage)
##
## 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
```

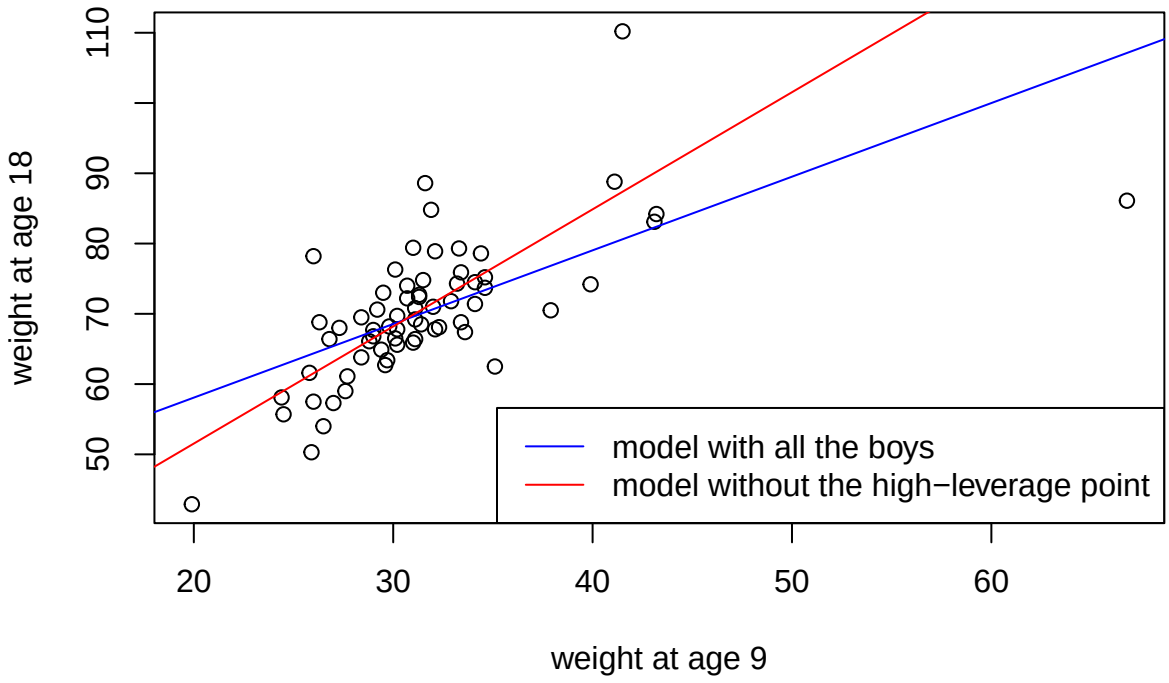
Answer:

The model "fit1" uses all of the boys in the dataset, "fit2" removes the high-leverage point.

The coefficient on Weight at Age 9 of the model "fit1" is 1.0481 is smaller than the coefficient on Weight at Age 9 of the model "fit2"(1.6667), and also has smaller standard error. Both coefficients are significant under $\alpha = 0.05$.

- **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(BGS.boy$WT9, BGS.boy$WT18,
     xlab = "weight at age 9", ylab = "weight at age 18")
abline(fit1, col = "blue")
abline(fit2, col = "red")
legend("bottomright",
       legend = c("model with all the boys","model without the high-leverage point"),
       lty = 1,
       col = c("blue", "red"))
```



- **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.**

I think the regression line from the model without the high-leverage point fitst the data better.

The intercept is 18.2029, which means that the estimated average weight at age 18 is 18.2029kg for the boy whose weight at age 9 is 0kg.

The coefficient on WT9 is 1.6667, which means that if we compare two groups of boys whose age 9 weight differs by 1kg, on average, the expected value of age 18 weight over the two groups would differ by 1.6667kg.

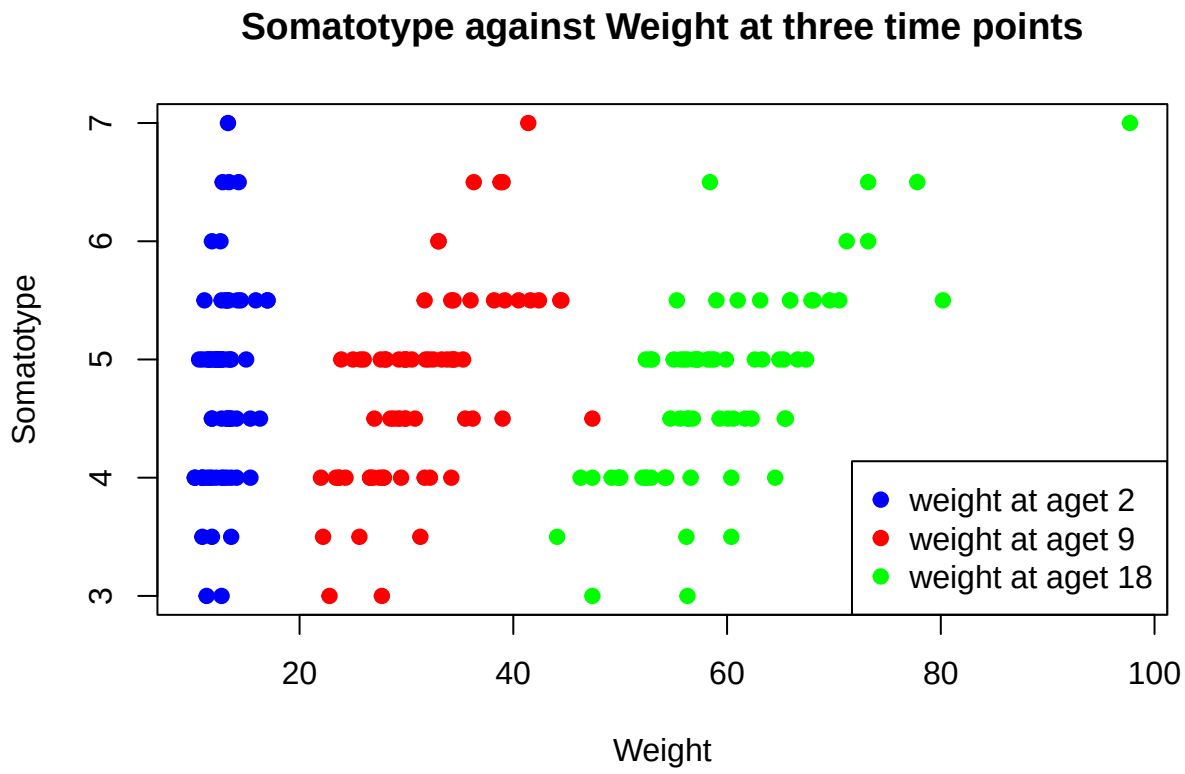
4. Create a new dataset that includes only the girls in the sample. Use this new dataset to model Somatotype in the following ways.

```
BGS.girl <- BGS %>%
  filter(Sex == 1)
```

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

```
# Plot somatotype against weight at each of the three time points
plot(BGS.girl$WT2, BGS.girl$Soma,
     xlim = c(min(c(BGS.girl$WT2, BGS.girl$WT9, BGS.girl$WT18)),
              max(c(BGS.girl$WT2, BGS.girl$WT9, BGS.girl$WT18))),
     main = "Somatotype against Weight at three time points",
     xlab = "Weight", ylab = "Somatotype",
     pch = 19, col = "blue")
points(BGS.girl$WT9, BGS.girl$Soma,
       pch = 19, col = "red")
```

```
points(BGS.girl$WT18, BGS.girl$Soma,
       pch = 19, col = "green")

# legend
legend("bottomright",
       legend = c("weight at aget 2","weight at aget 9", "weight at aget 18"),
       pch = 19,
       col = c("blue", "red", "green"))
```



Answer:

From the plot, we can see that as the age increases, with the same range of somatotype, the variation of the weight seems to be larger. And as the age increases, the larger weight is more likely related to a larger somatotype.

(b) Create new variables:

$$DW9 = WT9 - WT2$$

$$DW18 = WT18 - WT9$$

$$AVE = \frac{1}{3}(WT2 + WT9 + WT18)$$

$$LIN = WT18 - WT2$$

$$QUAD = WT2 - 2WT9 + WT18$$

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).

```
BGS.girl <- BGS.girl %>%
  mutate(DW9 = WT9 - WT2,
         DW18 = WT18 - WT9,
         AVE =(WT2 + WT9 + WT18)/3,
         LIN = WT18 - WT2,
         QUAD = WT2 - 2*WT9 + WT18)
```

(c) Fit the following three models:

$$M1 : Somatotype \sim WT2 + WT9 + WT18$$

$$M2 : Somatotype \sim WT2 + DW9 + DW18$$

$$M3 : Somatotype \sim AVE + LIN + QUAD$$

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

summary(M1)
```

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

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

```
##
## Call:
## lm(formula = Soma ~ AVE + LIN + QUAD, data = BGS.girl)
##
## 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 intercept of the models are the same across all three models. The other coefficients are different.

- **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?**

Let the regression result of M1 be

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

and let the regression result of M2 be

$$Somatotype = \beta_0^{'} + \beta_1^{'} WT2 + \beta_2^{'} DW9 + \beta_3^{'} DW18$$

The variable DW9 and DW18 are created based on WT2, WT9, and WT18 so we can transform the regression result of M2 to

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

Comparing the equations between the two models, we can see that they are essentially doing regression on the same variables. Then, we have

$$\beta_0 = \beta_0^{'}$$

$$\beta_1 = \beta_1^{'} - \beta_2^{'}$$
$$\beta_2 = \beta_2^{'} - \beta_3^{'}$$
$$\beta_3 = \beta_3^{'}$$

Since $\beta_3^{'} \neq 0$, the coefficient for DW18 in model 2 equals the coefficient for WT18 in model 1, and the coefficient for DW9 in model 2 does not equal the coefficient for WT9 in model 1.

- **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.**

Let the regression result of M1 be

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

and let the regression result of M3 be

$$Somatotype = \beta_0^{''} + \beta_1^{''} AVE + \beta_2^{''} LIN + \beta_3^{''} QUAD$$

The variables AVE, LIN, and QUAD are all created based on WT2, WT9, and WT18. So we can transform the regression result of M3 to

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

From the equation, we can tell that both models are essentially doing regression on the same variables. And we have

$$\beta_0 = \beta_0^{''}$$
$$\beta_1 = \frac{1}{3}\beta_1^{''} - \beta_2^{''} + \beta_3^{''}$$
$$\beta_2 = \frac{1}{3}\beta_1^{''} - 2\beta_3^{''}$$
$$\beta_3 = \frac{1}{3}\beta_1^{''} + \beta_2^{''} + \beta_3^{''}$$

So the coefficients in M3 can be obtained by algebraically manipulating the coefficients in M1 and thus M1 and M3 are equivalent.

(d) Fit the following model:

$$M4 : Somatotype \sim WT2 + WT9 + WT18 + DW9$$

Explain why some parameters are not estimated.

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

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

```
# check the correlation between the two predictor variables
cor(BGS.girl$WT9, BGS.girl$DW9)
```

```
## [1] 0.9757681
```

Answer:

The correlation between WT9 and DW9 is 0.9757681 which is close to 1. So they are collinear and this collinearity would make R unable to estimate DW9.

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 Y_i represent log PSA and $x_i = (x_{i1}, ..., x_{i8})$ denote the eight covariates for individual i, $i = 1, ..., n = 97$.

The freely available `R` software was used to fit the model

$$y_i = \beta_0 + \sum_{j=1}^{8} \beta_j x_{ij} + \epsilon_i, i = 1, ..., n$$

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

Answer:

The intercept is 0.669399, which means that the estimated average log PSA is 0.669399 for the men whose all 8 covariates are 0.

The coefficient on lcavol is 0.587023, which means that if we compare two groups of men whose other 7 covariates are the same but log cancer volume differs by 1 point, on average, the expected value of log PSA over the two groups would differ by 0.587023.

The coefficient on lweight is 0.454461, which means that if we compare two groups of men whose other 7 covariates are the same but log prostate weight differs by 1 point, on average, the expected value of log PSA over the two groups would differ by 0.454461.

The coefficient on age is -0.019637, which means that if we compare two groups of men whose other 7 covariates are the same but age differs by 1 year, on average, the expected value of log PSA over the two groups would differ by -0.019637.

The coefficient on lbph is 0.107054, which means that if we compare two groups of men whose other 7 covariates are the same but log benign prostatic hyperplasia amount differs by 1 point, on average, the expected value of log PSA over the two groups would differ by 0.107054.

The coefficient on svi is 0.766156, which means that if we compare two groups of men whose other 7 covariates are the same but seminal vesicle invasion differs by 1 point, on average, the expected value of log PSA over the two groups would differ by 0.766156.

The coefficient on lcp is -0.105474, which means that if we compare two groups of men whose other 7 covariates are the same but log capsular penetration differs by 1 point, on average, the expected value of log PSA over the two groups would differ by -0.105474.

The coefficient on gleason is 0.045136, which means that if we compare two groups of men whose other 7 covariates are the same but Gleason score differs by 1 point, on average, the expected value of log PSA over the two groups would differ by 0.045136.

The coefficient on pgg45 is 0.004525, which means that if we compare two groups of men whose other 7 covariates are the same but percentage Gleason scores 4 or 5 differs by 1 point, on average, the expected value of log PSA over the two groups would differ by 0.004525.

(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.

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

Y <- as.matrix(Prostate$lpsa)

X <- as.matrix(cbind(1, Prostate[,1:(ncol(Prostate)-1)]))

n <- nrow(X)
p <- ncol(X)-1

# Estimates
Beta.Est <- (solve(t(X) %*% X)) %*% t(X) %*% Y

# Residuals
epsilon <- Y - X %*% Beta.Est

# Std. Error
Beta.SE <- as.matrix(sqrt(c((t(epsilon) %*% epsilon)/(n-(p+1))) * diag((solve(t(X) %*% X)))))

t.value <- (Beta.Est - 0)/Beta.SE # t value
t.p <- pt(abs(t.value), n-(p+1), lower.tail = FALSE) * 2 # Pr(>|t|)
t.p[2] <- signif(t.p[2], 9)

RSE <- sqrt((t(epsilon) %*% epsilon)/(n-(p+1))) # residual standard error

DoF <- (n-(p+1)) # degrees of freedom

RSS <- sum(epsilon^2)
SYY <- sum((Y - mean(Y))^2)
R.squared <- 1 - RSS/SYY # Multiple R-squared
R.squared.adjusted <- R.squared - (1-R.squared)*p/(n-(p+1)) # Adjusted R-squared

SSreg <- SYY - RSS
MSreg <- SSreg/p
sigma.squared.hat <- RSS/(n-(p+1))
```

```
F.statistic <- MSreg/sigma.squared.hat # F-statistic
F.p <- pf(F.statistic, p, n-(p+1), lower.tail = FALSE)
```

The summary of the residuals is shown below which presents the "Min", "1Q", "Median", "3Q" and "Max" from the original output.

```
summary(epsilon)
```

```
##        V1
##  Min.   :-1.73316
##  1st Qu.:-0.37133
##  Median :-0.01702
##  Mean   : 0.00000
##  3rd Qu.: 0.41414
##  Max.   : 1.63811
```

The coefficients are shown in the table below.

```
coefficient <- cbind(Beta.Est, Beta.SE, t.value, t.p)
colnames(coefficient) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")
rownames(coefficient)[1] <- "(Intercept)"

# make sure the p-value of lcavol isn't presented as 0.00000
coefficient[2, 4] <- format(coefficient[2, 4], digits = 3)

knitr::kable(coefficient)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.669399027184521	1.29638127733067	0.516359684369135	0.606898363237784
lcavol	0.587022880773449	0.0879203738428021	6.67675596811065	2.11e-09
lweight	0.454460640790026	0.170012070924301	2.67310808179248	0.00895620581170239
age	-0.0196372076738313	0.0111727430862198	-1.75759950106177	0.0822932121190488
lbph	0.107054351135309	0.0584493315638548	1.83157528531108	0.0703981907229473
svi	0.766155884609281	0.244309491854207	3.13600539542888	0.00232882271454564
lcp	-0.10547356953901	0.0910134842618407	-1.15887849360397	0.249640824286659
gleason	0.0451359643599669	0.157464466864357	0.28664222004351	0.775060071644532
pgg45	0.00452532362022757	0.00442118469364313	1.02355452979247	0.308851251292271

The residual standard error output is shown below.

```
paste0("Residual standard error: ", round(RSE,4), " on ", DoF, " degrees of freedom")
```

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

The R-squared output is shown below.

```
paste0("Multiple R-squared: ", round(R.squared, 4),
       ",   Adjusted R-squared: ", round(R.squared.adjusted, 4))
```

```
## [1] "Multiple R-squared: 0.6548,   Adjusted R-squared: 0.6234"
```

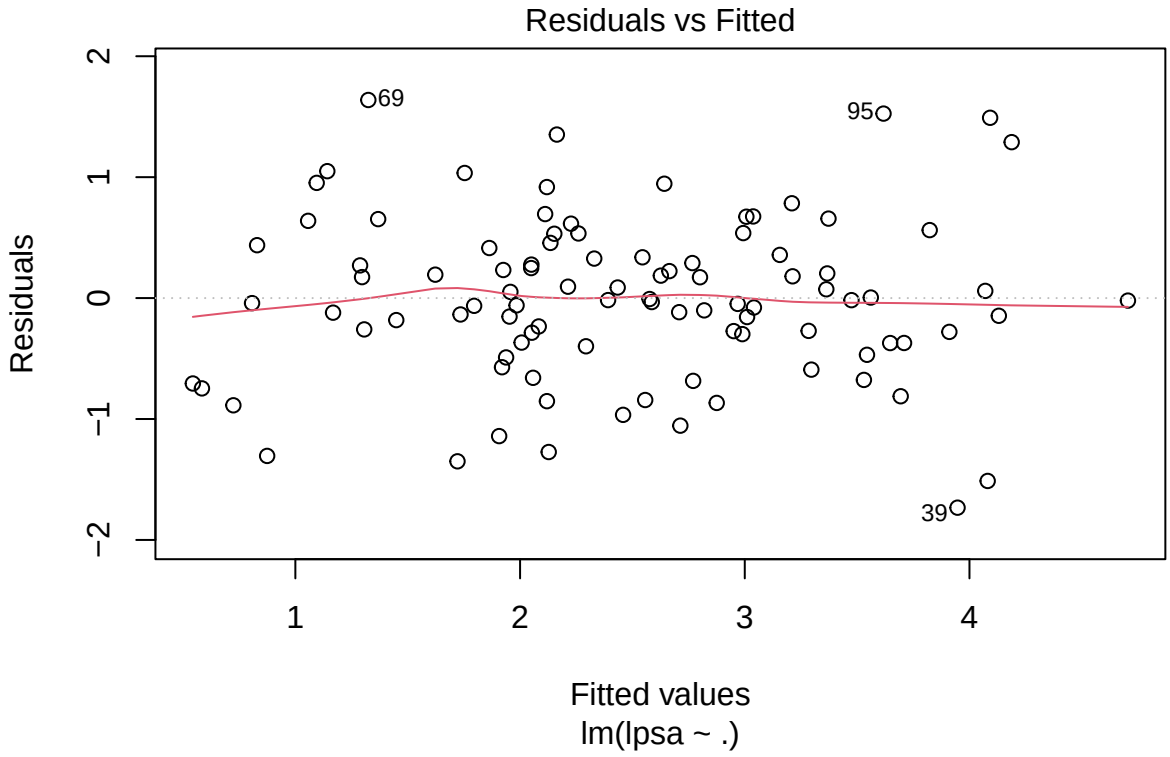
The F-statistic output is shown below.

```
paste0("F-statistic: ", round(F.statistic, 2),
       " on ", p, " and ", n-(p+1), " DF, p-value: ", F.p)
```

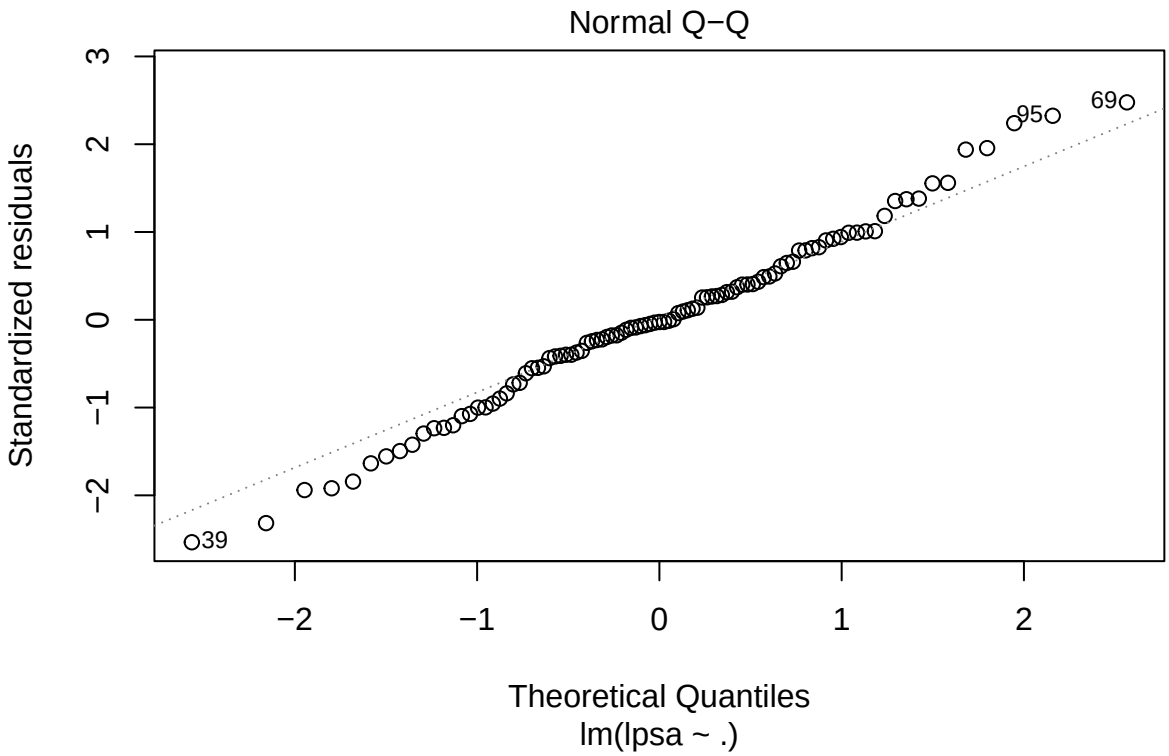
```
## [1] "F-statistic: 20.86 on 8 and 88 DF, p-value: 2.24484836792899e-17"
```

(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.

```
fit <- lm(formula = lpsa ~ ., data = Prostate)

plot(fit, which = 1)
```




```
plot(fit, which = 2)
```



Answer:

From the residuals against the fitted values plot, we can see that the residuals distribute around 0 and the residual points are roughly have the same variance. The red line in the plot also proves this conclusion. Thus the homoscedasticity assumption is plausible.

From the Q-Q plot we can tell that most of the points are on the $y = x$ line (when the standardized residuals is around [-2,-1] and [1,3], the points are not on the line but still close to the line) which indicates that the normality assumption is basically plausible.

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 plot.

Proceed as follows: * Generate variables: X_1 and $X_2 \sim N(0, 1)$

- Generate variable X_3 which is a Normal(0,1) random variable, but is correlated with X_1 at $\rho_1 = 0.5$.
- Generate variable X_4 which is a Normal(0,1) random variable, but is correlated with X_2 at $\rho_2 = 0.7$. – To generate a variable W that is correlated with X at ρ, you may use the equation:

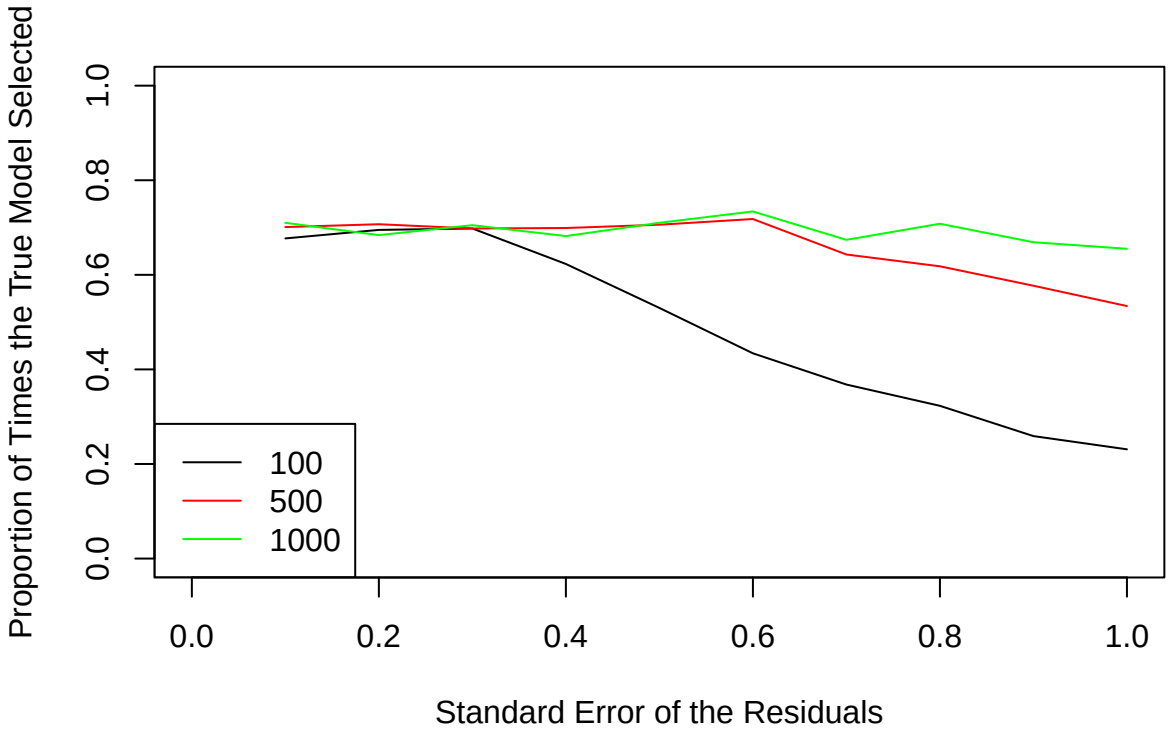
$$W = \frac{\rho X}{sd(X)} + \sqrt{1 - \rho^2} rnorm(0, 1)$$

- For each of 1,000 iterations, generate data from the true model $Y = 4 + 3X_1 - 0.1X_2 + \epsilon$ where $\epsilon \sim N(0, \sigma_e^2)$ and σ_e takes on values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.
- Note that you can generate X_1, X_2, X_3, X_4 once (outside your iteration for loop), but ϵ and Y should be generated for each iteration.
- For each iteration and for each value of σ_e, run the `stepAIC` procedure starting with the full model $lm(Y \sim X1 + X2 + X3 + X4)$ and check whether the output of `stepAIC` (the model selected) is the true model. You can use the default settings of `stepAIC`. You may want to use $call to do check whether the true model is selected.
- Finally, calculate the proportion of time across your 1,000 iterations for which the true model was selected for n = 100, 500, 1000. Plot them to reproduce the above plot.
- Write in complete sentences the moral of the story. What have you learned about the `stepAIC` process.

```
library("MASS")
```

```
n <- c(100, 500, 1000)
Itr <- 1000
sigma_e <- c(1:10)/10

generate_variables <- function(n){
  rho1 <- 0.5
  rho2 <- 0.7
  X_1 <- rnorm(n, mean = 0, sd = 1)
  X_2 <- rnorm(n, mean = 0, sd = 1)
  X_3 <- rho1*X_1/sd(X_1)+sqrt(1-rho1^2)*rnorm(n, mean = 0, sd = 1)
  X_4 <- rho2*X_2/sd(X_2)+sqrt(1-rho2^2)*rnorm(n, mean = 0, sd = 1)
  return(cbind(X_1, X_2, X_3, X_4))
}

true_model_prop <- matrix(rep(NA, length(n)*length(sigma_e)),
                          nrow = length(sigma_e), ncol = length(n))
colnames(true_model_prop) <- n
rownames(true_model_prop) <- sigma_e

for (i in 1:length(n)){
  X <- generate_variables(n[i])

  for (j in 1:length(sigma_e)){

    if_true_model <- rep(NA, Itr)

    for (k in 1:Itr){
      epsilon <- rnorm(n[i], mean = 0, sd = sigma_e[j])
      Y <- 4 + 3*X[,1] - 0.1*X[,2] + epsilon
      model <- lm(Y ~ X[,1] + X[,2] + X[,3] + X[,4])
      AIC_r <- stepAIC(model, trace = FALSE)
      if_true_model[k] <- AIC_r$call == "lm(formula = Y ~ X[, 1] + X[, 2])"
    }

    true_model_prop[j, i] <- sum(if_true_model)/Itr
  }
}

# draw the plot
```

```r
plot(c(0,1), c(0, 1), type = "n",
     xlab = "Standard Error of the Residuals",
     ylab = "Proportion of Times the True Model Selected")
lines(sigma_e, true_model_prop[,1], col = "black")
lines(sigma_e, true_model_prop[,2], col = "red")
lines(sigma_e, true_model_prop[,3], col = "green")
# Legend
legend("bottomleft",
       legend = n,
       lty = 1,
       col = c("black", "red", "green"))
```



For each n and for each value of σ_e, I ran the stepAIC procedure 1,000 times and drew the plot of proportion of times the true model $Y \sim X_1 + X_2$ was selected.

From the stepAIC process, I learned that:

1. For a fixed n(sample size), with larger standard error of the residuals, the stepAIC process would generally get lower proportion of times the true model selected.
2. For a fixed standard error of the residuals, when the standard error of the residuals is small(around 0.2 in this case), the proportion of times the true model selected for different sample size would roughly be the same. But with larger standard error of the residuals, stepAIC process would generally get higher proportion of times the true model selected when the sample size is larger.