Yuhan Zhu

STT863 Homework 3

2.23. Refer to **Grade point average** Problem 1.19.

a. Set up the ANOVA table.

b. What is estimated by *MSR* in your ANOVA table? by *MSE*? Under what condition do *MSR* and *MSE* estimate the same quantity?

c. Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.

d. What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model? What is the relative reduction? What is the name of the latter measure?

e. Obtain r and attach the appropriate sign.

f. Which measure, R^2 or r, has the more clear-cut operational interpretation? Explain.

a.

```
```{r 2.23.a}
# Read the data file
data <- read.table("CH01PR19.txt", header=FALSE, col.names=c("GPA", "ACT"))

# Linear regression
model <- lm(GPA ~ ACT, data=data)

# ANOVA table
anova_table <- anova(model)
anova_table
```
```

```
Analysis of Variance Table

Response: GPA
           Df Sum Sq Mean Sq F value   Pr(>F)
ACT         1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$SS = \Sigma(gpa - \overline{gpa})^2$ $\quad SSR = \Sigma(\widehat{gpa} - \overline{gpa})^2$ $\quad SSE = SS - SSR$

$MSR = \frac{SSR}{1}$ $\quad MSE = \frac{SSE}{n-1-1}$ $\quad F\_value = \frac{MSR}{MSE}$

b. MSR (Mean Square Regression) $= \frac{SSR}{1} = 3.588$

MSE (Mean Square Error (residual) $= \frac{SSE}{(n-p-1)} = 0.3883$

When $MSR = MSE$

$$SSR = \frac{SSE}{n-p-1}$$

$$\Sigma(\hat{Y} - \bar{Y})^2 = \frac{\Sigma(Y_i - \bar{Y})^2 - \Sigma(\hat{Y} - \bar{Y})^2}{c}$$

MSR and MSE estimate the same quantity when the slope $\beta_1 = 0$, means there's no relationship between X and Y. Simply use the $\bar{Y}$ does the same job.

c. $H_0: \beta_1 = 0$ (slope = 0, no relationship between GPA & ACT)

$H_a: \beta_1 \neq 0$ (regression Model works)

If F value > F critical value then reject H_0

```{r 2.23.c}
alpha <- 0.01
if (anova_table$`Pr(>F)`[1] < alpha) {
  print("Reject H0")
} else {
  print("Fail to reject H0")
}
```

```
[1] "Reject H0"
```

d. The absolute reduction in this case is $SSR = 3.588$

The relative reduction is $R^2 = \frac{SSR}{SST} = 0.07262$

```{r 2.23.d}
summary_model <- summary(model)
R2 <- summary_model$r.squared
R2
```

```
[1] 0.07262044
```

R^2 the coefficient of determination represent "the proportion the the variance in the dependent variable that is predictable from the independent variable"

e. $r = \beta_1 \times \sqrt{R^2}$

or we can simply do this

```
```{r 2.23.e}
cor(data$GPA,data$ACT)
```
```

```
[1] 0.2694818
```

f. I think R^2 offers a clearer interpretation for the regression model pepective.

2.26. Refer to **Plastic hardness** Problem 1.22.

a. Set up the ANOVA table.

b. Test by means of an F test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. Plot the deviations $Y_i - \hat{Y}_i$ against X_i on a graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph, using the same scales as for the first graph. From your two graphs, does SSE or SSR appear to be the larger component of $SSTO$? What does this imply about the magnitude of R^2?

d. Calculate R^2 and r.

a.
```
```{r 2.26.a}
data <- read.table("CH01PR22.txt", header=FALSE, col.names=c("Hardness", "Time"))

model <- lm(Hardness ~ Time, data=data)

# Display the ANOVA table
anova_table <- anova(model)
print(anova_table)
```
```

```
Analysis of Variance Table

Response: Hardness
          Df Sum Sq Mean Sq F value    Pr(>F)    
Time       1 5297.5  5297.5  506.51 2.159e-12 ***
Residuals 14  146.4    10.5                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

b. $H_0: \beta_1 = 0$ there's no linear relationship

$H_1: \beta_1 \neq 0$ there is linear relationship

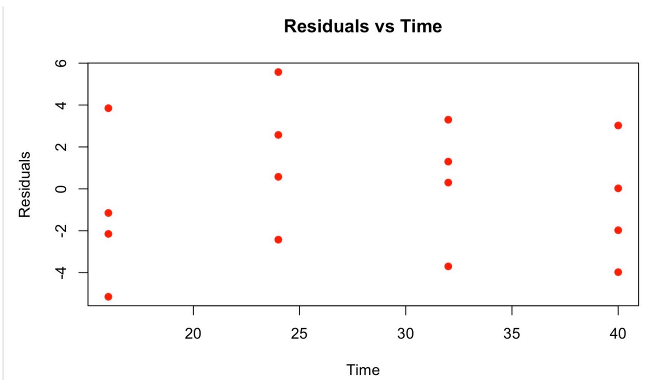
If F value > critical F value for $\alpha = 0.01$ then we reject H_0.

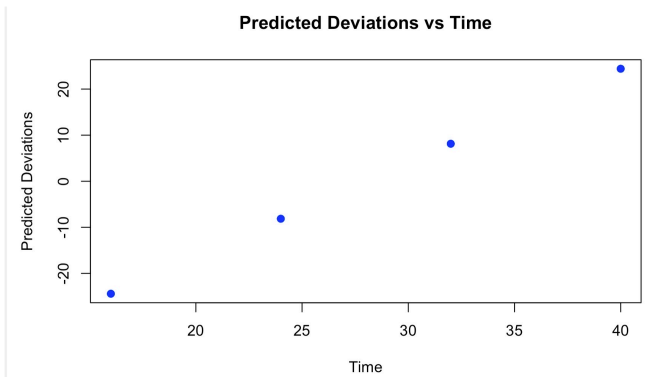
$F_{critical} \approx 8.8615$

```{r 2.26.b}
alpha <- 0.01

# Display the p-value
p_value <- anova_table$`Pr(>F)`[1]
print(paste("P-value:", p_value))

# Decision rule
if (p_value < alpha) {
  print("Reject H0: There is a linear association between hardness and elapsed time.")
} else {
  print("Fail to reject H0: There is no linear association between hardness and elapsed time.")
}
```

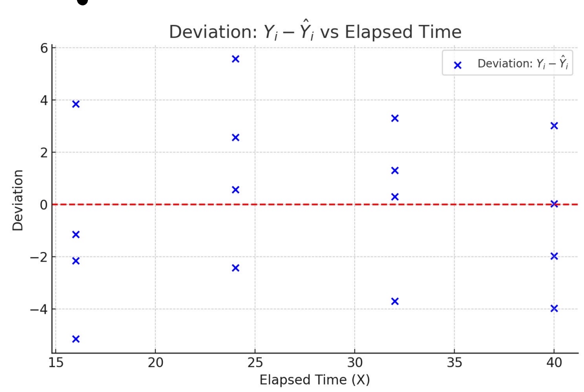
```
[1] "P-value: 2.15881368252505e-12"
[1] "Reject H0: There is a linear association between hardness and elapsed time."
```

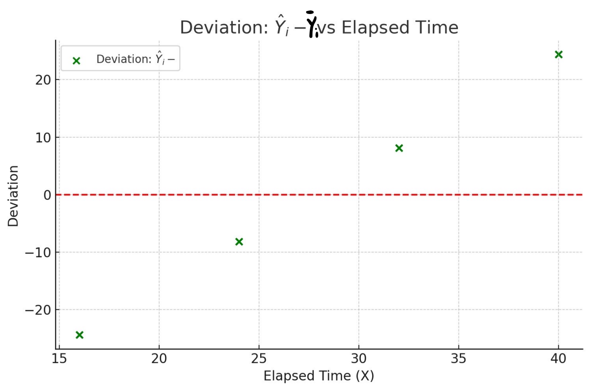
c.





From Python





SSTO (Total Sum of Squares) = SSR + SSE = 5297.5 + 146.4 = 5443.9

SSR is larger component of SSTO. This implies regression model explained significant amount of the total variation in the observed value. And this suggest we will get a high R^2 value.

d.

```{r 2.26.d}
summary_model <- summary(model)

R_squared <- summary_model$r.squared
r <- cor(data$Hardness, data$Time)

R_squared
r
```

```
[1] 0.9731031
[1] 0.9864599
```

2.57. The normal error regression model (2.1) is assumed to be applicable.

a. When testing H_0: $\beta_1 = 5$ versus H_a: $\beta_1 \neq 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

b. When testing H_0: $\beta_0 = 2$, $\beta_1 = 5$ versus H_a: not both $\beta_0 = 2$ and $\beta_1 = 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

a. $H_0: \beta_1 = 5$ $H_a: \beta_1 \neq 5$ slope of regression line equal or not equal to 5

Reduced model: $Y = \beta_0 + 5X + \varepsilon$ β_1 is constant 5

$df_R = n - 1$ since we only need to consider β_0 in the reduced model.

n: number of observations.

b. $H_0: \beta_0 = 2, \beta_1 = 5$ H_a: not both $\beta_0 = 2$ and $\beta_1 = 5$

Reduced model $Y = 2 + 5X + \varepsilon$

$df_R = n$ since we need to consider both β_0 and β_1

2.61. Show that the ratio $SSR/SSTO$ is the same whether Y_1 is regressed on Y_2 or Y_2 is regressed on Y_1. [*Hint*: Use (1.10a) and (2.51).]

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_1 - X)^2}$$

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

set $X = Y_2$ and $Y = Y_1$

Let Y_1 regressed on Y_2:

$$b_1 = \frac{\sum (Y_2 - \bar{Y}_2)(Y_1 - \bar{Y}_1)}{\sum (Y_2 - \bar{Y}_2)^2}$$

$$SSR = b_1^2 \sum (Y_2 - \bar{Y}_2)^2 \qquad \frac{SSR}{SSTO} = \frac{b_1^2 \sum (Y_2 - \bar{Y}_2)^2}{\sum (Y_1 - \bar{Y}_1)^2}$$

$$SSTO = \sum (Y_1 - \bar{Y}_1)^2$$

same for Y_2 regressed on Y_1

$$b_1 = \frac{\sum (Y_1 - \bar{Y}_1)(Y_2 - \bar{Y}_2)}{\sum (Y_1 - \bar{Y}_1)^2}$$

$$SSR = b_1^2 \sum (Y_1 - \bar{Y}_1)^2 \qquad \frac{SSR}{SSTO} = \frac{b_1^2 \sum (Y_1 - \bar{Y}_1)^2}{\sum (Y_2 - \bar{Y}_2)^2}$$

$$SSTO = \sum (Y_2 - \bar{Y}_2)^2$$

If we plug in b_1 knowing $\sum (Y_2 - \bar{Y}_2)(Y_1 - \bar{Y}_1) = \sum (Y_1 - \bar{Y}_1)(Y_2 - \bar{Y}_2)$

$$\Rightarrow \frac{\left(\sum (Y_2 - \bar{Y}_2)(Y_1 - \bar{Y}_1)\right)^2}{\sum (Y_1 - \bar{Y}_1)^2 \sum (Y_2 - \bar{Y}_2)^2} = \frac{\left(\sum (Y_1 - \bar{Y}_1)(Y_2 - Y_2 - \bar{Y}_2)\right)^2}{\sum (Y_1 - \bar{Y}_1)^2 \sum (Y_2 - \bar{Y}_2)^2}$$

$\therefore \frac{SSR}{SSTO}$ ratio is the same wheather Y_1 regress on Y_2 or Y_2 regress on Y_1