THEORETICAL PORTION

PROBLEM #1.

a. $Y = \beta_0 + \beta_{1M}X_{1M} + \beta_{1F}X_{1F} + \beta_{2N}X_{2N} + \beta_{2T}X_{2T} + \beta_{2F}X_{2F} + \beta_{2S}X_{2S} + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \epsilon$

With $\beta_{1M/F}$ & $\beta_{2N/T/F/S}$ representing the categorical covariates and the remaining as the continuous covariates. (Though β_0 will remain constant unless fitted differently)

b. β_0 = intercept
β_{1i} = Gender; where β_{1M} = Male and β_{1F} = Female
β_{2j} = Amount of alcohol per week; where β_{2N} = None, β_{2T} = 1-2 Drinks, β_{2F} = 3-5 Drinks, and β_{2S} = 6+ Drinks
β_3 = Amount of Fish consumed: average, in oz.
β_4 = Amount of Red meat consumed: average, in oz.
β_5 = Weight, in kg

c. Design Matrix X:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 6 & 0 & 93 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 12 & 9 & 70 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 8 & 3 & 83 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 7 & 77 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 3 & 1 & 71 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 4 & 4 & 84 \end{bmatrix}$$

d. Regression Model:

$$\begin{bmatrix} 102 \\ 95 \\ 92 \\ 113 \\ 132 \\ 148 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 6 & 0 & 93 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 12 & 9 & 70 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 8 & 3 & 83 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 7 & 77 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 3 & 1 & 71 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 4 & 4 & 84 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{1M} \\ \beta_{1F} \\ \beta_{2N} \\ \beta_{2T} \\ \beta_{2F} \\ \beta_{2S} \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

PROBLEM #2. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i1} * x_{i2} + \epsilon_i$

a. β_3 is added to the model to exponential create an additional effect on the model, with a higher significance as age is increased. The response variable increase in relation to the quadratic associated with β_3 must play a important role with the response time in terms of an exponentially increase with age. Since x_{i1} is centered around the average age, the further away from the central age, the greater the overall value $\beta_3 x_{i1}$.
b. It is assumed that β_4 positive value while β_2 is a negative value, which are both factored only when x_{i2} is 1. If the task has already been done before, then it must be that it takes less time to complete the task while also offsetting by a value of β_4. I believe it makes greater sense to write the equation as $x_{i2}(\beta_2 x_{i2} + \beta_4)$, since both only apply if the subject has completed the task previously.
c. β_0 is the base value when all x's = 0. Though this is an unlikely case since age would be zero. I'm sure there is a threshold at which age is first determined since a young age could produce a longer response time as well as an elderly age. However, since the average age is centered around the average age, then this should correct the issue.

   β_1 - an one unit increase in x_{i1} would increment β_1 by that amount dependent on the age.

   β_2 - this parameter is a set value, which only is included when the task has been completed before; whether it be negative or positive though suggestive negative since it should ideally decrease the response time if the task has been completed before.

   β_3 - a parameter that will exponentially be affected as age increases - or if age is centered around the mean of the age that has the lowest overall response time.

   β_4 - similar to β_2, this parameter will only factor when the task has been completed before and may be used to offset β_2 since age is a factor.

PROBLEM #4.

a. For 1 inch increase (β_1), beginning with 62 inches = 5.48*62 = 117.28 pounds. Every inch above this value will increase the weight by 5.48.
b. For each 1 inch increase in height above 62 inches, the value simply goes up form the y-intercept (β_0) by a weight increase of 5.48 from. When the x-value is 0, it's reflecting that the height is 62. This is a simple adjustment from the original predicted y-value.
c. Subtracting the mean height value from given x-value will reflect the average y-value when the x-value is zero. Thus, when the x-value is zero, the average weight is 150.57 pounds. Otherwise, the weight value as the x-value is increased or decreased by 1 will show the amount for which the person's weight is away from the average weight.
d. The third model is optimal since the y-intercept reflects the average weight and would show a better fit.

COMPUTATIONAL PORTION

Code and graphs included with each part, not indexed.

PROBLEM #1 - COMPUTATIONAL

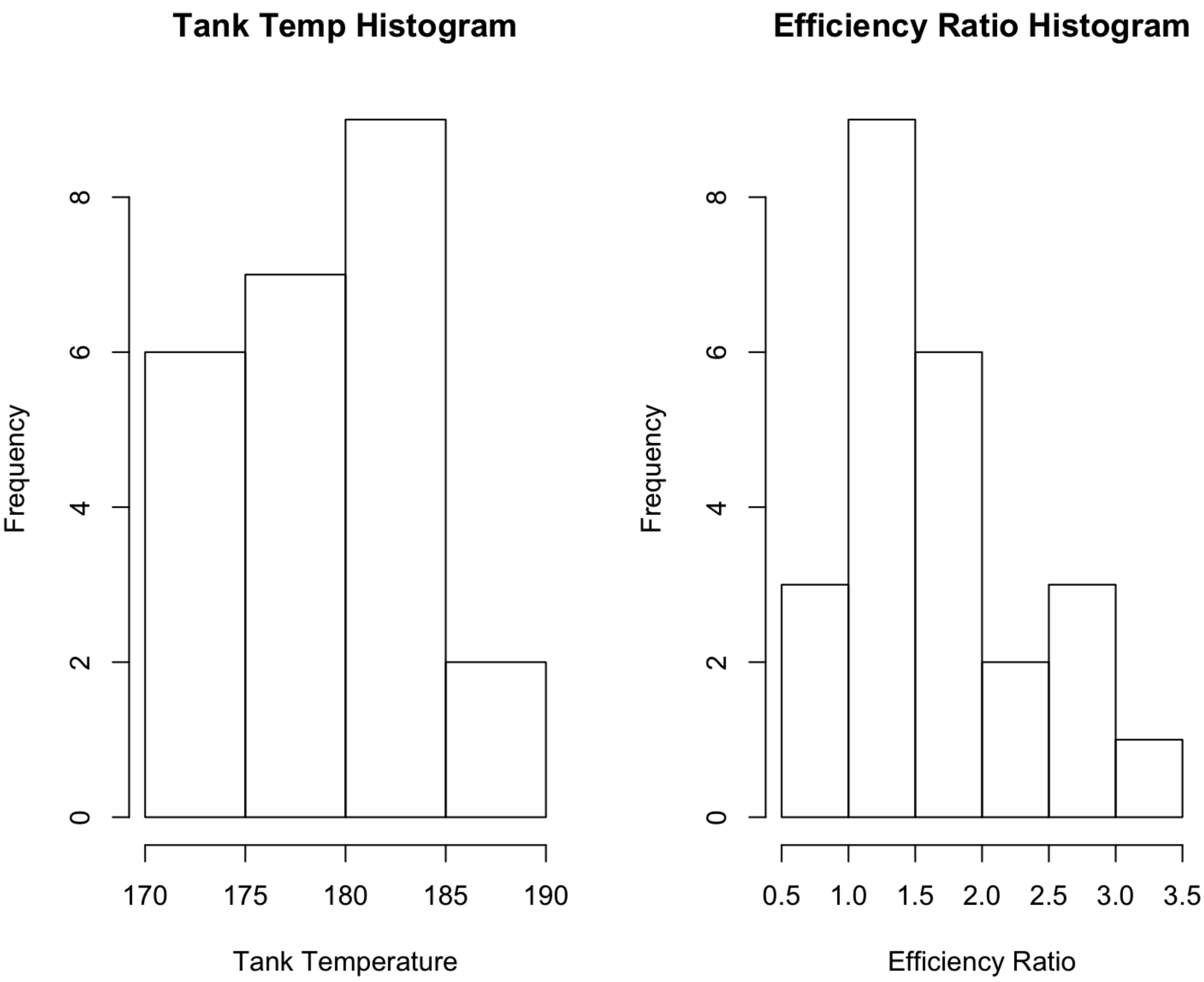
#1, part *a.*

code:

```
steeldata = read.table("steeldata.txt",header=TRUE)

par(mfrow=c(1,2))
hist(steeldata[,1], main="Tank Temp Histogram", xlab="Tank Temperature") #TankTemp
hist(steeldata[,2], main="Efficiency Ratio Histoogram", xlab="Efficiency Ratio") #EfficiencyRatio
```

output:



Both of these appear normally distributed about their mean value.
(TankTemp mean = 179.5, Efficiency Ratio mean = 1.6704)

#1, part b.

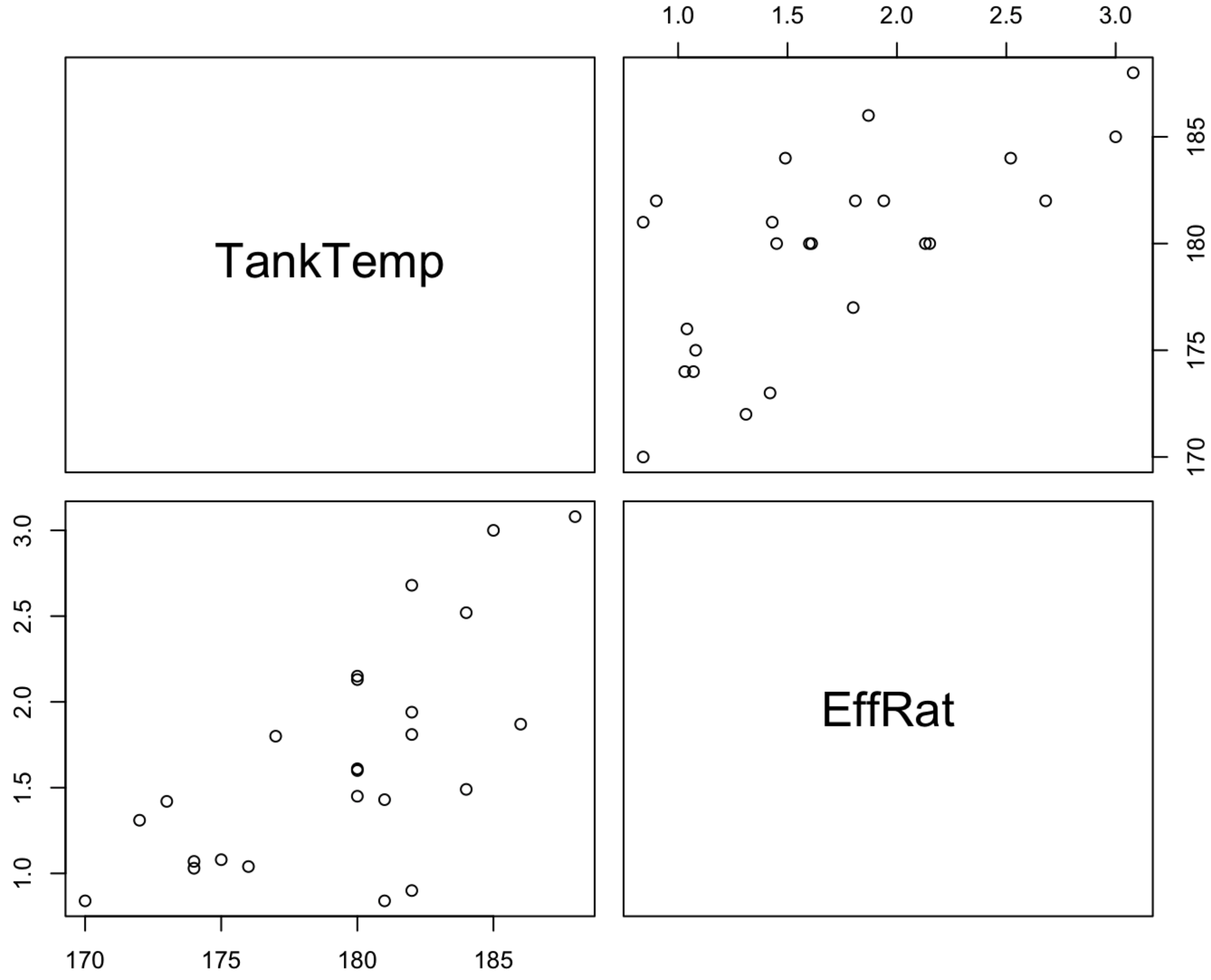
code:

```
cor(steeldata)
pairs(steeldata)

steel1=lm(EffRat~TankTemp, steeldata)
summary(steel1)
```

output:

```
            TankTemp    EffRat
TankTemp 1.0000000 0.6718615
EffRat     0.6718615 1.0000000
```



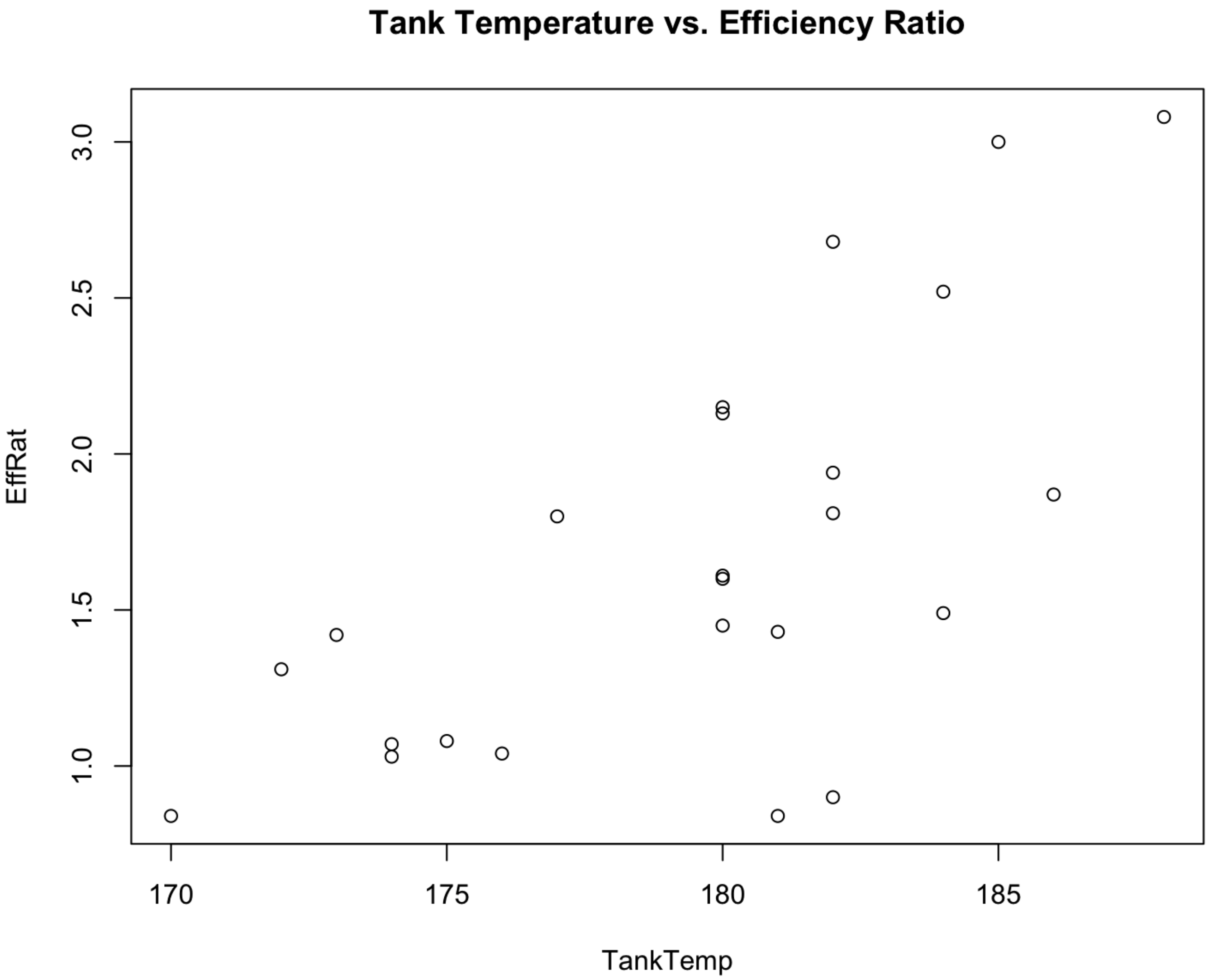
There does not appear to a complete and unique relationship between the efficiency ratio and the tank temperature for the data on steel. However, there is some type of linear relationship between the two.

#1, part c.

code:

```
plot(steeldata, main="Tank Temperature vs. Efficiency Ratio")
```

output:



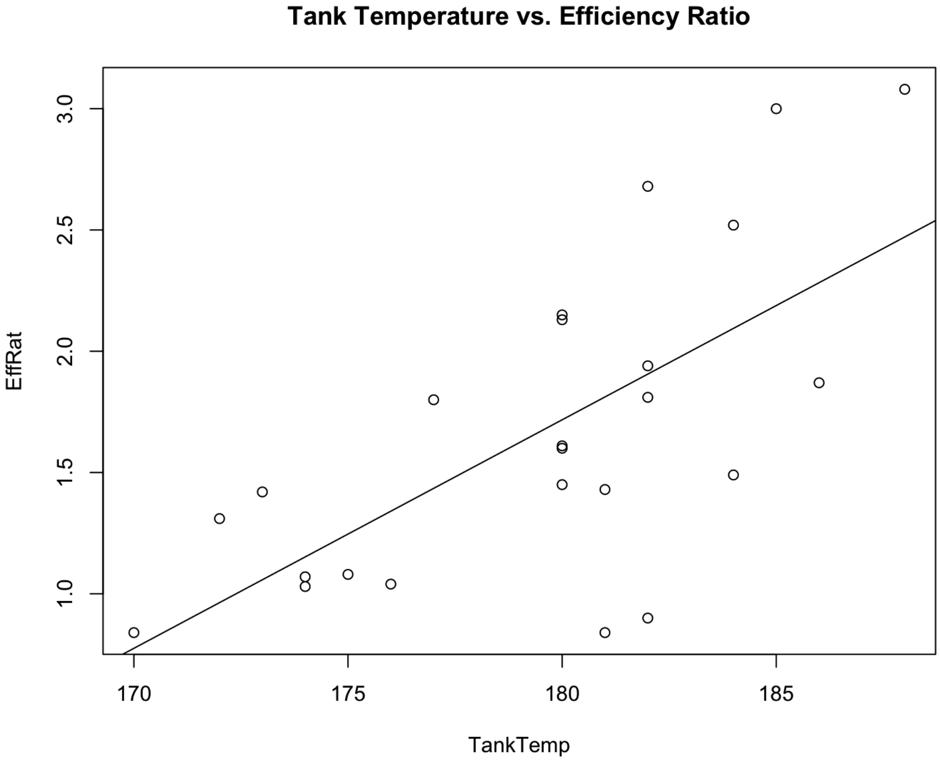
There appears to be a linear relationship between the efficiency ratio and the value of temperature. The efficiency ratio could be predicted by the value of the temperature but it may not be very accurate.

#1, part d(i).

code:

```
plot(steeldata, main="Tank Temperature vs. Efficiency Ratio");abline(steel1)
summary(steel1)
```

output:



```
> summary(steel1)

Call:
lm(formula = EffRat ~ TankTemp, data = steeldata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00601 -0.27580 -0.08906  0.37700  0.81128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.24497    3.97705  -3.833 0.000905 ***
TankTemp      0.09424    0.02215   4.255 0.000324 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4972 on 22 degrees of freedom
Multiple R-squared:  0.4514,	Adjusted R-squared:  0.4265
F-statistic:  18.1 on 1 and 22 DF,  p-value: 0.0003239
```

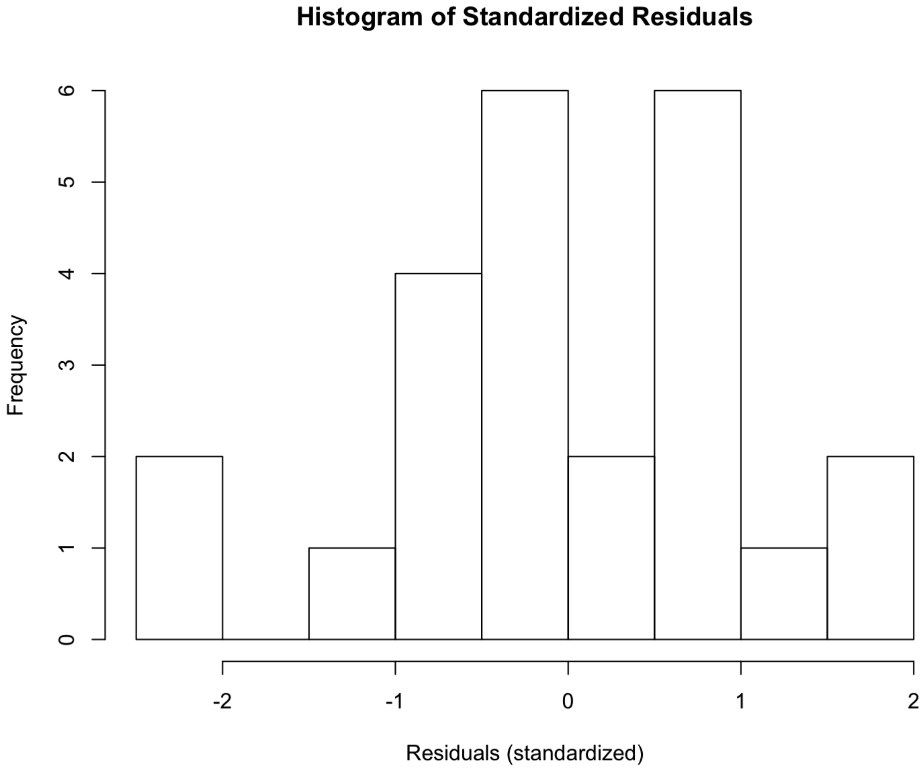
Equation for estimated regression line: $\hat{Y} = -15.245 + 0.0942x$

#1, part d(ii).

code:

```
hist(rstandard(steel1), main="Histogram of Standardized Residuals", xlab="Residuals (standardized)")
```

output:



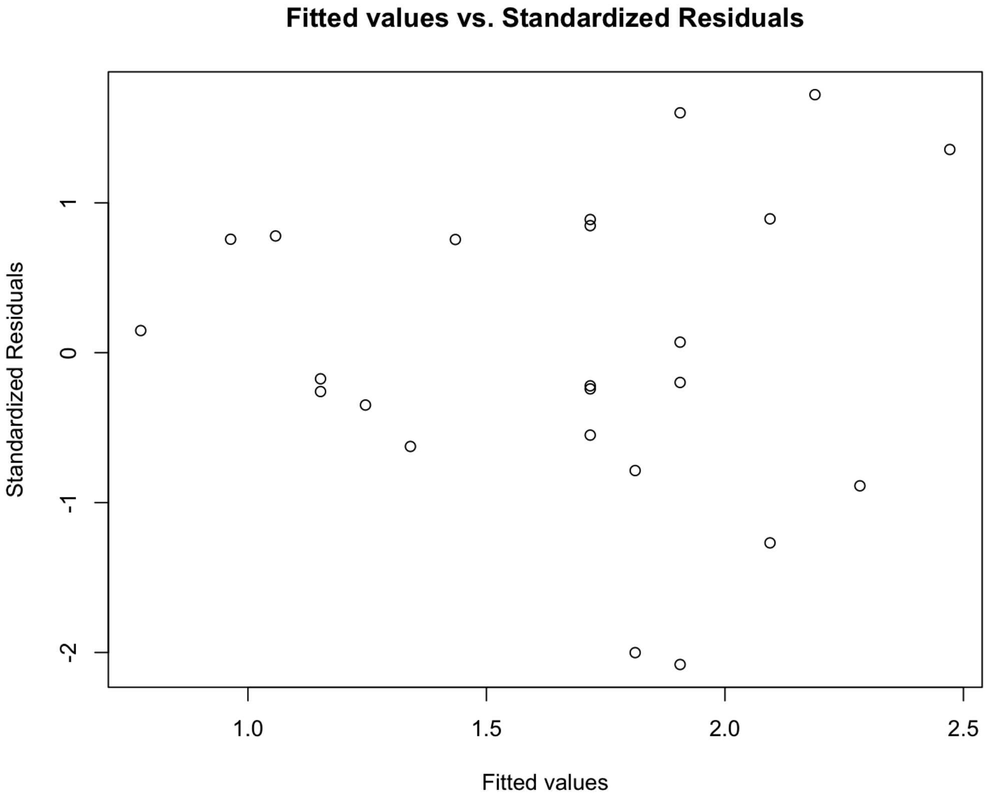
The error does appear to be slightly normally distributed but the values on the left and right side as well as the dip in the middle make it not very well distributed.

#1, part d(iii).

code:

```
plot(steel1$fitted,rstandard(steel1), xlab="Fiited values", ylab="Standardized Residuals", main="Fitted values vs. Standardized Residuals")
```

output:



The homoscedastic assumption does not seem reasonable as the plot above looks to be a heteroscedastic linear model. The values begin at zero but tend to spread towards the right.

#1, part e.

$$\textit{When } x = 182$$

$$\hat{Y} = -15.245 + 0.0942(182) = 1.899$$

#1, part f.

$$\textit{Residual of temperature } 182$$

$$\hat{Y} = -15.245 + 0.0942(182) = 1.899$$

$$e_1 = Y_1 - \hat{Y}_1 = 0.9 - 1.899 = -0.999$$

$$e_2 = Y_2 - \hat{Y}_2 = 1.81 - 1.899 = -0.089$$

$$e_3 = Y_3 - \hat{Y}_3 = 1.94 - 1.899 = 0.041$$

$$e_4 = Y_4 - \hat{Y}_4 = 2.68 - 1.899 = 0.781$$

The negative values exist because the parameter is centered around the median value, so those that are less will be negative.

#1, part g.

The proportion of the observed variation in the efficiency ratio can be attributed to the simple linear regression 45.14% of the time. (provided from the `summary` function in R)

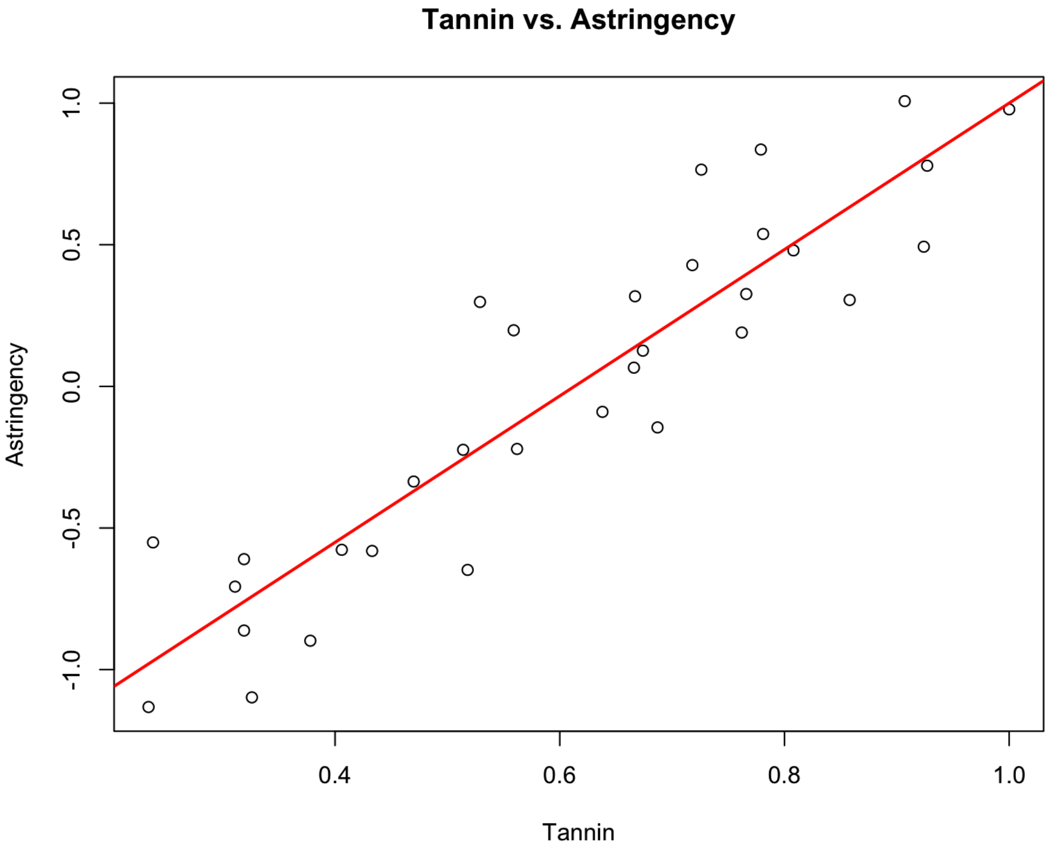
PROBLEM #2 - COMPUTATIONAL

#2, part a.

code:

```
winedata = read.table(“winedata.txt",header=TRUE)
wine1=lm(Astringency~Tannin,data=winedata)
plot(winedata[,2],winedata[,1],main="Tannin vs. Astringency", xlab=“Tannin", ylab=“Astringency")
abline(wine1,col=“red",lwd=2)
summary(wine1)
```

output:



From the Multiple R-squared summary output, the proportion is: 83.73%

#2, part b(i).

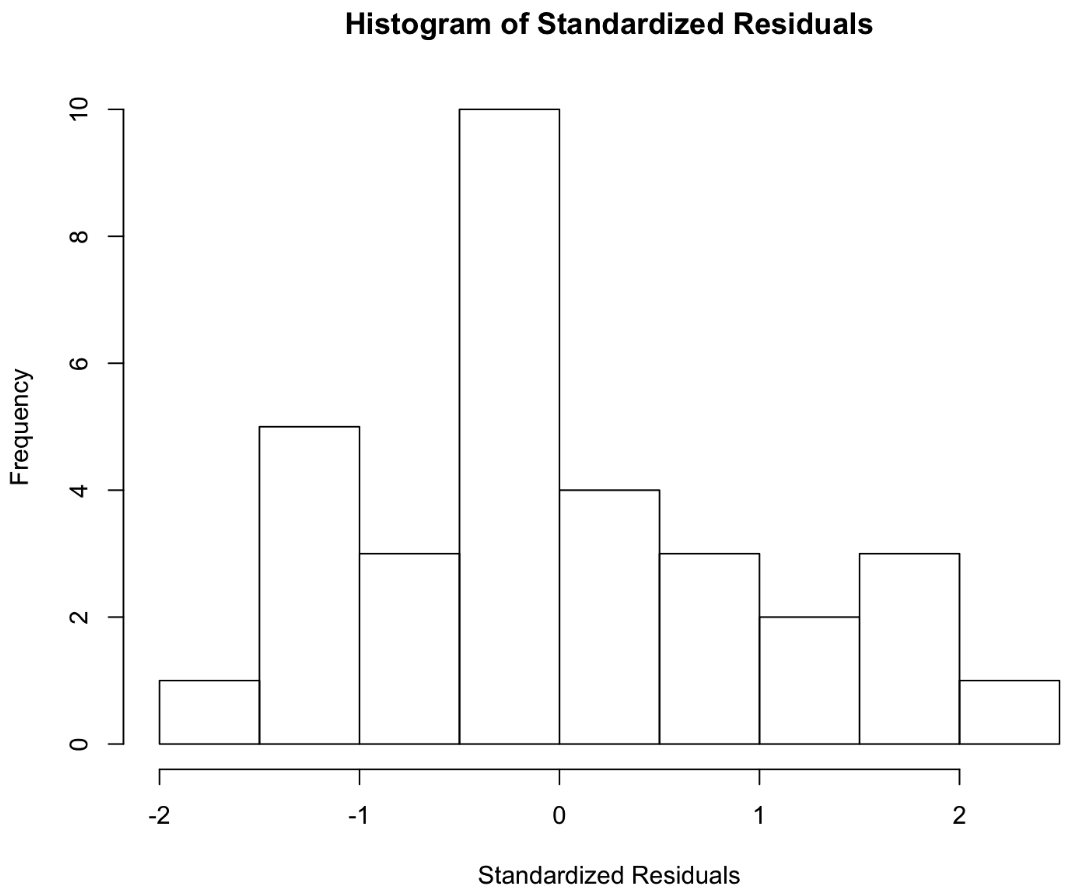
Observing the scatterplot from part a, the simple linear regression model appears to be reasonable.

#2, part b(ii).

<u>code:</u>

```
hist(rstandard(wine1), main="Histogram of Standardized Residuals", xlab="Standardized Residuals")
```

<u>output:</u>



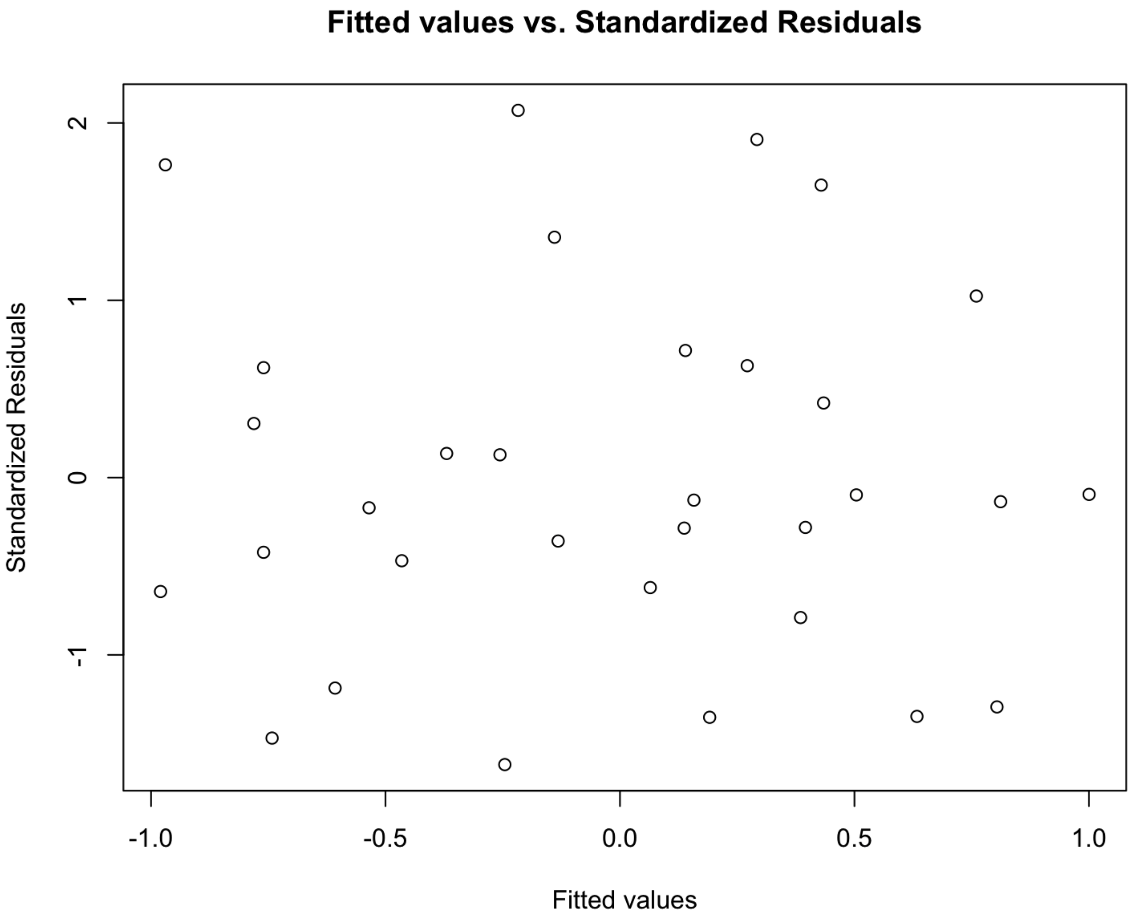
This histogram of the error does appear to be normally distributed.

#2, part b(iii).

code:

```
plot(wine1$fitted.values,rstandard(wine1),xlab="Fitted values", ylab="Standardized Residuals", main="Fitted values vs. Standardized Residuals")
```

output:



Observing the above fitted vs. standardized residuals of the win data, the homoscedastic assumption seems reasonable.

#2, part c.

code:

```
c=coef(summary(wine1))
cbind(low=c[,1]-qt(p=.975,df=30)*c[,2],high=c[,1]+qt(p=.975, df=30)*c[,2])
```

output:

```
                 low         high
(Intercept)   -1.857981    -1.311223
Tannin         2.160067     3.009823
```

With Tannin being the slope for this data, I am 95% confident that the true value of the slope lies between 2.160 and 3.010 with a significant level of .05.

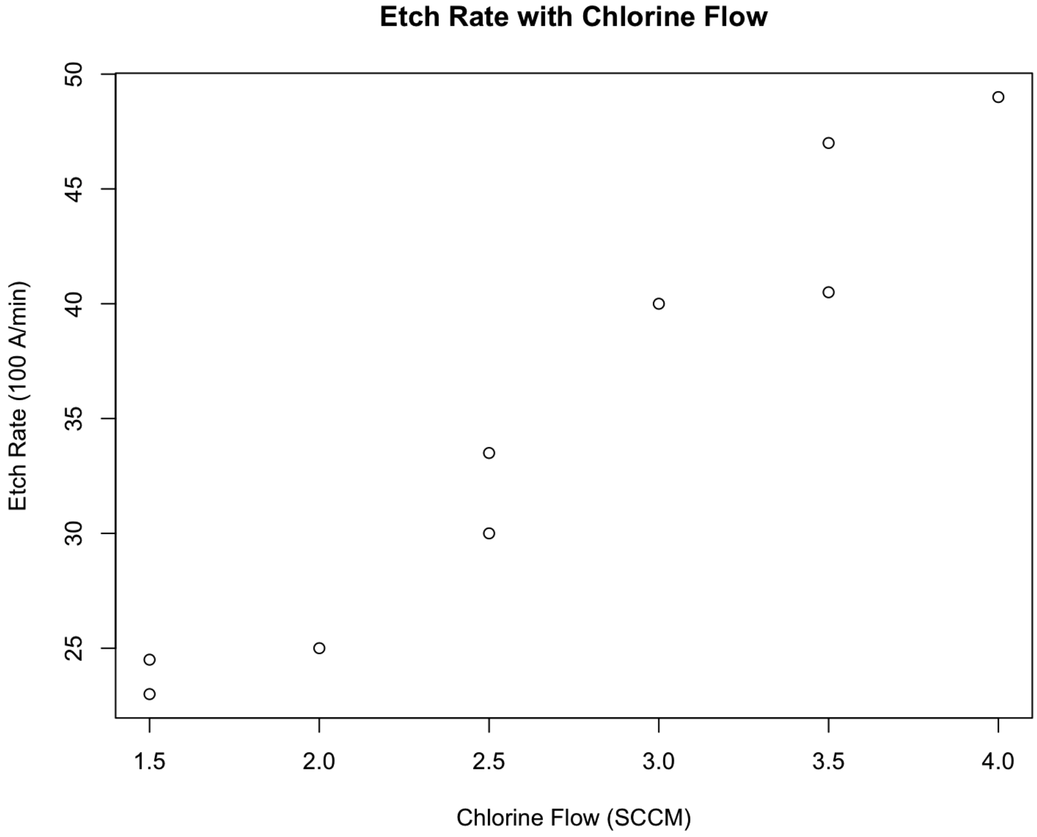
PROBLEM #3 - COMPUTATIONAL

#3, part a(i).

code:

```
chlorine_flow=c(1.5,1.5,2.0,2.5,2.5,3.0,3.5,3.5,4.0)
etch_rate=c(23.0,24.5,25.0,30.0,33.5,40.0,40.5,47.0,49.0)
plasma=data.frame(chlorine_flow,etch_rate)
plasma_lm=lm(etch_rate~chlorine_flow,data=plasma)
plot(plasma,main="Etch Rate with Chlorine Flow",xlab="Chlorine Flow (SCCM)", ylab="Etch Rate (100 A/min)”)
```

output:



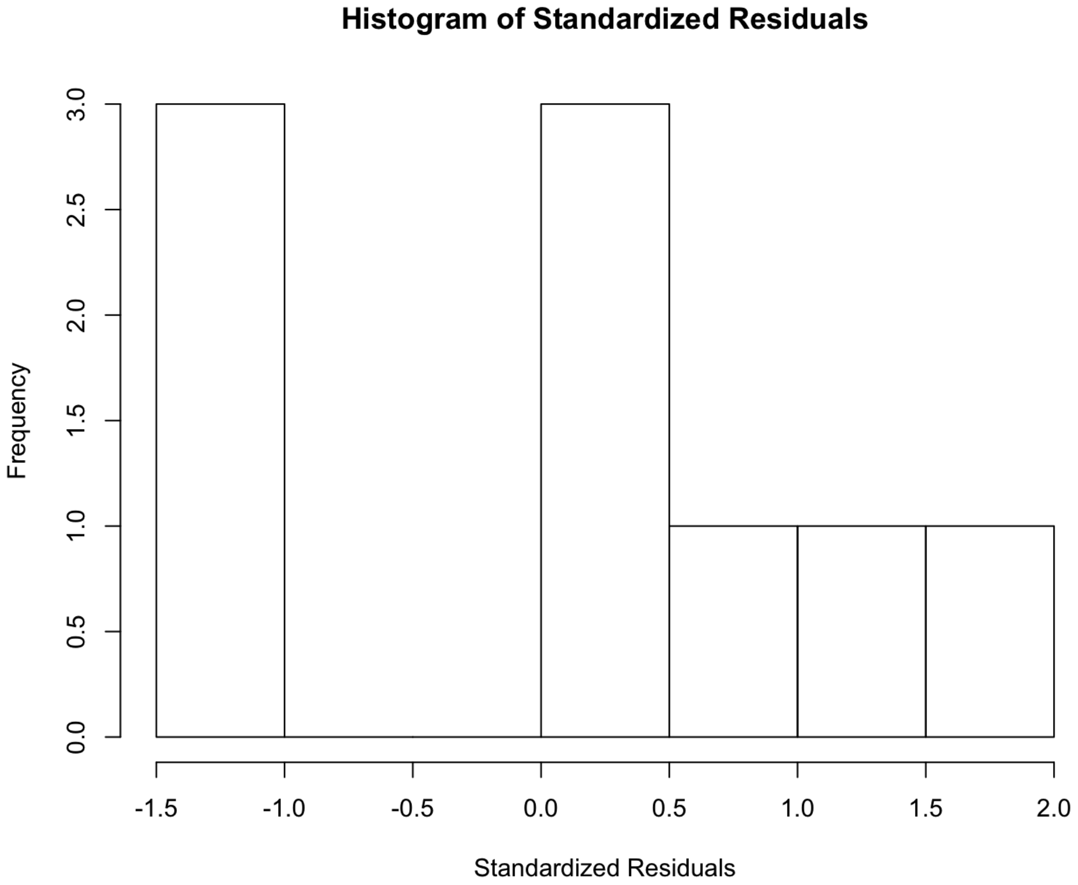
The linear regression model appears to be reasonable for this data.

#3, part a(ii).

code:

```
hist(rstandard(plasma_lm), main="Histogram of Standardized Residuals", xlab="Standardized Residuals")
```

output:



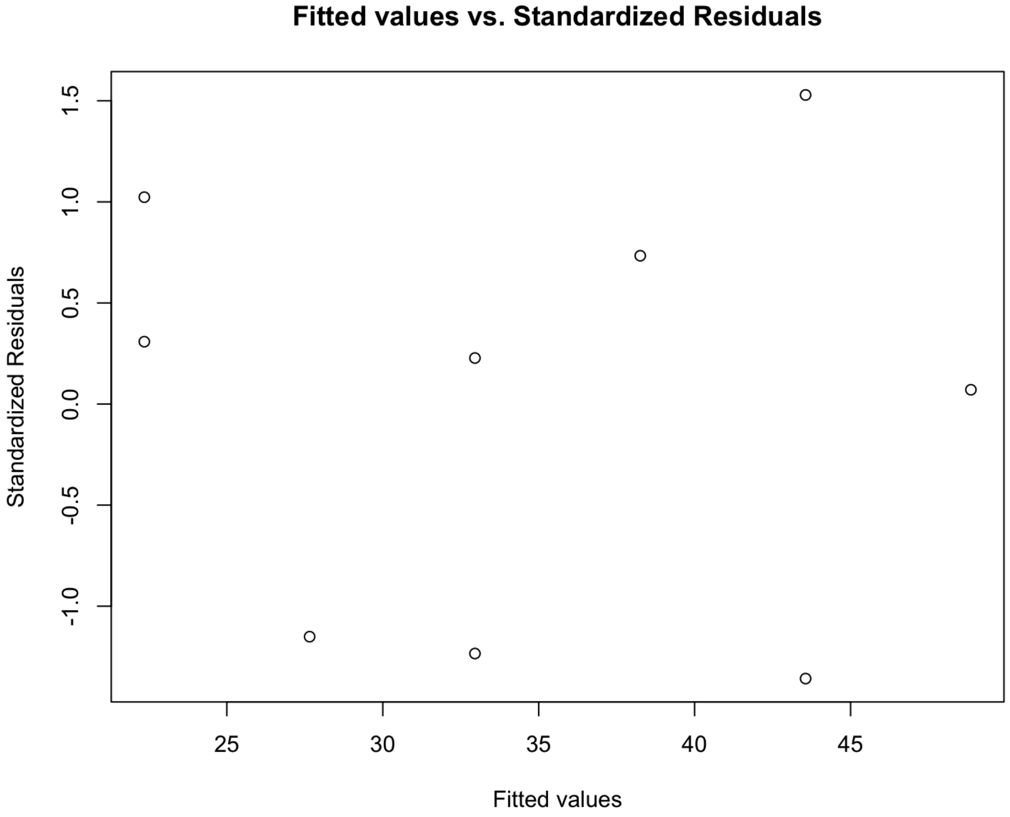
The error does appear to be slightly normally distributed though the increased error on the left makes it not a very strong normal distribution.

#3, part a(iii).

<u>code:</u>

```
plot(plasma_lm$fitted,rstandard(plasma_lm),xlab="Fitted values", ylab="Standardized Residuals",
main="Fitted values vs. Standardized Residuals")
```

<u>output:</u>



The homoscedastic assumption seems reasonable though with very few data points, it's more difficult to accurately determine visually.

#3, part a(iv).

<u>code:</u>

```
summary(plasma_lm)
```

<u>output:</u>

```
.
. ***(truncated data)***
.
Residual standard error: 2.546 on 7 degrees of freedom
Multiple R-squared:  0.9415,  Adjusted R-squared:  0.9332
F-statistic: 112.8 on 1 and 7 DF,  p-value: 1.438e-05
```

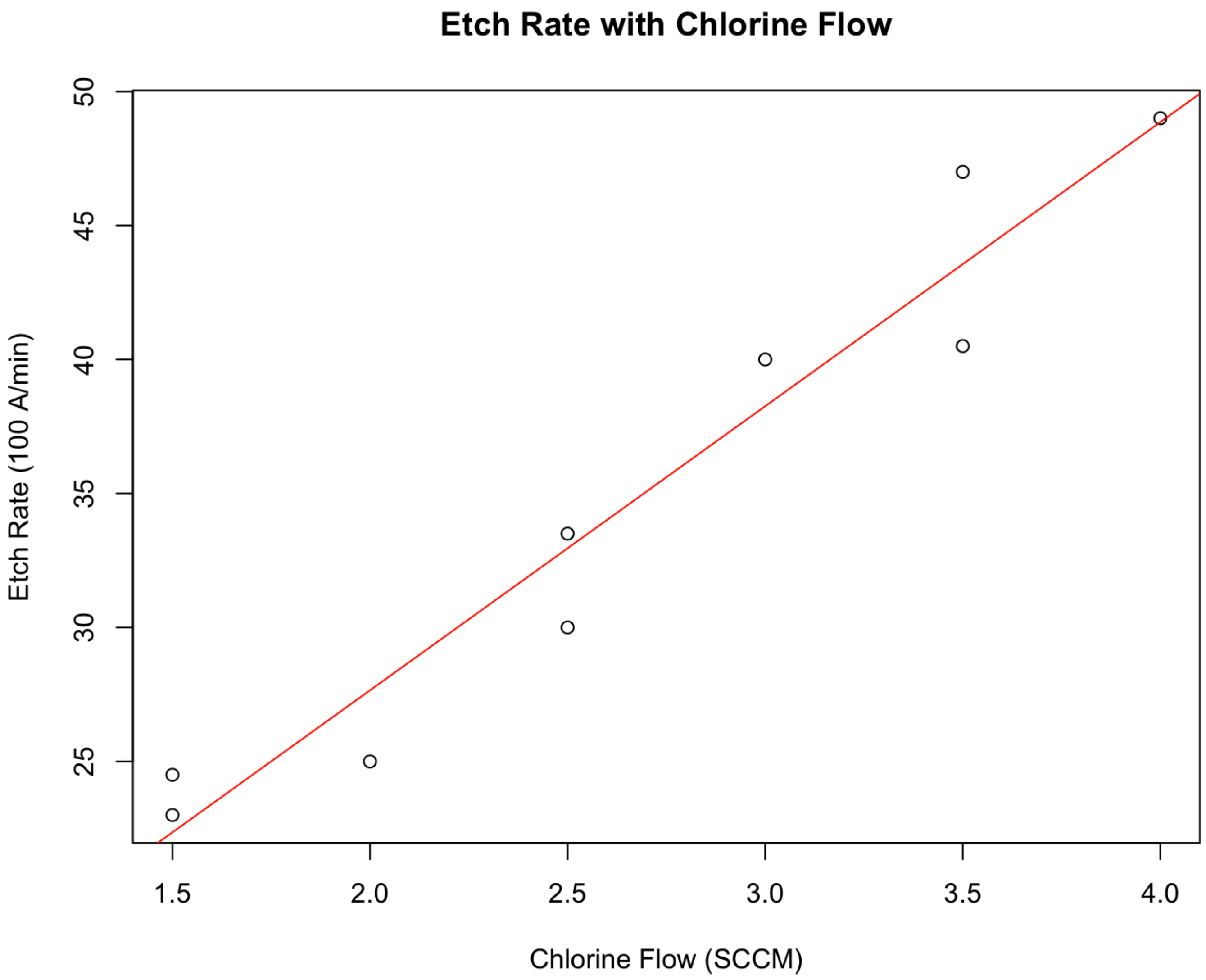
From the r^2 statistic, 94.15% of observed variation in etch rate can be explained by the approximate linear relationship between the two variables.

#3, part b.

code:

```
plot(plasma,main="Etch Rate with Chlorine Flow",xlab="Chlorine Flow (SCCM)", ylab="Etch Rate (100 A/min)")
abline(plasma_lm, col="red")
```

output:



There appears to be a useful (linear) relationship between chlorine flow and etch rate.