1) a) $Y \sim N_n(\mu, \sigma^2 I_n)$

$\Rightarrow AY \sim N_n(A\mu, A\sigma^2 I_n A^T)$

$BY \sim N_n(B\mu, B\sigma^2 I_n B^T)$

$Cov(AY, BY) = E[(AY - A\mu)(BY - B\mu)^T]$

$= E[A(Y-\mu)[B(Y-\mu)]^T]$

$= E[A(Y-\mu)(Y-\mu)^T B^T]$

$= A\, E[(Y-\mu)(Y-\mu)^T]\, B^T$

$= A\sigma^2 B^T = \sigma^2(AB^T) = \sigma^2 \cdot 0 = 0$

$\Rightarrow$ AY and BY are independent

b) AY and BY are independent $\Rightarrow (AY)(BY)^T = 0$

We have $(Y^TAY)^T(Y^TBY)$

$= Y^TA^TY\, Y^TBY$

$= Y^TAY(BY)^TY$ (as $A^T = A$, $B^T = B$ by assumption)

$= Y^T(0)Y = 0$

$\Rightarrow$ Y^TAY and Y^TBY are also independent

2) a) With $p = 3$, $Y = X\beta + \varepsilon$

$$\Rightarrow \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \epsilon_1 \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \epsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} 1_n & X_{11} & X_{12} \\ & \vdots & \vdots \\ & X_{n1} & X_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon$$

Scanned with CamScanner

(i) With $H_0: \beta_2 = 0$

$$\Rightarrow \underset{n\times 3}{X} = \begin{bmatrix} 1_n & x_{11} & x_{12} \\ & \vdots & \vdots \\ & x_{n1} & x_{n2} \end{bmatrix} = [1_n \; X_1 \; X_2]_{n\times 3}$$

And $L_{n\times 2} = \begin{bmatrix} 1_n & x_{11} \\ & \vdots \\ & x_{n1} \end{bmatrix} = [1_n \; X_1]_{n\times 2}$

- Prove $Im(L) \subset Im(X)$

Let $a \in \mathbb{R}^n$ is a vector such that $a \in Im(L)$

$\Rightarrow a = \underset{n\times 2}{L} \underset{2\times 1}{v}$ for some $v \in \mathbb{R}^2$

$$\Rightarrow a = [1_n \; X_1] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1 + v_2 x_{11} \\ \vdots \\ v_1 + v_2 x_{n1} \end{bmatrix}$$

$$= \begin{bmatrix} v_1 + v_2 x_{11} + 0 x_{12} \\ \vdots \\ v_1 + v_2 x_{n1} + 0 x_{n2} \end{bmatrix} = [1_n \; X_1 \; X_2] \begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix}$$

$\Rightarrow$ a can be written as $a = X \begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix}$ for $\begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix} \in \mathbb{R}^3$

$\Rightarrow a \in Im(X) \Rightarrow Im(L) \subset Im(X)$

(ii) With $H_0: \beta_1 + \beta_2 = 0$, $X_{n\times 3} = [1_n \; X_1 \; X_2]$

$\Rightarrow H_0: \beta_1 = -\beta_2$

$\Rightarrow$ Our model becomes $y_i = \beta_0 + \beta_1 (X_{i1} - X_{i2}) + \varepsilon$

$\Rightarrow L_{n\times 2} = [1_n \; (X_1 - X_2)]_{n\times 2}$

Let $b \in \mathbb{R}^n$ is a vector in $Im(L)$

$\Rightarrow b = Lv$ for some $v \in \mathbb{R}^2$

$$\Rightarrow b = [1_n \; (X_1 - X_2)] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 & (x_{11} - x_{12}) \\ \vdots & \vdots \\ 1 & (x_{n1} - x_{n2}) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$= \begin{bmatrix} v_1 + v_2(x_{11} - x_{12}) \\ \vdots \\ v_1 + v_2(x_{n1} - x_{n2}) \end{bmatrix} = \begin{bmatrix} v_1 + v_2 x_{11} - v_2 x_{12} \\ \vdots \\ v_1 + v_2 x_{n1} - v_2 x_{n2} \end{bmatrix}$$

$$= \underset{n\times 3}{[1_n \; X_1 \; X_2]} \begin{bmatrix} v_1 \\ v_2 \\ -v_2 \end{bmatrix}_{3\times 1} = X \begin{bmatrix} v_1 \\ v_2 \\ -v_2 \end{bmatrix}$$

$\Rightarrow b \in Im(X) \Rightarrow Im(L) \subset Im(X)$

Scanned with CamScanner

b) $SSE_X = Y^T(I - H_X)Y$

$SSE_L = Y^T(I - H_L)Y$

$\Rightarrow SSE_L - SSE_X = Y^T(I - H_L)Y - Y^T(I - H_X)Y$

$= Y^TY - Y^TH_LY - Y^TY + Y^TH_XY = Y^TH_XY - Y^TH_LY$

$= Y^T(H_X - H_L)Y$

$$\Rightarrow f = \frac{(SSE_L - SSE_X)/(p-s)}{SSE_X/(n-p)} = \frac{Y^T(H_X - H_L)Y/(p-s)}{Y^T(I - H_X)Y/(n-p)}$$

c) • Symmetric

$(H_X - H_L)^T = H_X^T - H_L^T$

But H_X and H_L are symmetric themselves $\Rightarrow H_X = H_X^T$, $H_L = H_L^T$

$\Rightarrow (H_X - H_L)^T = H_X - H_L$

• Idempotent

$(H_X - H_L)^2 = (H_X - H_L)(H_X - H_L) = H_X^2 - H_XH_L - H_LH_X + H_L^2$

$= H_X - H_XH_L - H_LH_X + H_L$

(because H_X, H_L are idempotent themselves $\Rightarrow H_X = H_X^2, H_L = H_L^2$)

Also $Im(L) \subset Im(X) \Rightarrow H_XH_L = H_LH_X = H_L$

$\Rightarrow (H_X - H_L)^2 = H_X - H_L - H_L + H_L = H_X - H_L$

$\Rightarrow H_X - H_L$ is idempotent

• $(I_n - H_X)(H_X - H_L) = I_nH_X - I_nH_L - H_X^2 + H_XH_L$

$= H_X - H_L - H_X + H_L = (H_X - H_X) + (H_L - H_L) = 0$

d) From question 1) if $AB^T = 0 \Rightarrow AY$ & BY are independent

From part c, $(I_n - H_X)(H_X - H_L)^T = 0$

$\Rightarrow (I_n - H_X)Y$ and $(H_X - H_L)Y$ are independent

Also, $Y^T(H_X - H_L)Y$ is a quadractic form $\sim \chi^2$ with

$df = tr(H_X - H_L) = tr(H_X) - tr(H_L) = p - s$

$Y^T(I_n - H_X)Y$ is also $\sim \chi^2$ with $df = tr(I_n) - tr(H_X)$

$= n - p$

$$\Rightarrow f = \frac{Y^T(H_X - H_L)Y/(p-s)}{Y^T(I - H_X)Y/(n-p)} \sim \frac{\chi^2_{(p-s)}/(p-s)}{\chi^2_{(n-p)}/(n-p)}$$

$\Rightarrow f \sim F_{(p-s),(n-p)}$

3) a) We want $\bar{\hat{Y}} = \bar{Y}$ or $n^{-1} 1_n^T \hat{Y} = n^{-1} 1_n^T Y$

We have $1_n \in Im(X) \Rightarrow \underset{(n\times 1)}{1_n} = \underset{n\times n}{X} \underset{n\times 1}{v}$ for $v \in \mathbb{R}^n$

$$1_n^T \hat{Y} = v^T X^T \hat{Y} = v^T X^T (X(X^TX)^{-1}X^TY) \quad (\text{because } \hat{Y} = HY)$$

$$= v^T(X^TX(X^TX)^{-1})X^TY = v^T I X^T Y = v^T X^T Y$$

$$1_n^T Y = v^T X^T Y$$

$$\Rightarrow 1_n^T \hat{Y} = 1_n^T Y \Rightarrow n^{-1} 1_n^T \hat{Y} = n^{-1} 1_n^T Y \Rightarrow \bar{\hat{Y}} = \bar{Y}$$

b) We have $\sum(\hat{Y}_i - \bar{Y})^2 = SSR = Y^T(H - \frac{1}{n}J)Y$

$\sum(Y_i - \bar{Y})^2 = SSTO = Y^T(I_n - \frac{1}{n}J)Y$

with $J = 1_n 1_n^T$

as well as $\sum(\hat{Y}_i - Y_i)^2 = SSE = Y^T(I - H)Y$

and $SSTO = SSR + SSE$

Let $(\hat{Y}_i - \bar{\hat{Y}}) = (\hat{Y}_i - \bar{Y}) = A$ (as $\bar{\hat{Y}} = \bar{Y}$ in 3a))

$(Y_i - \bar{Y}) = B$

$$\Rightarrow \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})(Y_i - \bar{Y}) = \sum_{i=1}^n AB$$

$$= \cancel{\sum_{i=1}^n (A^2 + B^2 - A^2 - B^2 - 2AB}$$

Let's look at $(A-B)^2 = A^2 - 2AB + B^2$

$$\Rightarrow AB = \frac{A^2 + B^2 - (A-B)^2}{2}$$

$$\Rightarrow \sum_{i=1}^n AB = \frac{1}{2}\left(\sum A^2 + \sum B^2 - \sum(A-B)^2\right)$$

$$= \frac{1}{2}\left(\sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \bar{Y})^2 - \sum(\hat{Y}_i - \cancel{\bar{Y}} - Y_i + \cancel{\bar{Y}})^2\right)$$

$$= \frac{1}{2}(SSR + SSTO - SSE)$$

$$= \frac{1}{2}(SSR + SSR + SSE - SSE) = \frac{1}{2} \cdot 2 \cdot SSR$$

$$= SSR = Y^T(H - \frac{1}{n} 1_n 1_n^T)Y$$

We have $SSR \geq 0$, $\sqrt{\sum(\hat{Y}_i - \bar{\hat{Y}})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2} = \sqrt{SSR} \cdot \sqrt{SSTO} > 0$

(because $SSR \geq 0$, $SSTO \geq 0$, and the term is in denominator)

$$\Rightarrow r_{\hat{Y},Y} = \frac{SSR}{\sqrt{SSR}\sqrt{SSTO}} \geq 0 \Rightarrow \text{it will never be} < 0$$

c) $$r^2_{\hat{Y}Y} = \left(\frac{SSR}{\sqrt{SSR}\sqrt{SSTO}}\right)^2 = \frac{SSR^2}{SSR \cdot SSTO} = \frac{SSR}{SSTO}$$

$$= \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO} = R^2$$

Scanned with CamScanner

hw6

Giang Vu

10/1/2020

HOMEWORK 6

4)

a)

```
#read data
hw6_dt <- read.delim("/Users/giangvu/Desktop/STAT 2131 - Applied Stat Methods 1/HW/hw6/steam_text-1.txt
#regress steam (Y) onto fat (X1) and glycerine (X2)
hw6_md <- lm(steam ~ fat + glycerine, data = hw6_dt)
#i)
hw6_sm<-summary(hw6_md) #mean model
hw6_anova <- summary(aov(hw6_md)) #variance model
hw6_sm
```

```
##
## Call:
## lm(formula = steam ~ fat + glycerine, data = hw6_dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7977 -1.0015 -0.4424  1.0575  3.2397
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.625      2.247   2.058   0.0516 .
## fat            1.728      1.168   1.480   0.1529
## glycerine     -6.628      7.578  -0.875   0.3912
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.546 on 22 degrees of freedom
## Multiple R-squared:  0.1755, Adjusted R-squared:  0.1005
## F-statistic: 2.341 on 2 and 22 DF,  p-value: 0.1197
```

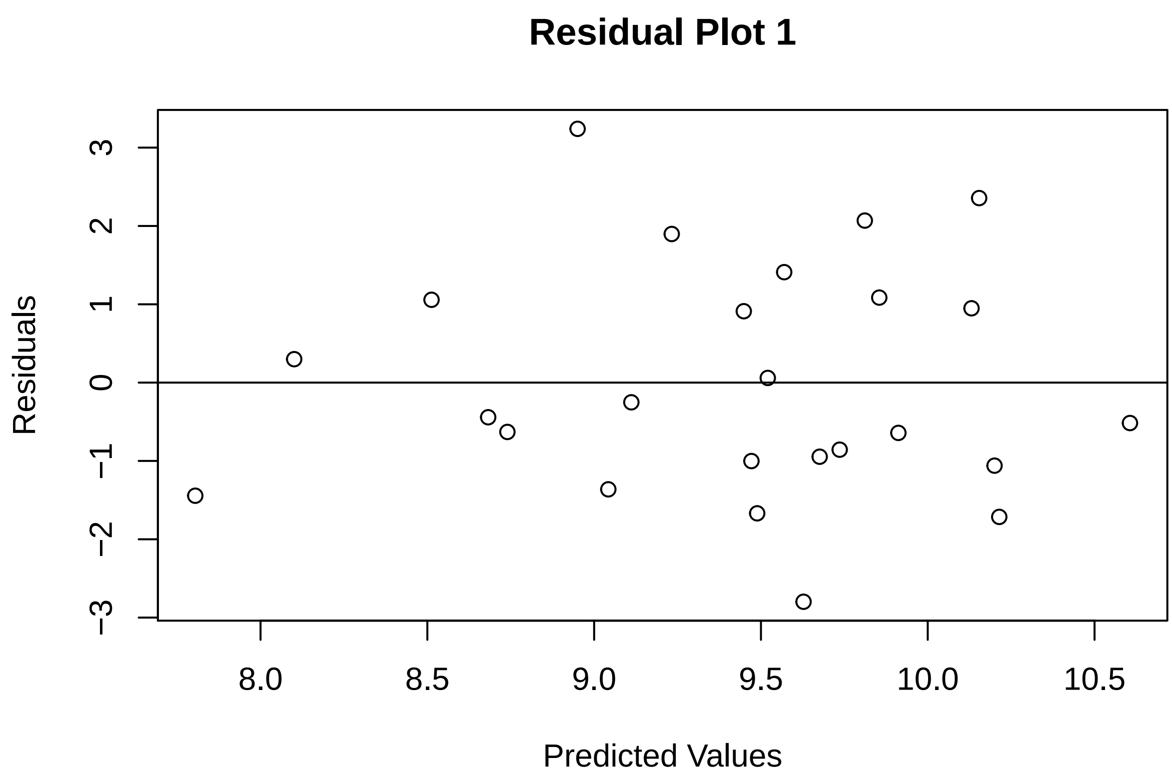
```
hw6_anova
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fat          1   9.37   9.370   3.918 0.0604 .
```

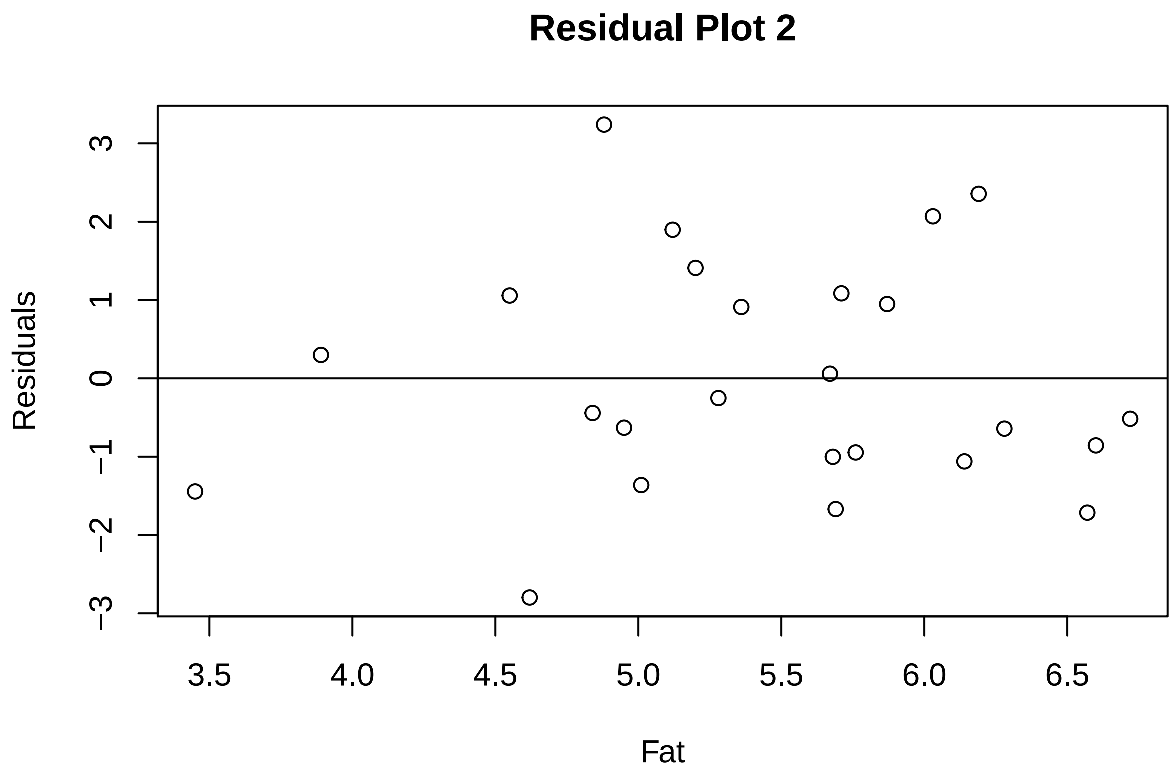
```
## glycerine    1   1.83   1.829   0.765 0.3912
## Residuals   22  52.62   2.392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the coefficients: As the value for variable fat increases by 1 unit, the value for steam is expected to increase by 1.728 units. As the value for variable glycerin increases by 1 unit, the value for steam is expected to decreease by -6.628 units. However, the coefficients estimates for both variables fat and glycerin in this model are not statistically significant given alpha = 0.05

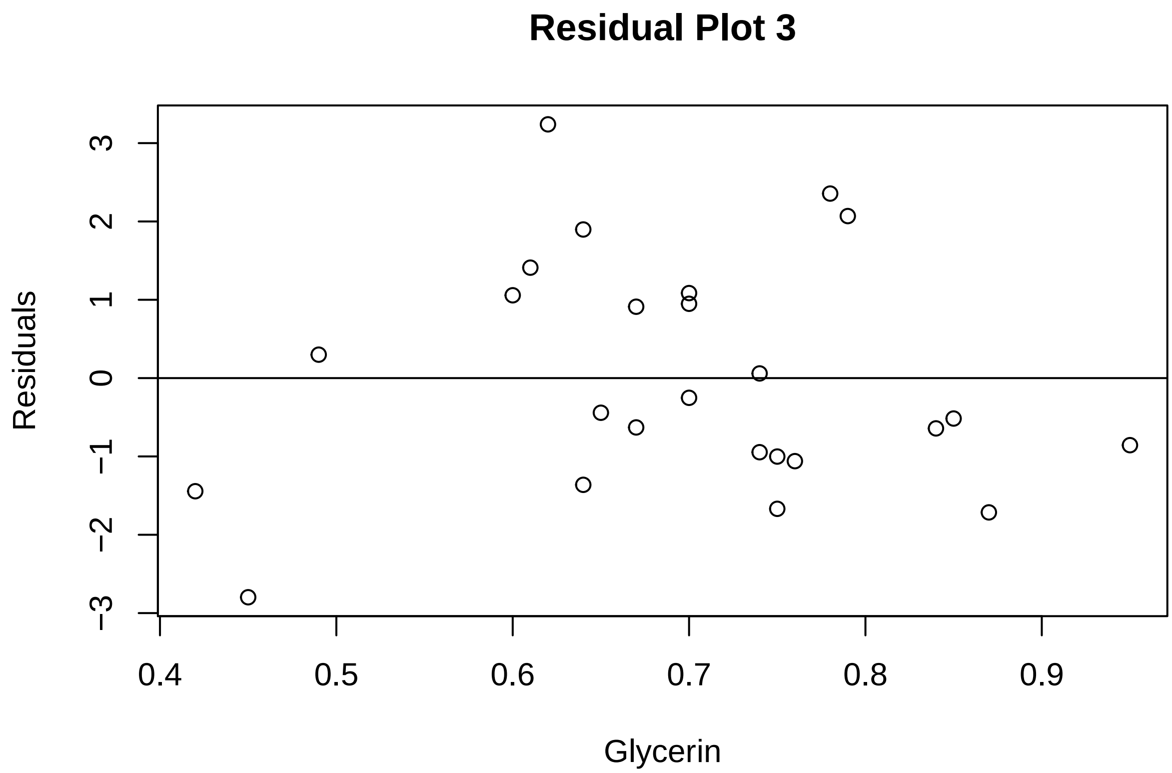
```
#ii) plot residual as fcn of Y^, fat, glycerine
plot(predict(hw6_md), resid(hw6_md),
     ylab  = "Residuals", xlab = "Predicted Values", main ="Residual Plot 1")
abline(0,0)
```




```
plot(hw6_dt$fat, resid(hw6_md),
     ylab  = "Residuals", xlab = "Fat", main ="Residual Plot 2")
abline(0,0)
```




```r
plot(hw6_dt$glycerine, resid(hw6_md),
     ylab  = "Residuals", xlab = "Glycerin", main ="Residual Plot 3")
abline(0,0)
```



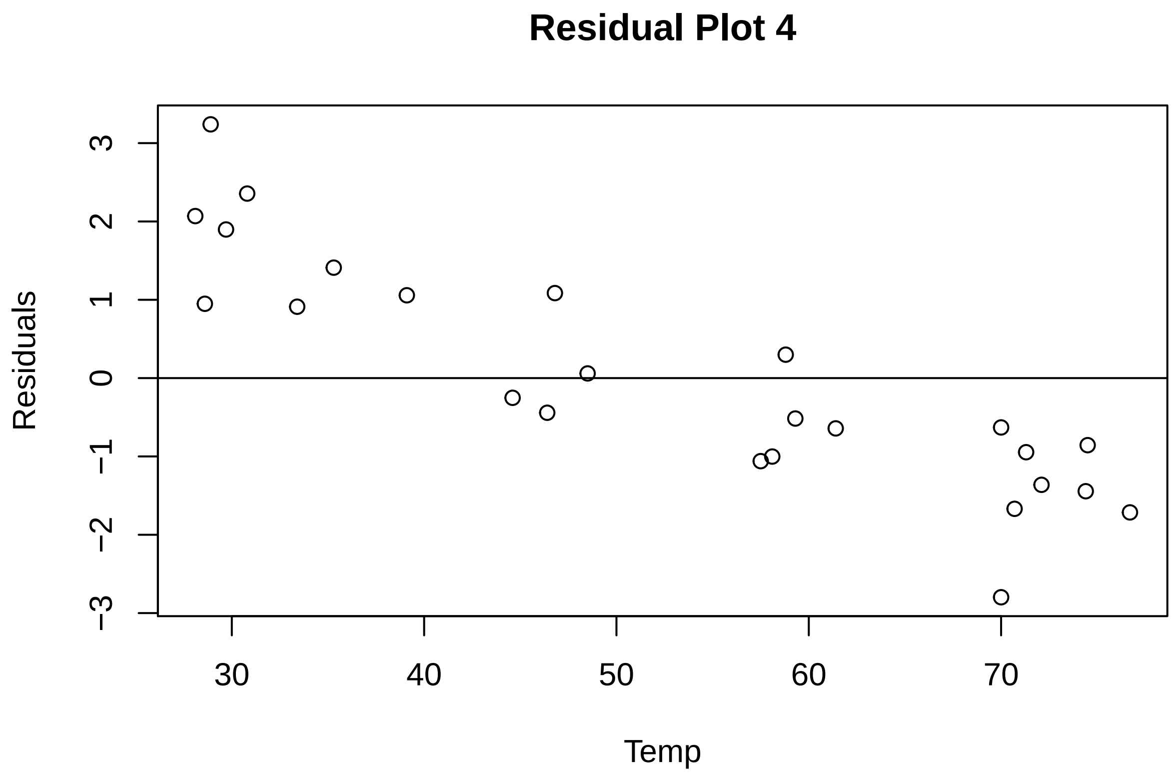
Looking at the 3 plots, I don't see any relationship between residuals and glycerin, fat, or predicted Y.

```
#iii)
linearHypothesis(hw6_md,c("fat=0","glycerine=0")) #F-test
```

```
## Linear hypothesis test
##
## Hypothesis:
## fat = 0
## glycerine = 0
##
## Model 1: restricted model
## Model 2: steam ~ fat + glycerine
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     24 63.816
## 2     22 52.617  2    11.199 2.3413 0.1197
```

p-value for F test is 0.1197 > 0.05 => fail to reject H0, also I noticed that R square of this model is really small too (0.1755) We can conclude that the two variables fat and glycerin cannot fully explain the variability of Y (steam). Their coefficients are not statistically significant.

```
#iv) temp against residual
plot(hw6_dt$temp, resid(hw6_md),
     ylab  = "Residuals", xlab = "Temp", main ="Residual Plot 4")
abline(0,0)
```



There's a decreasing pattern, so we can say that residuals can be explained by temp. This is probably what makes our original model a poor model.

b)

```r
#i) regress steam (Y) onto fat (X1), glycerine (X2) and temp (X3)
hw6_md2 <- lm(steam ~ fat + glycerine + temp, data = hw6_dt)
summary(hw6_md2)
```

```
##
## Call:
## lm(formula = steam ~ fat + glycerine + temp, data = hw6_dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2348 -0.4116  0.1240  0.3744  1.2979
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.514814   1.062969   8.951 1.30e-08 ***
## fat          0.713592   0.502297   1.421     0.17
## glycerine    0.330497   3.267694   0.101     0.92
## temp        -0.079928   0.007884 -10.138 1.52e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.652 on 21 degrees of freedom
```

```
## Multiple R-squared:  0.8601, Adjusted R-squared:  0.8401
## F-statistic: 43.04 on 3 and 21 DF,  p-value: 3.794e-09
```

```r
linearHypothesis(hw6_md2,c("fat=0","glycerine=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## fat = 0
## glycerine = 0
##
## Model 1: restricted model
## Model 2: steam ~ fat + glycerine + temp
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     23 18.223
## 2     21  8.927  2    9.2964 10.934 0.0005569 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 1: steam ~ temp
Model 2: steam ~ fat + glycerin + temp

p value for F test = 0.0005569 < 0.05 => reject H0 We can say that fat and glycerine do have some impact on steam now that temp is included in the model. However, with a new variable temp added into the model and alpha = 0.05, the coefficient estimates for fat and glycerine are sill not statistically significant (based on lm() results), but the coefficient estimate for temp is statistically significant.

ii) Because of the inclusion of variable X3 (temp). Our original model is underfitting because we leave outt an important variable - temp, so we saw that a lot of residuals are now accounted for by temp. With temp included, we have a better model, higher R squared.