STAT 2131
Orly Olbum

HOMEWORK 6

1. $Y \sim N(\mu, \sigma^2 I_n)$, $\mu \in \mathbb{R}^n$
$A, B \in \mathbb{R}^{n \times n}$

(a) Show that if $AB^T = 0_{n\times n}$, AY and BY are ind.
For Y, A, B above,
$AY \sim N(A\mu_y, A\sigma^2 I_n A^T)$, $BY \sim N(B\mu_y, B\sigma^2 I_n B^T)$
(Proof: $E(AY) = AE(Y) = A\mu_y$
$$\begin{aligned}Var(AY) &= E[(AY - A\mu_y)(AY - A\mu_y)^T] \\ &= E[(AY - A\mu_y)(Y - \mu_y)^T A^T] \\ &= AE[(Y - \mu_y)(Y - \mu_y)^T]A^T \\ &= A\sigma^2 I_n A^T \text{ , and same for } BY)\end{aligned}$$
If Y is as above, and $AB^T = 0$, the joint distribution of AY & BY is bivariate N
To show independence of AY, BY, show that $Cov(AY, BY) = 0$
$$\begin{aligned}\to Cov(AY, BY) &= E[(AY - A\mu_y)(BY - B\mu_y)^T] \\ &= E[A(Y - \mu_y)(Y - \mu_y)^T B^T] \\ &= AE[(Y - \mu_y)(Y - \mu_y)^T]B^T = A\sigma^2 B^T = AB^T\sigma^2 = 0\end{aligned}$$
So, AY and BY are independent

(b) Suppose A, B are symmetric | idempotent, $AB = 0_{n\times n}$
Show Y^TAY, Y^TBY are independent.
Say $q_1 = Y^TAY$, $q_2 = Y^TBY$ and $AB = ATT^TB = 0$
$\Rightarrow T^TATT^TBT = 0$, $C = T^TAT$ and $D = T^TBT$
Making $CK = (T^TAT)(T^TBT) = T^TABT = T^TT = 0$, so $CK = KC$
And $Q^TCQ = \begin{bmatrix} E_1 & 0 \\ 0 & 0 \end{bmatrix}$, $Q^TDQ = \begin{bmatrix} 0 & 0 \\ 0 & E_2 \end{bmatrix}$

say $V = Q^T T^{-1} Y$, $E(V) = Q^T T^{-1} Y$, ($V \to$ vector of standard normals)
$var(V) = Q^T T^{-1}\mu = Q^T T^{-1} T T^T T^{-1} Q = I$
so $Y = TQV$ and $Y^T = V^T Q^T T^T$
$q_1 = Y^TAY = V^TQ^TT^TATQV = V^TQ^TT^T(T^{T-1}CT^{-1})TQV$
$= V^TQ^TCQV = V_1^TE_1V_1$, and $q_2 = V_2^TE_2V_2$
where V_1 is first half, V_2 is second half of V,
so q_1 & q_2 are independent

2. $X \in \mathbb{R}^{n\times p}$, $p \leq n$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = Y = X\beta + \epsilon, \quad \epsilon \sim N_n(0_n, \sigma^2 I_n)$$

for some non-random $\beta \in \mathbb{R}^p$

$L \in \mathbb{R}^{n\times s}$ (full-rank design matrix), $s<p$

$im(L) \subset im(X)$

$E(Y) = L\gamma$ for $\gamma \in \mathbb{R}^s$

Test $H_0: E(Y) = L\gamma$ for $\gamma \in \mathbb{R}^s$

@ $p=3$, $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$

Find design matrices X, L, show that $im(L) \subset im(X)$

For $p=3$: $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$, 3×1 matrix $\quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, and $\underset{n\times3}{X} = \begin{bmatrix} 1 & X_1 & X_2 \\ 1 & X_1 & X_2 \\ \vdots & \vdots & \vdots \\ 1 & X_1 & X_2 \end{bmatrix}$

(i) $H_0: \beta_2 = 0$

$\underset{n\times2}{L} = \begin{bmatrix} 1 & X_1 \\ 1 & X_1 \\ \vdots & \vdots \\ 1 & X_1 \end{bmatrix}$

X, the same

(essentially removing a term)

For V, a 2×1 vector and $a \in im(L)$

$\Rightarrow a = [1_n \; X_1]V \rightarrow a = [1_n \; X_1]\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$

$= \begin{bmatrix} V_1 + X_{11}V_2 + X_{12}\,0 \\ \vdots \\ V_1 + X_{n1}V_2 + X_{n2}\,0 \end{bmatrix} = [1 \; X_1 \; X_2]\begin{bmatrix} V_1 \\ V_2 \\ 0 \end{bmatrix}$

$\Rightarrow a \in im(X)$

(ii) $H_0: \beta_1 + \beta_2 = 0$

$Y = \beta_0 + \beta_c(X_{i1} - X_{i2}) + \epsilon_i$

$\underset{n\times1}{L} = \begin{bmatrix} 1 & X_{11}+X_{12} \\ 1 & \vdots \\ \vdots & \vdots \\ 1 & X_{n1}+X_{n2} \end{bmatrix}$

same X

(terms cancel)

$\underset{2\times1}{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \rightarrow a \in im(L)$

$a = [1_n \; X_1 \; X_2]\begin{bmatrix} V_1 \\ V_2 \\ V_2 \end{bmatrix}$

$= \begin{bmatrix} V_1 + V_2(X_{11}+X_{12}) & 0 \\ \vdots & \vdots \\ V_1 + V_2(X_{n1}+X_{n2}) & 0 \end{bmatrix} \rightarrow a \in im(X)$

2, continued

(b) $H_X = X(X^TX)^{-1}X^T$, $H_L = L(L^TL)^{-1}L^T$

SSE_X = SSE when design matrix is X

SSE_L = SSE when design matrix is L

F-stat: $f = \dfrac{(SSE_L - SSE_X)/(p-s)}{SSE_X/(n-p)}$

Show that we can write f as:

$$f = \frac{Y^T(H_X - H_L)Y/(p-s)}{Y^T(I_n - H_X)Y/(n-p)}$$

<u>Denominator</u>: $SSE_X = \sum \epsilon_i^2 = \sum(Y_i - \hat{Y}_i)^2$

$= \epsilon^T\epsilon = (Y - XB)^T(Y - XB) = Y^TY - 2B^TX^TY + B^TX^TXB$

$= Y^TY - 2B^TX^TY + B^TX^TX(X^TX)^{-1}X^TY$ (sub for B)

$= Y^TY - 2B^TX^TY + B^TIX^TY = Y^TY - Y^TX(X^TX)^{-1}X^TY$

$= Y^T(I - H)Y$

and rank $(I - H) = n - p$ (degrees of freedom)

so denominator: $SSE_X/(n-p) = Y^T(I - H_X)Y/(n-p)$

<u>Numerator</u>:

$SSE_X = Y^T(I - H_X)Y$, so $SSE_L = Y^T(I - H_L)Y$

expand: $Y^T(I - H_L)Y = Y^TY - Y^TL(L^TL)^{-1}L^TY$

$SSE_L - SSE_X = (Y^TY - Y^TL(L^TL)^{-1}L^TY) - (Y^TY - Y^TX(X^TX)^{-1}X^TY)$

$= -Y^TL(L^TL)^{-1}L^TY + Y^TX(X^TX)^{-1}X^TY$

$= -Y^TH_LY + Y^TH_XY = Y^TH_XY - Y^TH_LY = Y^T(H_X - H_L)Y$

and rank $(H_X - H_L) = (n-s) - (n-p) = -s + p = p - s$

so $f = \dfrac{(SSE_L - SSE_X)/(p-s)}{SSE_X/(n-p)} = \dfrac{Y^T(H_X - H_L)Y/(p-s)}{Y^T(I_n - H_X)Y/(n-p)}$

(c) Show that $(H_X - H_L)$ is symmetric & idempotent and that $(I_n - H_X)(H_X - H_L) = 0_{n \times n}$

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

$(H_X - H_L)^T = H_X^T - H_L^T = [(X(X^TX)^{-1}X^T)^T] - [(L(L^TL)^{-1}L^T)^T]$

$= X[(X^TX)^{-1}]^TX^T - L[(L^TL)^{-1}]^TL = X[(X^TX)^T]^{-1}X^T - L[(L^TL)^T]^{-1}L^T$

$= X(X^TX)^{-1}X^T - L(L^TL)^{-1}L^T = H_X - H_L$

next page →

20, continued

Show that $(H_x - H_L)$ is idempotent

$(H_x - H_L) = (H_x - H_L)^2 = (H_x - H_L)(H_x - H_L)$

$= (X(X^TX)^{-1}X - L(L^TL)^{-1}L^T)(X(X^TX)^{-1}X^T - L(L^TL)^{-1}L^T)$

$= H_x^2 - X(X^TX)^{-1}X^T L(L^TL)^{-1}L^T - L(L^TL)^{-1}L^T X(X^TX)^{-1}X^T + H_L^2$

$= H_x - H_xH_L - H_LH_x + H_L$ because $im(L) \in im(X)$

$= H_x - 2H_L + H_L = H_x - H_L$ by properties of H

H are idempotent, proof in HW5

Show that $(I_n - H_x)(H_x - H_L) = 0_{n \times n}$

$(I_n - H_x)(H_x - H_L) = IH_x - IH_L - H_xH_x + H_xH_L$

$= H_x - H_L - H_x + H_L = 0_{n \times n}$

(d) Use (1) to prove $f \sim F(p-s, n-p)$ when $H_0: E(Y) = L\gamma$, $\gamma \in \mathbb{R}^s$ is true

From (1), $AB^T = 0$ means AY and BY are independent, and from the results in (c), $(I_n - H_x)(H_x - H_L)^T = (I_n - H_x)(H_x - H_L) = 0$ so the numerator part & denominator part: $(H_x - H_L)Y$ & $(I_n - H_x)Y$ are independent (under H_0)

Additionally, $Y^T(H_x - H_L)Y$ is quadratic form, χ^2 distributed with df $= tr(H_x - H_L)$

$tr(H_x - H_L) = tr(H_x) - tr(H_L) = p - s$

$Y^T(I_n - H_x)Y \rightarrow tr(I_n - H_x) = tr(I_n) - tr(H_x) = n - p$

Which tells us that under H_0,

$$f \sim F_{p-s, n-p} = \frac{Y^T(H_x - H_L)Y / p-s}{Y^T(I_n - H_x)Y / n-p}$$

3. $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ full-rank design matrix and
$R^2 = 1 - \frac{SSE}{SSTO}$ (uegress Y onto X) $\left(= \frac{SSR}{SSTo}\right)$
$\bar{\hat{Y}} = \frac{1}{n} 1_n^T \hat{Y}$, $\bar{Y} = \frac{1}{n} 1_n^T Y$, and
$$r_{\hat{Y}Y} = \frac{\sum (\hat{Y_i} - \bar{\hat{Y}})(Y_i - \bar{Y})}{\sqrt{\sum (\hat{Y_i} - \bar{\hat{Y}})^2}\sqrt{\sum (Y_i - \bar{Y})^2}}$$

ⓐ Use assumption $1_n \in im(X)$ to show that $\bar{\hat{Y}} = \bar{Y}$

$1_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ of n rows, $\in im(X)$

Show that $\bar{\hat{Y}} = \bar{Y} \rightarrow \frac{1}{n} 1_n^T \hat{Y} = \frac{1}{n} 1_n^T Y$

$\frac{1}{n} 1_n^T \hat{Y} \Rightarrow \frac{1}{n}$ cancels from both sides

$\rightarrow 1_n^T \hat{Y} \Rightarrow$ $1_n^T \in im(X)$, so $1_n^T = Xv$ for some vector v

$\rightarrow (Xv)^T \hat{Y} = v^T X^T \hat{Y} = v^T X^T X (X^T X)^{-1} X^T Y$ $(\hat{Y} = HY)$

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

next page →

3, continued

(b) Use (a) to show that $\sum(\hat{Y_i}-\bar{\hat{Y}})(Y_i-\bar{Y}) = Y^T(H-\frac{1}{n}1_n1_n^T)Y$
will $r^2_{\hat{Y},Y}$ ever be smaller than 0? why?

From (a), $\bar{Y}=\bar{\hat{Y}}$, so we have

$$\left[\sum(\hat{Y_i}-\bar{Y})(Y_i-\bar{Y})\right]^2 = Y^T(H-\tfrac{1}{n}1_n1_n^T)Y$$

we have $J=1_n1_n^T$

$$\sum(\hat{Y_i}-\bar{Y})(Y_i-\bar{Y})\cdot\sum(\hat{Y_i}-\bar{Y})(Y_i-\bar{Y}) = (SSR)(SSTo)$$
$$= [Y^T(H-\tfrac{1}{n}J)Y][Y^T(I-\tfrac{1}{n}J)Y]$$
$$= Y^T(H-\tfrac{1}{n}J)YY^T(I-\tfrac{1}{n}J)Y$$
$$= Y^T(H-\tfrac{1}{n}J)(I-\tfrac{1}{n}J)Y$$
$$= Y^T(HI-\tfrac{1}{n}HJ-\tfrac{1}{n}JI+(\tfrac{1}{n}J)^2)Y$$
$$= Y^T(H-\tfrac{1}{n}J-\cancel{\tfrac{1}{n}J}+\cancel{\tfrac{1}{n}J})Y \qquad HJ=J \text{ and } (\tfrac{1}{n}J)^2=\tfrac{1}{n}J$$
$$= Y^T(H-\tfrac{1}{n}J)Y$$

$r_{\hat{Y},Y}$ will always be ≥ 0 because it's the square root of SSR (always ≥ 0), and denominator is squared so also always ≥ 0. So, $r_{\hat{Y},Y}$ always ≥ 0

(c) Show that $R^2 = r^2_{\hat{Y},Y}$

$r_{\hat{Y},Y} = \frac{\sum(\hat{Y_i}-\bar{Y})(Y_i-\bar{Y})}{\sqrt{\sum(\hat{Y_i}-\bar{Y})^2}\sqrt{\sum(Y_i-\bar{Y})^2}}$ and from (a), $\bar{\hat{Y}}=\bar{Y}$

so $(r_{\hat{Y},Y})^2 = \dfrac{\left(\sum(\hat{Y_i}-\bar{Y})(Y_i-\bar{Y})\right)^2}{\sum(\hat{Y_i}-\bar{Y})^2\sum(Y_i-\bar{Y})^2}$

and from (b), $\sum(\hat{Y_i}-\bar{Y})(Y_i-\bar{Y}) = Y^T(H-\frac{1}{n}J)Y$

$$= \frac{[Y^T(H-\frac{1}{n}J)Y]^2}{[Y^T(H-\frac{1}{n}J)Y][Y^T(I-\frac{1}{n}J)Y]} \leftarrow \text{sub matrix form, SSR \& SSTo}$$

$$= \frac{Y^T(H-\frac{1}{n}J)Y}{Y^T(I-\frac{1}{n}J)Y} \Rightarrow \frac{SSR}{SSTo} = 1-\frac{SSE}{SSTo} = R^2$$

2131 HW6

Orly Olbum

Question 4

(a) Suppose you regress steam (Y) onto fat (X1) and glycerine (X2).

(i) Write down the model you are assuming when performing this regression (i.e. what is the mean and variance model). Provide an interpretation for the coefficients in the mean model.

```r
model = lm(steam ~ fat + glycerine, data = steam)
summary(model)
```

```
##
## Call:
## lm(formula = steam ~ fat + glycerine, data = steam)
##
## 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
```

```r
summary(aov(model))
```

```
##             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
```

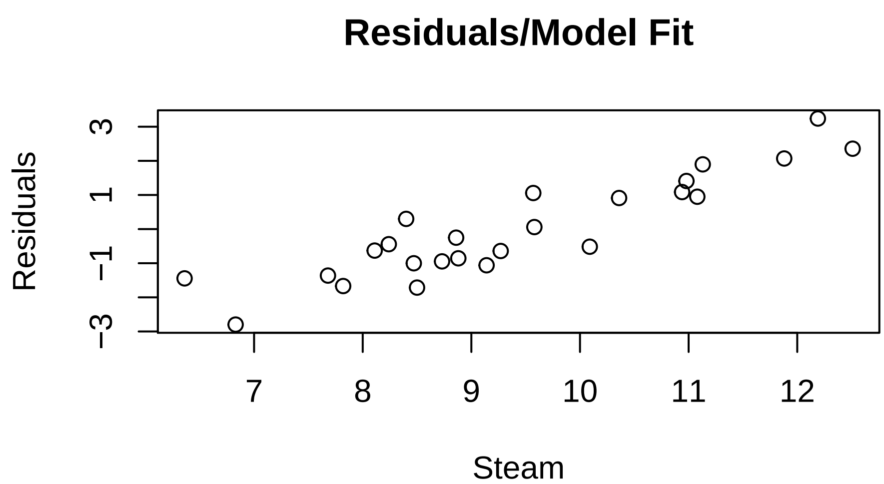
For a first order linear regression model, we assume: - the mean model is linear in both X variables - the regression function is a plane - the regression function is linear in both X variables - the association between Y and one of the X variables does not depend on the other X variable - we assume independent residuals (and normally distributed)

The fitted linear model above shows an equation of: steam = 4.625 + 1.728xfat + -6.628xglycerine, meaning for every unit increase in steam, there is a 1.728 unit increase in fat and a 6.628 unit decrease in glycerine. At alpha = 0.05, neither of these variables prove to be significant as linear predictors for steam output. This leaves us with an underfit model, because we are leaving out variable that could explain the variation in steam.

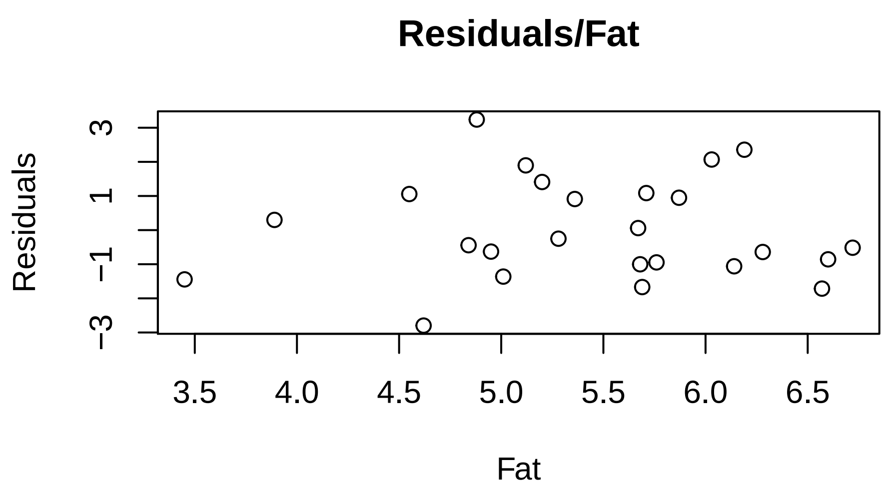
The variance model (ANOVA output) also does not show significance at alpha = 0.05, but it does show us that there is more variation due to the fat variable than the glycerine variable in steam.

(ii) In separate plots, plot e-hat as a function of Y-hat, fat and glycerine. Do you see any evidence that the mean or variance model is incorrect?

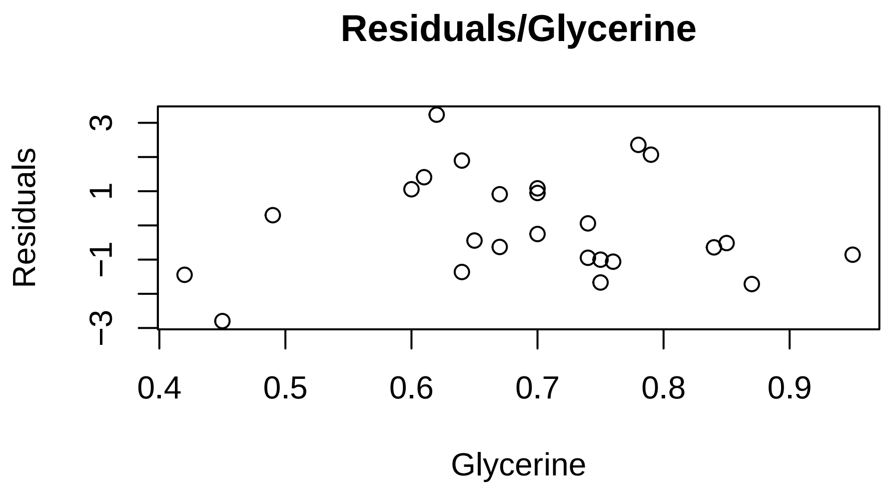
```r
model.res = resid(model)
plot(steam$steam, model.res, ylab = "Residuals", xlab = "Steam", main="Residuals/Model Fit")
```




```r
plot(steam$fat, model.res, ylab = "Residuals", xlab = "Fat", main="Residuals/Fat")
```



```r
plot(steam$glycerine, model.res, ylab = "Residuals", xlab = "Glycerine", main="Residuals/Glycerine")
```



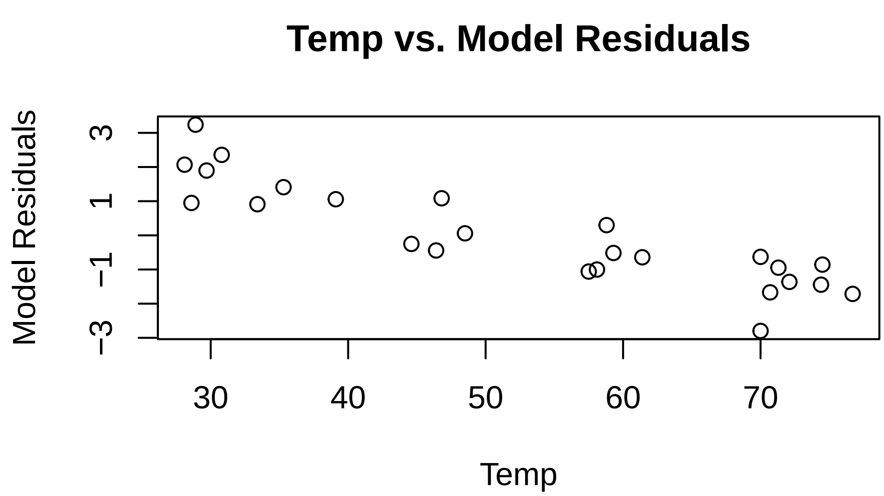
The first plot shows a strong positive association between steam and the model residuals. Because the R^2 of the model is very low, this relationship indicates a potentially poor model. If the R^2 was higher, the dependent variable's variation would be explained more by the independent variables. This is not the case here. The second plot and third plots show no (obvious) association between the variables and the model residuals, indicating a correct model.

(iii) Consider the null hypothesis that the coefficients for both fat and glycerine are 0. At a significance level of alpha = 0.05, what do you conclude about these coefficients?

At alpha = 0.05, we fail to reject the hypothesis of coefficients being 0 and conclude that we do not have evidence to support fat and glycerine being significant predictors for steam.

(iv) Plot the variable "temp" against the residuals from this regression. What can you conclude from this plot?

```r
plot(steam$temp, model.res, ylab = "Model Residuals", xlab = "Temp", main = "Temp vs. Model Residuals")
```



We see in this plot that there is a negative association between temp and the model residuals.

(b) Now regress steam (Y) onto fat (X1), glycerine (X2) and temp (X3).

```
model2 = lm(steam ~ fat + glycerine + temp, data = steam)
summary(model2)
```

```
##
## Call:
## lm(formula = steam ~ fat + glycerine + temp, data = steam)
##
## 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
```

```
summary(aov(model2))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## fat          1   9.37    9.37  22.042 0.000124 ***
## glycerine    1   1.83    1.83   4.304 0.050512 .
## temp         1  43.69   43.69 102.776 1.52e-09 ***
## Residuals   21   8.93    0.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i) Consider the null hypothesis that the coefficients for both fat and glycerine are 0. At a significance level of alpha = 0.05, what do you conclude about these coefficients?

From the mean model, we can see that both fat and glycerine still are not significant predictors of steam, but the variance model shows that with temp in the mix as another independent variable, fat is a significant predictor. Glycerine remains insignificant, at alpha = 0.05.

(ii) Why are the P values from this test so much smaller than those from part (a)?

The R^2 of this model has risen by a lot, indicating that the independent variables are now accounting for much more of the variation in the dependent variable (steam) than in the prior model. Because of this, the p-values will be lower, because they are dependent on the F-values, which in turn depends on the mean squared error (and therefore variance). We see that in both the mean and variance models, temp is a significant predictor for steam. A plot of the residuals for this model against steam will show no obvious association, telling us that the variation in steam is explained by the model itself (moreso than the first model).

```
model2.res = resid(model2)
plot(steam$steam, model2.res, ylab = "Residuals", xlab = "Steam", main="Residuals/Model 2 Fit")
```