STAT 420

1. Consider the population of high school graduates who were admitted to a particular university during a ten-year time period and who completed at least the first year of coursework after being admitted. We are interested in investigating how well Y, the first year grade point average (GPA), can be predicted by using the following quantities with $n = 20$ students:

X_1 = the score on the mathematics part of the SAT (SATmath)
X_2 = the score on the verbal part of the SAT (SATverbal)
X_3 = the grade point average of all high school mathematics courses (HSmath)
X_4 = the grade point average of all high school English courses (HSenglish)

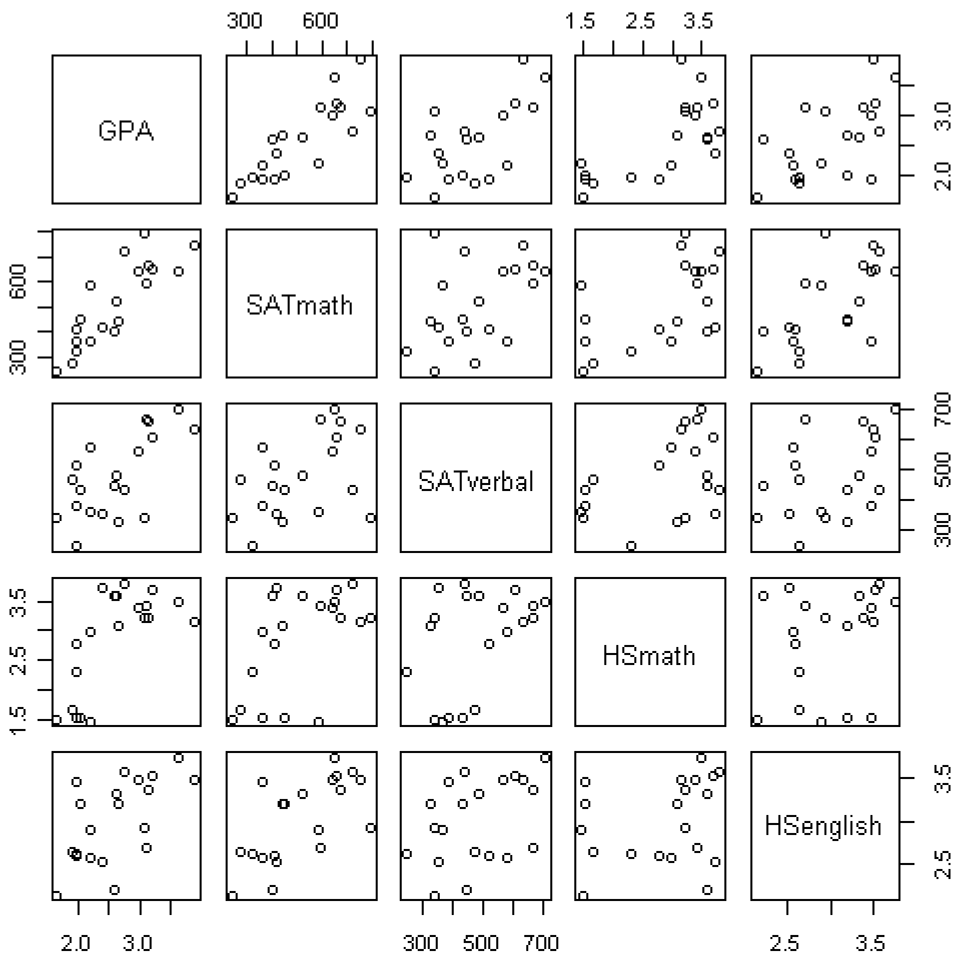
Consider the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, 20,$$

where ε_i's are independent $N(0, \sigma^2)$ random variables.

```
> fit = lm(GPA ~ SATmath + SATverbal + HSmath + HSenglish)
> summary(fit)

Call:
lm(formula = GPA ~ SATmath + SATverbal + HSmath + HSenglish)

Residuals:
      Min         1Q    Median         3Q        Max
-0.443283 -0.128374  0.002571  0.133996  0.538996

Coefficients:
               Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   0.1615496    0.4375321     0.369    0.71712
SATmath       0.0020102    0.0005844     3.439    0.00365  **
SATverbal     0.0012522    0.0005515     2.270    0.03835  *
HSmath        0.1894402    0.0918680     2.062    0.05697  .
HSenglish     0.0875637    0.1764963     0.496    0.62700
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2685 on 15 degrees of freedom
Multiple R-squared: 0.8528,     Adjusted R-squared: 0.8135
F-statistic: 21.72 on 4 and 15 DF,  p-value: 4.255e-06
```

```
> pairs(GPA ~ SATmath+SATverbal+HSmath+HSenglish)
```



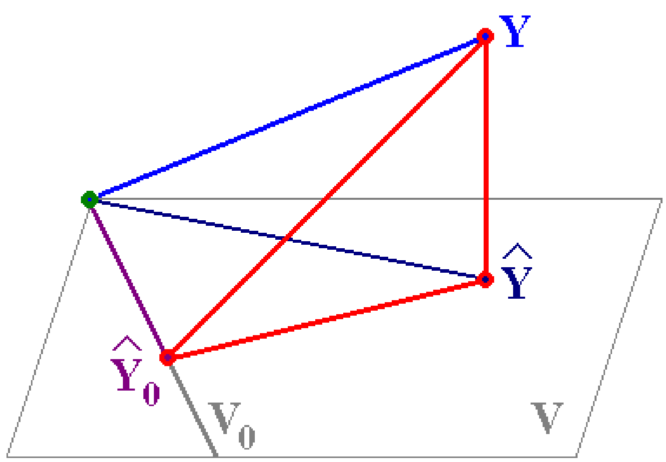
GPA vs. SATmath, GPA vs. SATverbal, and GPA vs. HSenglish suggest a linear relationship. GPA vs. HSmath does not look linear. Either additional higher-order terms in HSmath (for example, the second-order term) are needed, or the values of one or more variables should be transformed before analysis.

```
> fit2 = lm(GPA ~ SATmath + SATverbal)
> fit3 = lm(GPA ~ SATmath + HSmath)
> sum(fit$residuals^2)
[1] 1.081499
> sum(fit2$residuals^2)
[1] 1.388384
> sum(fit3$residuals^2)
[1] 1.528179
```

Suppose we wish to test the claim that SATverbal and HSenglish do not affect the first year GPA. That is, we wish to test $H_0: \beta_2 = \beta_4 = 0$ vs. H_a: at least one of β_2 and β_4 is significantly different from 0. Perform the test at a 10% level of significance.

Full Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$.

Null Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i$.



$$V = \{ a_0 \mathbf{1} + a_1 \boldsymbol{x_1} + a_2 \boldsymbol{x_2} + a_3 \boldsymbol{x_3} + a_4 \boldsymbol{x_4}, \quad a_0, a_1, a_2, a_3, a_4 \in \mathbf{R} \},$$

$$\dim(V) = 5.$$

$$V_0 = \{ a_0 \mathbf{1} + a_1 \boldsymbol{x_1} + a_3 \boldsymbol{x_3}, \quad a_0, a_1, a_3 \in \mathbf{R} \}, \qquad \dim(V_0) = 3.$$

Numerator d.f. $= \dim(V) - \dim(V_0) = 5 - 3 = \mathbf{2}$.

Denominator d.f. $= n - \dim(V) = 20 - 5 = \mathbf{15}$.

	SS	*DF*	*MS*	*F*
Diff.	$\text{SSResid}_{\text{null}} - \text{SSResid}_{\text{full}}$	$\dim(\text{V}) - \dim(\text{V}_0)$	…	…
Full	$\text{SSResid}_{\text{full}}$	$n - \dim(\text{V})$	…	
Null	$\text{SSResid}_{\text{null}}$	$n - \dim(\text{V}_0)$		

	SS	*DF*	*MS*	*F*	
Diff.	0.44668	2	0.22334	**3.098**	← Test Statistic
Full	1.08150	15	0.07210		
Null	1.52818	17			

Critical Value: $F_{0.10}(2, 15) = \mathbf{2.70}$.

Decision: **Reject $\mathbf{H_0}$.**

```
> anova(fit3,fit)
Analysis of Variance Table

Model 1: GPA ~ SATmath + HSmath
Model 2: GPA ~ SATmath + SATverbal + HSmath + HSenglish
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     17 1.5282
2     15 1.0815  2   0.44668 3.0976 0.0748 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Suppose we wish to test the claim that high school performance does not affect the first year GPA. That is, we wish to test $H_0: \beta_3 = \beta_4 = 0$ vs. H_a: at least one of β_3 and β_4 is significantly different from 0. Perform the test at a 10% level of significance.

Full Model: $\quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i.$

Null Model: $\quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$

$$V = \{ a_0 \mathbf{1} + a_1 \boldsymbol{x_1} + a_2 \boldsymbol{x_2} + a_3 \boldsymbol{x_3} + a_4 \boldsymbol{x_4}, \quad a_0, a_1, a_2, a_3, a_4 \in \mathbf{R} \},$$

$$\dim(V) = 5.$$

$$V_0 = \{ a_0 \mathbf{1} + a_1 \boldsymbol{x_1} + a_2 \boldsymbol{x_2}, \quad a_0, a_1, a_2 \in \mathbf{R} \}, \qquad \dim(V_0) = 3.$$

Numerator d.f. $= \dim(V) - \dim(V_0) = 5 - 3 = \mathbf{2}$.

Denominator d.f. $= n - \dim(V) = 20 - 5 = \mathbf{15}$.

	SS	*DF*	*MS*	*F*	
Diff.	0.30688	2	0.15344	**2.128**	← Test Statistic
Full	1.08150	15	0.07210		
Null	1.38838	17			

Critical Value: $F_{0.10}(2, 15) = \mathbf{2.70}$.

Decision: **Do NOT Reject H_0.**

```
> anova(fit2,fit)
Analysis of Variance Table

Model 1: GPA ~ SATmath + SATverbal
Model 2: GPA ~ SATmath + SATverbal + HSmath + HSenglish
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     17 1.3884
2     15 1.0815  2   0.30689 2.1282 0.1536
```

Test $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ vs. H_a: at least one of β_2, β_3, and β_4 is significantly different from 0. Use a 5% level of significance.

Full Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$.

Null Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \qquad\qquad + \varepsilon_i$.

```
> fit4 = lm(GPA ~ SATmath)
> sum(fit4$residuals^2)
[1] 2.044274
```

$$V = \{ a_0 \mathbf{1} + a_1 \boldsymbol{x}_1 + a_2 \boldsymbol{x}_2 + a_3 \boldsymbol{x}_3 + a_4 \boldsymbol{x}_4, \quad a_0, a_1, a_2, a_3, a_4 \in \mathbf{R} \},$$

$\dim(V) = 5$.

$$V_0 = \{ a_0 \mathbf{1} + a_1 \boldsymbol{x}_1, \quad a_0, a_1 \in \mathbf{R} \}, \qquad \dim(V_0) = 2.$$

Numerator d.f. = $\dim(V) - \dim(V_0)$ = $5 - 2$ = **3**.

Denominator d.f. = $n - \dim(V)$ = $20 - 5$ = **15**.

	SS	*DF*	*MS*	*F*	
Diff.	0.96277	3	0.3209233	**4.451**	← Test Statistic
Full	1.08150	15	0.07210		
Null	2.04427	18			

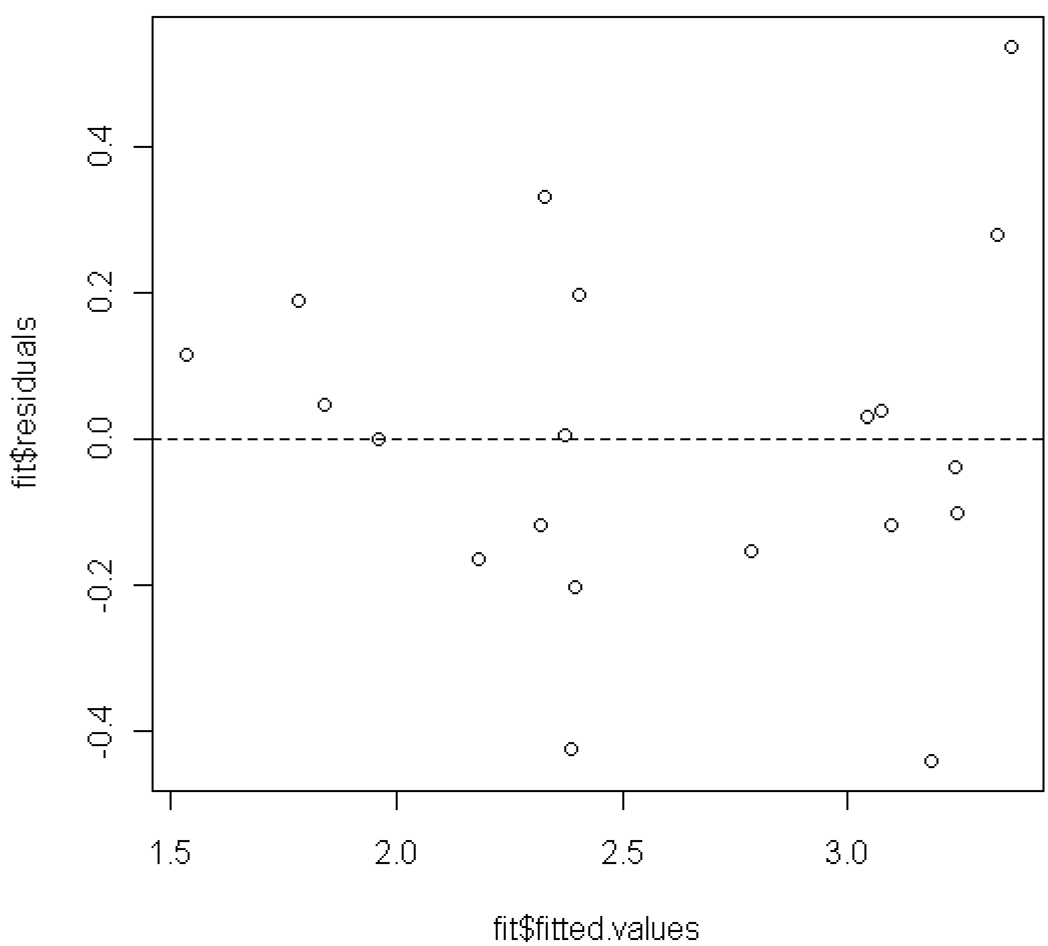
Critical Value: $F_{0.05}(3, 15) = \mathbf{3.29}$.

Decision: **Reject H_0.**

```
> anova(fit4,fit)
Analysis of Variance Table

Model 1: GPA ~ SATmath
Model 2: GPA ~ SATmath + SATverbal + HSmath + HSenglish
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     18 2.0443
2     15 1.0815  3   0.96277 4.4511 0.01994 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

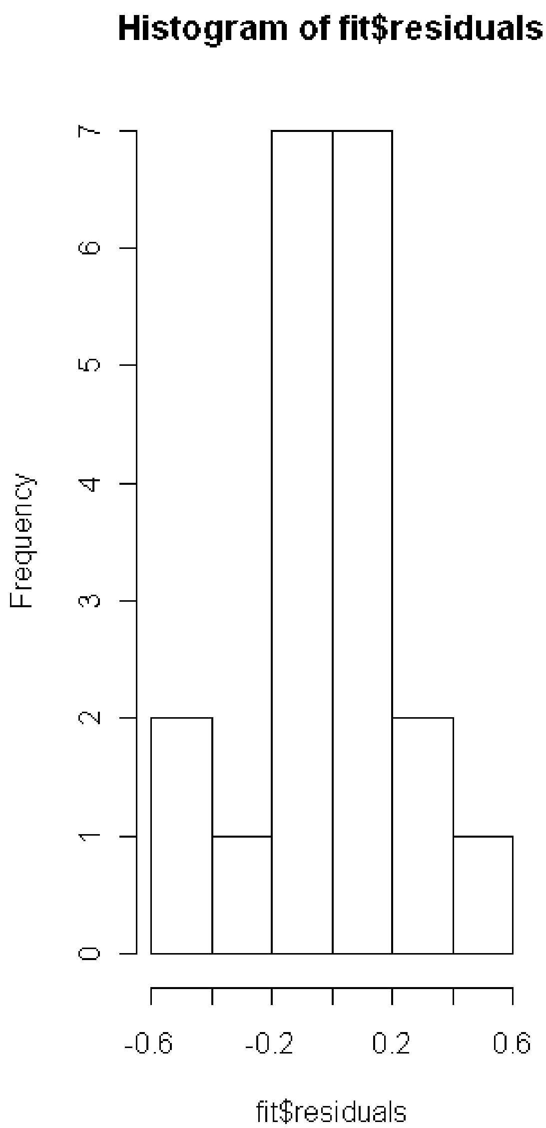
_ _

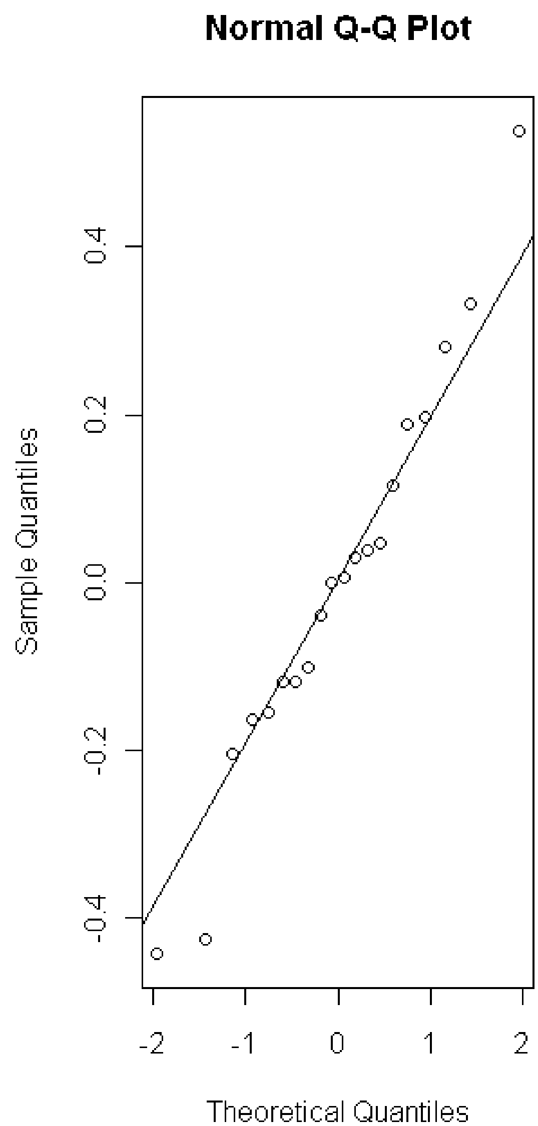
```
> plot(fit$fitted.values,fit$residuals)
> abline(h=0,lty=2)
```



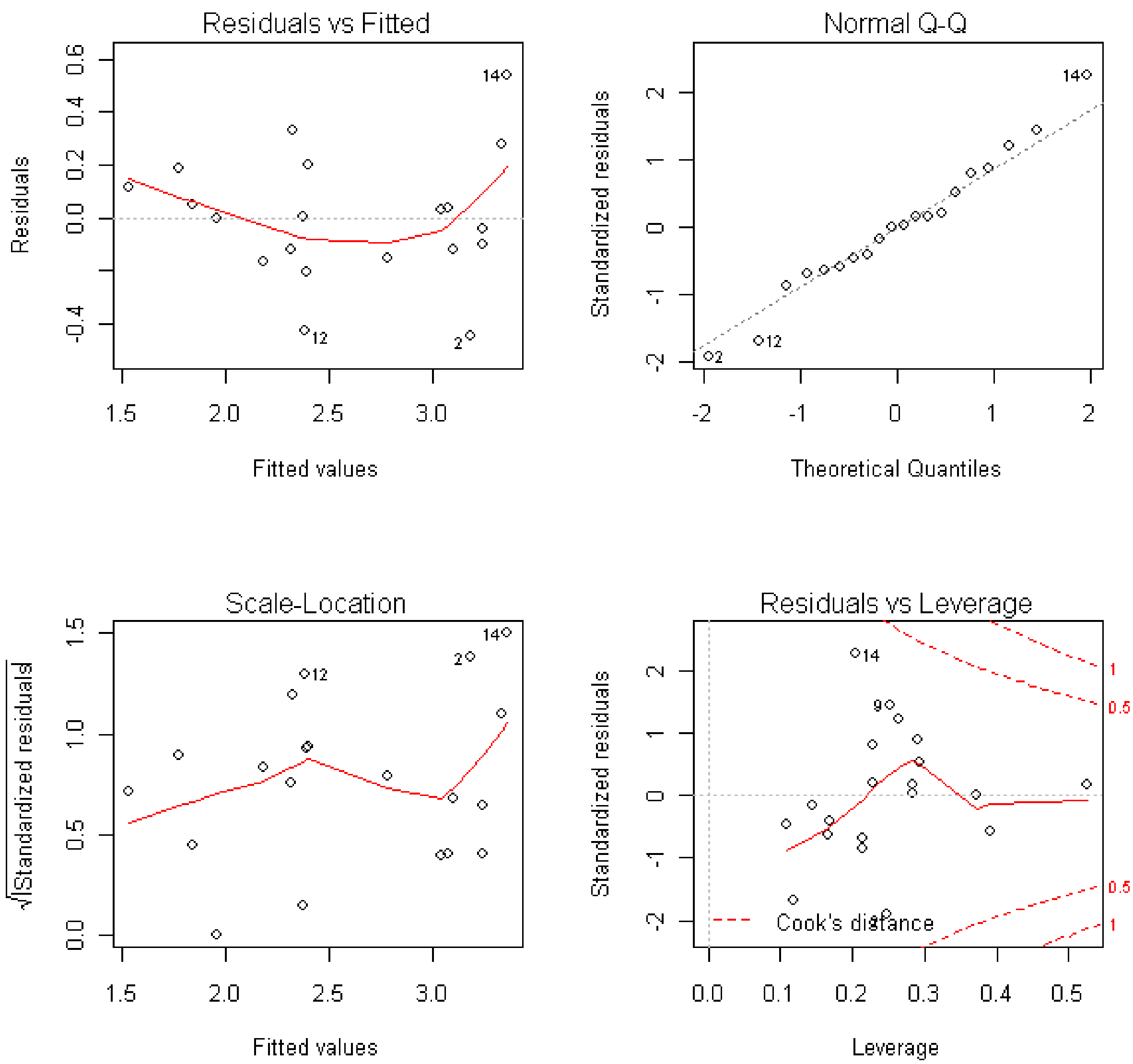
The residuals versus the fitted values plot suggests that the variance σ^2 is not constant.

```
> par(mfrow=c(1,2))
> hist(fit$residuals)
> qqnorm(fit$residuals)
> qqline(fit$residuals)
```





```
> par(mfrow=c(2,2))
> plot(fit)
```



2. The following data are indexed prices of gold and copper over a 10-year period. Assume that the indexed values constitute a random sample from a bivariate normal distribution.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})\cdot(y-\bar{y})$	$(y-\bar{y})^2$
76	80	16	12	256	192	144
62	68	2	0	4	0	0
70	73	10	5	100	50	25
59	60	– 1	– 8	1	8	64
53	64	– 7	– 4	49	28	16
54	68	– 6	0	36	0	0
55	65	– 5	– 3	25	15	9
58	62	– 2	– 6	4	12	36
57	67	– 3	– 1	9	3	1
56	73	– 4	5	16	– 20	25
600	680	0	0	500	288	320

a) Test for the existence of linear relationship between the indexed prices of the two metals. That is, test $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$. Use a 5% level of significance.

$\bar{x} = \frac{600}{10} = 60.$ $\quad \bar{y} = \frac{680}{10} = 68.$ $\quad r = \frac{288}{\sqrt{500}\sqrt{320}} = \mathbf{0.72}.$

Test Statistic: $\quad t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.72 \cdot \sqrt{10-2}}{\sqrt{1-0.72^2}} = \mathbf{2.9345}.$

Rejection Region: Rejects H_0 if $t < -t_{0.025}(8\text{ df})$ or $t > t_{0.025}(8\text{ df})$.

$\pm t_{0.025}(8\text{ df}) = \pm 2.306.$ **Reject H_0.**

Since $t_{0.01}(8\text{ df}) = 2.896 < 2.9345 < 3.355 = t_{0.005}(8\text{ df})$,

$2 \times 0.005 = 0.01 < \text{p-value} < 0.02 = 2 \times 0.01.$ $\quad (\text{p-value} \approx 0.01887)$

OR

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.72}{1-0.72}\right) = 0.907645.$$

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0}{1-0}\right) = 0,$

$$\sigma_W^2 = \frac{1}{n-3} = \frac{1}{7}.$$

Test Statistic: $z = \frac{W-\mu_W}{\sigma_W} = \frac{0.907645-0}{\sqrt{1/7}} = \mathbf{2.40}.$

Rejection Region: Rejects H_0 if $z < -z_{0.025}$ or $z > z_{0.025}$.

$\pm z_{0.025} = \pm 1.960.$ **Reject H_0.**

P-value = $2 \times P(Z > 2.40) = 2 \times 0.0082 = 0.0164.$

b) Is there enough evidence to conclude $\rho > 0.40$. That is, test $H_0: \rho = 0.40$ vs. $H_1: \rho > 0.40$. Use a 5% level of significance. What is the p-value of this test?

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.72}{1-0.72}\right) = 0.907645.$$

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0.40}{1-0.40}\right) = 0.423649,$

$$\sigma_W^2 = \frac{1}{n-3} = \frac{1}{7}.$$

Test Statistic: $z = \frac{W-\mu_W}{\sigma_W} = \frac{0.907645-0.423649}{\sqrt{1/7}} = \mathbf{1.2805}.$

Rejection Region: Rejects H_0 if $z > z_{0.05}$.

$z_{0.05} = 1.645.$ **Do NOT Reject H_0.**

P-value = right tail = $P(Z > 1.2805) = \mathbf{0.10}.$