

HW3: 563 Regression Methods

4

(a)

```
> mod<-read.table("4.txt",header=T)
> mod.lm<-lm(mod$Expe~mod$Earn)
> summary(mod.lm)
```

Call:

```
lm(formula = mod$Expe ~ mod$Earn)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.3281	-0.1862	-0.1443	0.1862	0.5454

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.62228	0.50670	1.228	0.265389
mod\$Earn	0.72646	0.09746	7.454	0.000301 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3264 on 6 degrees of freedom

Multiple R-Squared: 0.9025, Adjusted R-squared: 0.8863

F-statistic: 55.56 on 1 and 6 DF, p-value: 0.0003005

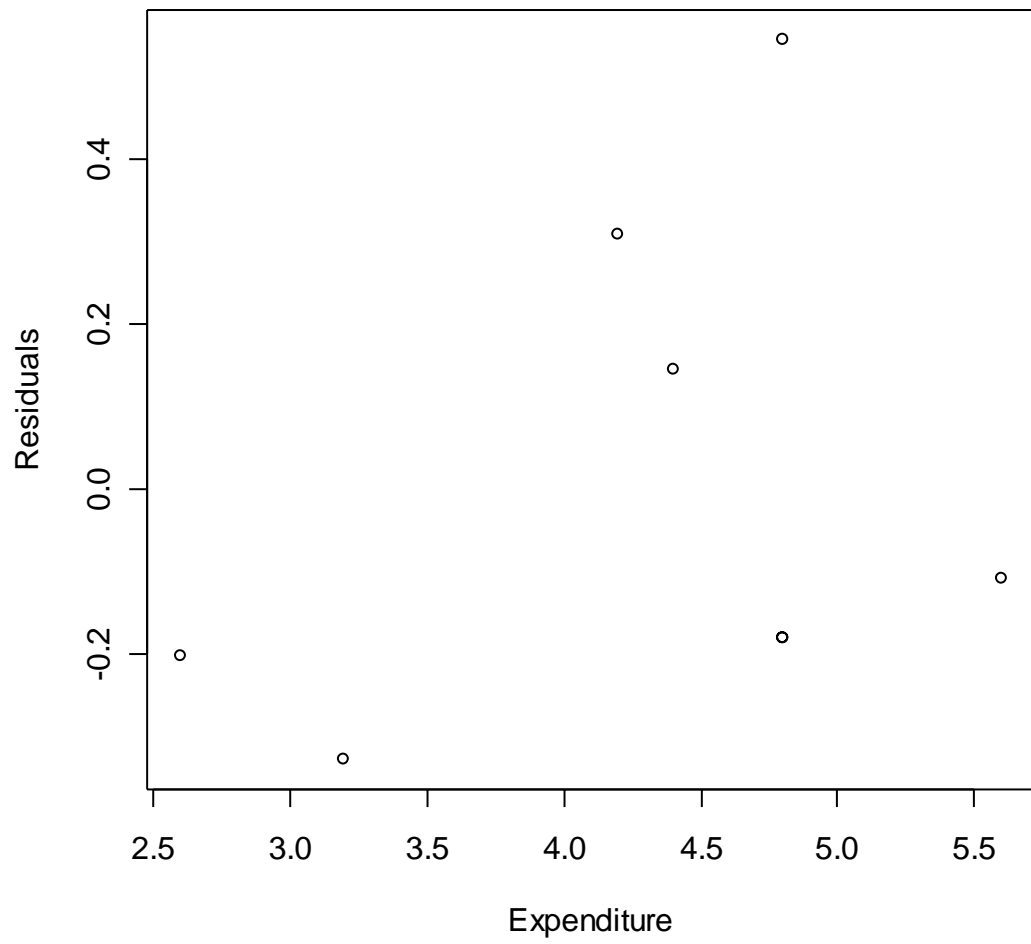
$\alpha = 0.62228$, $\beta = 0.72646$, $\sigma^2 = 0.1065$

```
> list(mod.lm$resi)
```

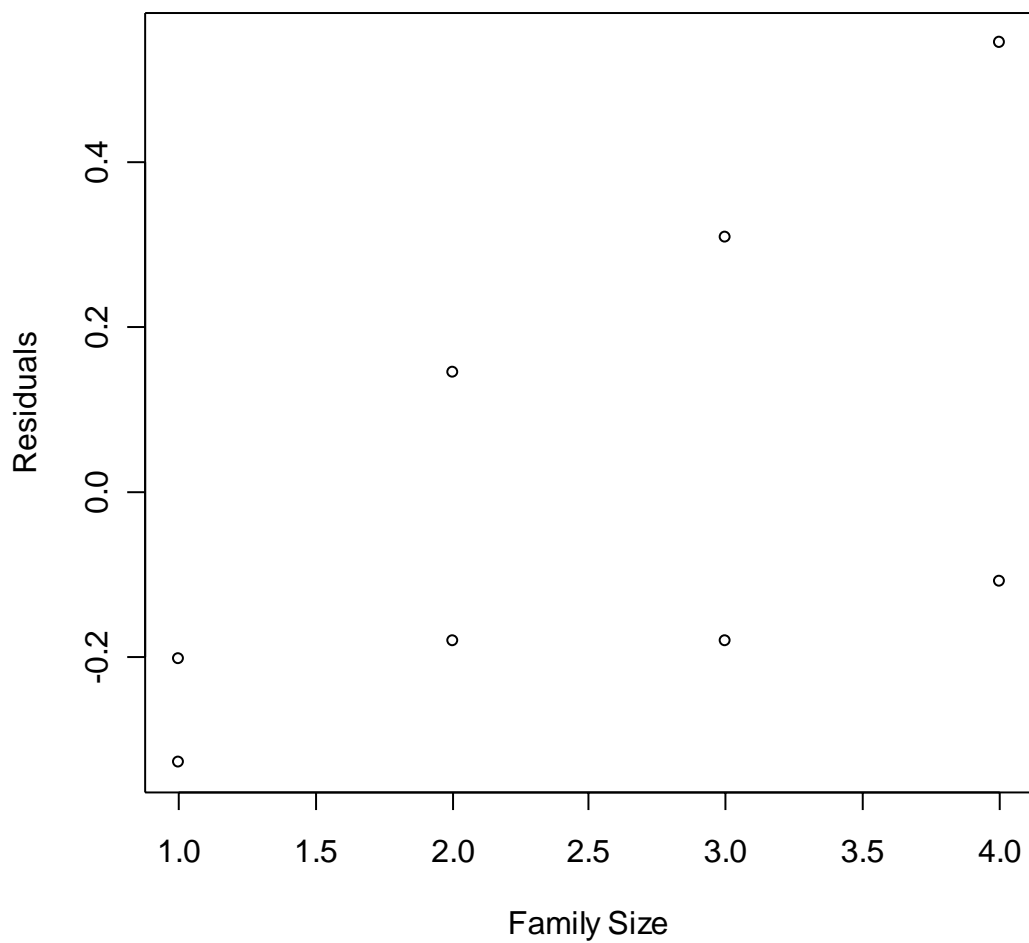
```
[[1]]
```

1	2	3	4	5	6	7	8
-0.2016713	-0.3281337	0.1454039	-0.1810585	0.3086351	-0.1810585	0.5454039	-0.1075209

```
> plot(mod$Expenditure,mod.lm$residuals,xlab="Expenditure",ylab="Residuals")
```



```
> plot(mod$Size,mod.lm$residuals,xlab="Family Size",ylab="Residuals")
```



#It seems that the error terms do not have constant variance, especially there exists certain pattern when it refers to #residuals VS the family size.

```
(b)
> mod$Expenditure<-mod$Expenditure/mod$Size
> mod$Earnings<-mod$Earnings/mod$Size
> mod1.lm<-lm(mod$Expenditure~mod$Earnings)
> summary(mod1.lm)
```

Call:
lm(formula = mod\$Expenditure ~ mod\$Earnings)

Residuals:

Min	1Q	Median	3Q	Max
-0.12604	-0.08438	0.01042	0.07396	0.12604

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.19896	0.11114	1.79	0.124
mod\$Earnings	0.75833	0.04396	17.25	2.43e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1077 on 6 degrees of freedom

Multiple R-Squared: 0.9802, Adjusted R-squared: 0.9769

F-statistic: 297.5 on 1 and 6 DF, p-value: 2.432e-06

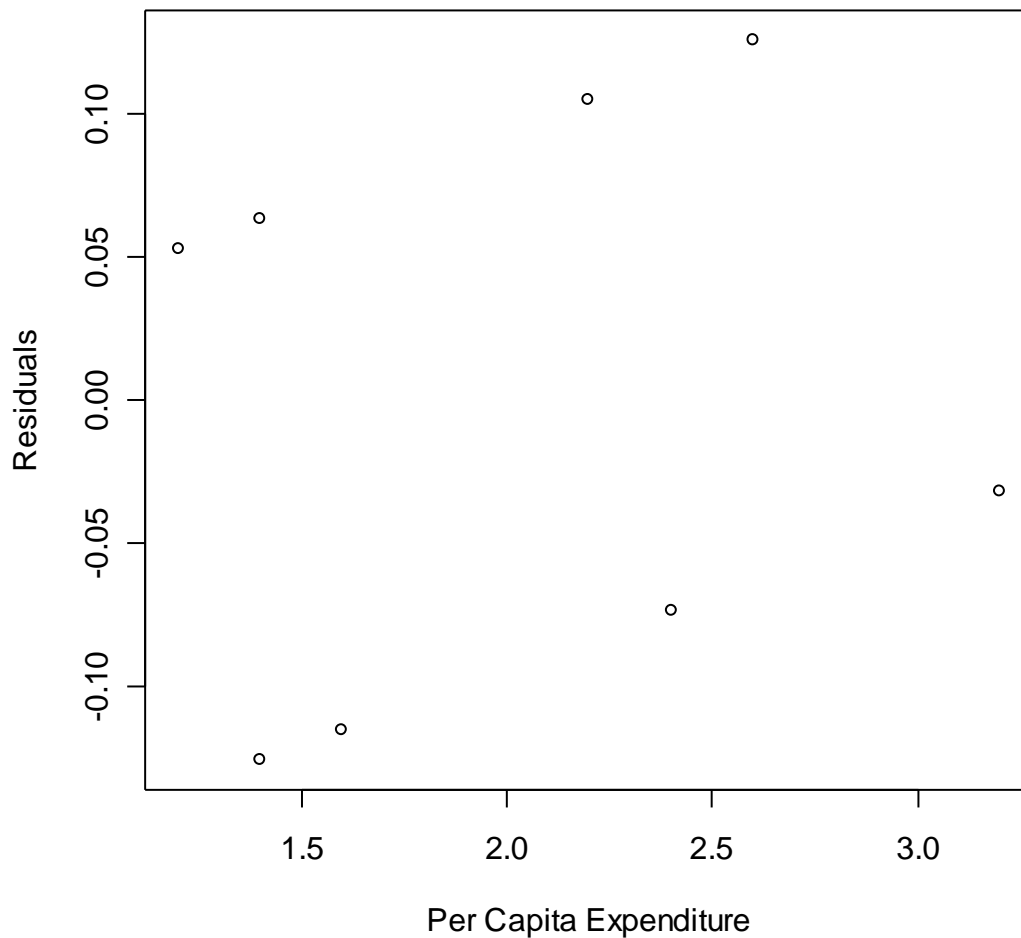
$\alpha = 0.19896$, $\beta = 0.75833$, $\sigma^2 = 0.0116$

> list(mod1.lm\$resi)

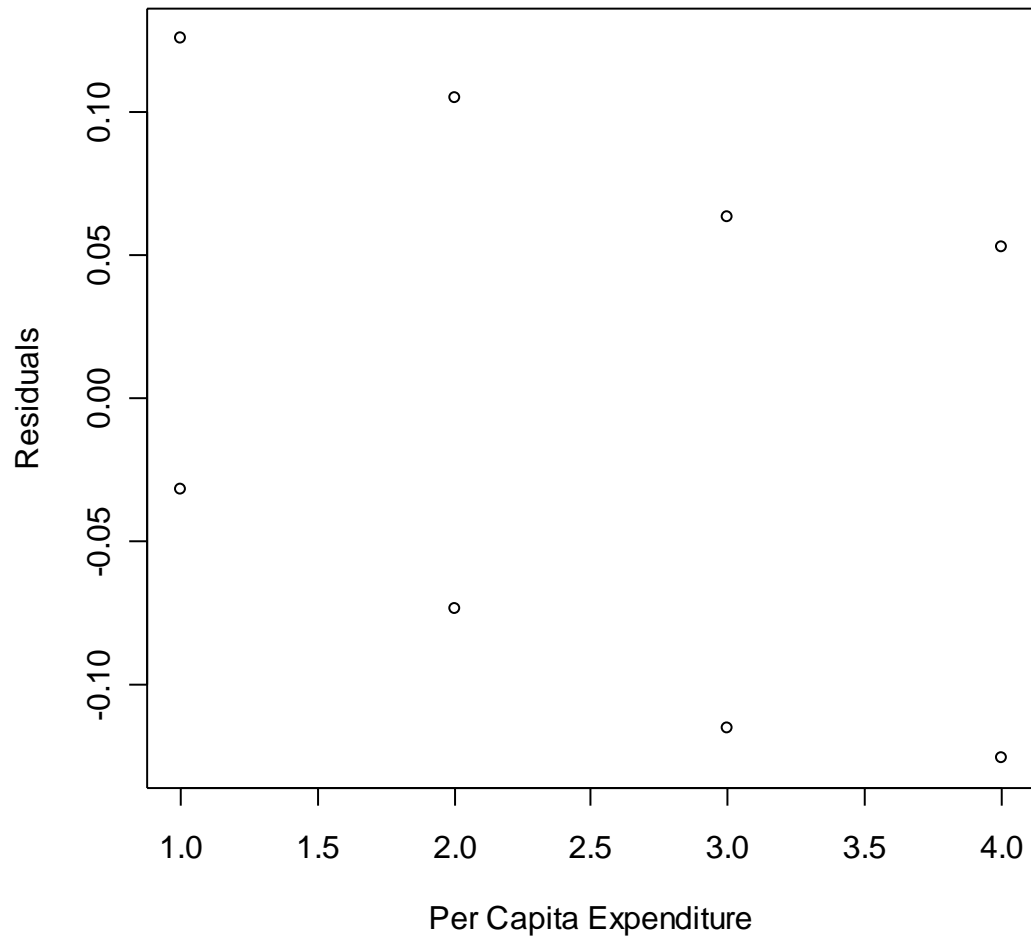
[[1]]

1	2	3	4	5	6	7	8
0.12604167	-0.03229167	0.10520833	-0.07395833	0.06354167	-0.11562500	0.05312500	-0.12604167

> plot(mod\$Expenditure,mod1.lm\$residuals,xlab="Per Capita Expenditure",ylab="Residuals")



```
> plot(mod$Size,mod1.lm$residuals,xlab="Per Capita Expenditure",ylab="Residuals")
```



#after the transformation of x and y, it seems that the error terms have constant variance.

5

```
> mod<-read.table("5.txt",header=T)
```

```
> mod
```

	y1	y2	y3	y4	x	x4
1	8.04	9.14	7.46	6.58	10	8
2	6.95	8.14	6.77	5.76	8	8
3	7.58	8.74	12.74	7.71	13	8
4	8.81	8.77	7.11	8.84	9	8
5	8.33	9.26	7.81	8.47	11	8
6	9.96	8.10	8.84	7.04	14	8
7	7.24	6.13	6.08	5.25	6	8
8	4.26	3.10	5.39	5.56	4	8
9	10.84	9.13	8.15	7.91	12	8
10	4.82	7.26	6.42	6.89	7	8

```
11 5.68 4.74 5.73 12.50 5 19
> mod1.lm<-lm(mod$y1~mod$x)
> summary(mod1.lm)
```

Call:

```
lm(formula = mod$y1 ~ mod$x)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.92127	-0.45577	-0.04136	0.70941	1.83882

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.0001	1.1247	2.667	0.02573 *
mod\$x	0.5001	0.1179	4.241	0.00217 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

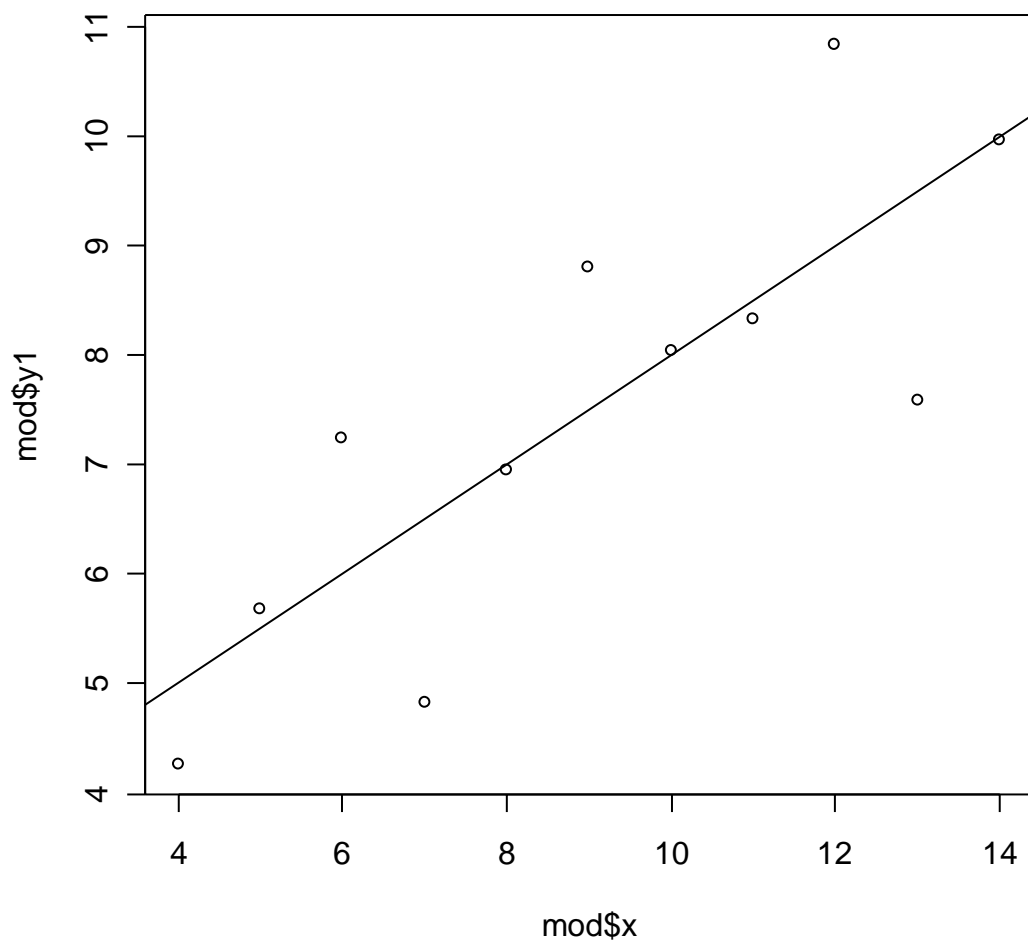
Residual standard error: 1.237 on 9 degrees of freedom

Multiple R-Squared: 0.6665, Adjusted R-squared: 0.6295

F-statistic: 17.99 on 1 and 9 DF, p-value: 0.002170

```
> plot(mod$x,mod$y1)
```

```
> abline(mod1.lm)
```



#obviously linearity is not appropriate for the data, even half of the data are far away from the fitted regression line.

```
> mod2.lm<-lm(mod$y2~mod$x)
> summary(mod2.lm)
```

Call:

```
lm(formula = mod$y2 ~ mod$x)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.9009	-0.7609	0.1291	0.9491	1.2691

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.001	1.125	2.667	0.02576 *
mod\$x	0.500	0.118	4.239	0.00218 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

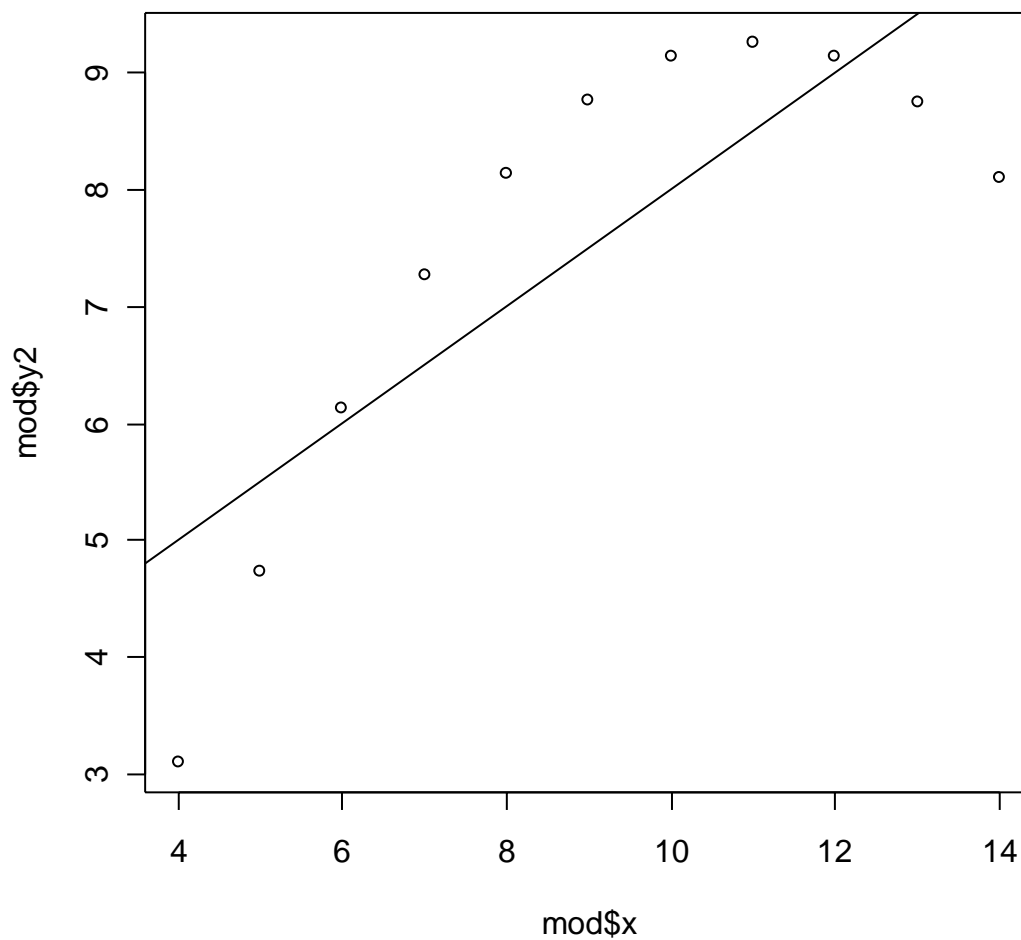
Residual standard error: 1.237 on 9 degrees of freedom

Multiple R-Squared: 0.6662, Adjusted R-squared: 0.6292

F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002179

```
> plot(mod$x,mod$y2)
```

```
> abline(mod2.lm)
```



#the true regression line should be a curvilinear one.

```
> mod3.lm<-lm(mod$y3~mod$x)
```

```
> summary(mod3.lm)
```

Call:

```
lm(formula = mod$y3 ~ mod$x)
```

Residuals:

Min 1Q Median 3Q Max
-1.1586 -0.6146 -0.2303 0.1540 3.2411

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.0025	1.1245	2.670	0.02562 *
mod\$x	0.4997	0.1179	4.239	0.00218 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

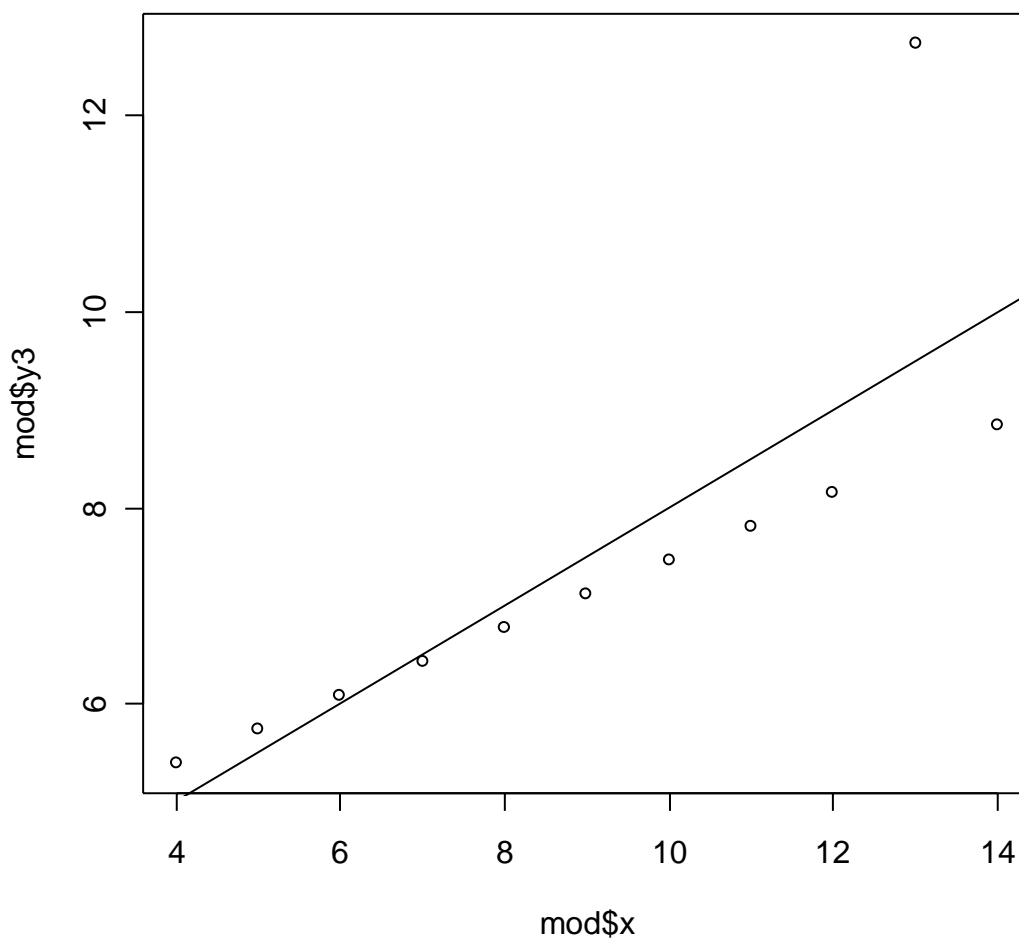
Residual standard error: 1.236 on 9 degrees of freedom

Multiple R-Squared: 0.6663, Adjusted R-squared: 0.6292

F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002176

```
> plot(mod$x,mod$y3)
```

```
> abline(mod3.lm)
```



#generally speaking, the linear regression line is appropriate, but there exists an outlier that pull the fitted regression

#line up. Obviously if we get rid of the outlier point, the result will be perfect.

```
> mod4.lm<-lm(mod$y4~mod$x4)
> summary(mod4.lm)
```

Call:

```
lm(formula = mod$y4 ~ mod$x4)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.751e+00	-8.310e-01	1.258e-16	8.090e-01	1.839e+00

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.0017	1.1239	2.671	0.02559 *
mod\$x4	0.4999	0.1178	4.243	0.00216 **

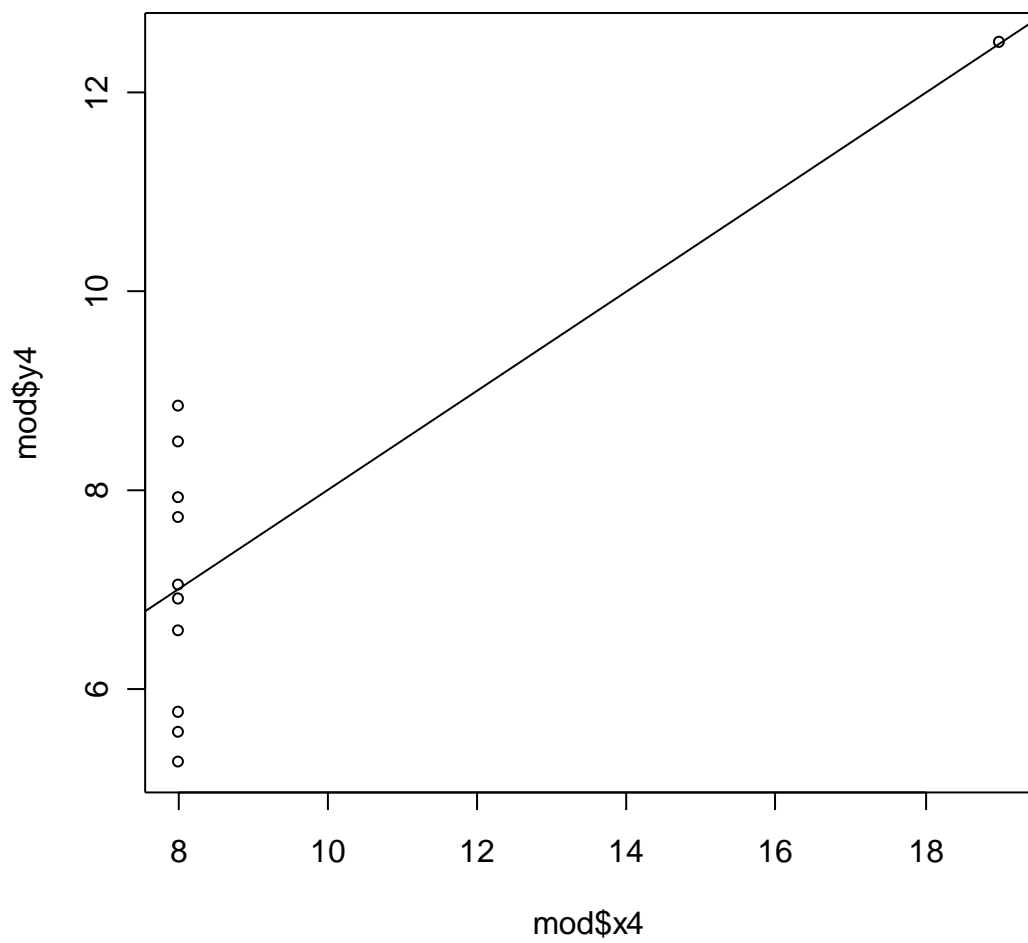
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom

Multiple R-Squared: 0.6667, Adjusted R-squared: 0.6297

F-statistic: 18 on 1 and 9 DF, p-value: 0.002165

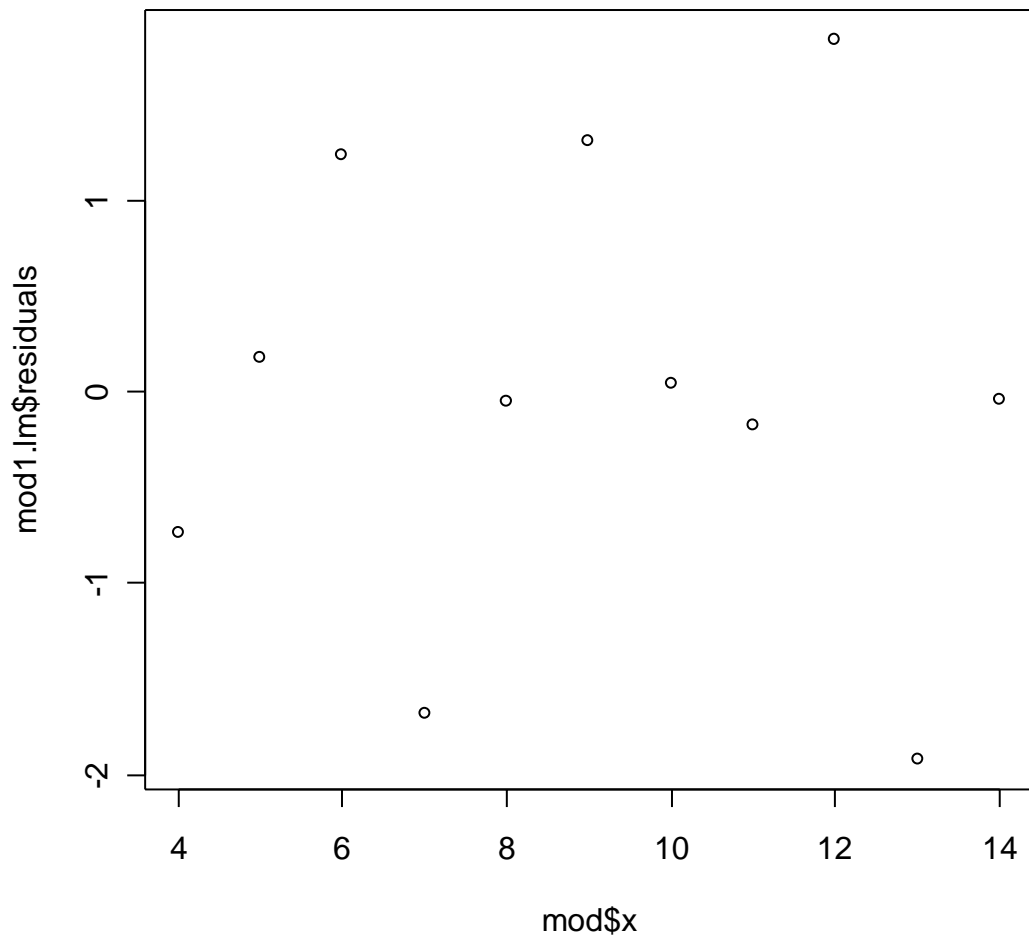
```
> plot(mod$x4,mod$y4)
> abline(mod4.lm)
```



#this figure is weird, that there exists the outlier (19,12.50), after getting rid of it, the fitted regression will be vertical to
 #the x4 coordinate.

#I am to analyze the residuals of mod1 and mod3 after getting rid of the outlier.

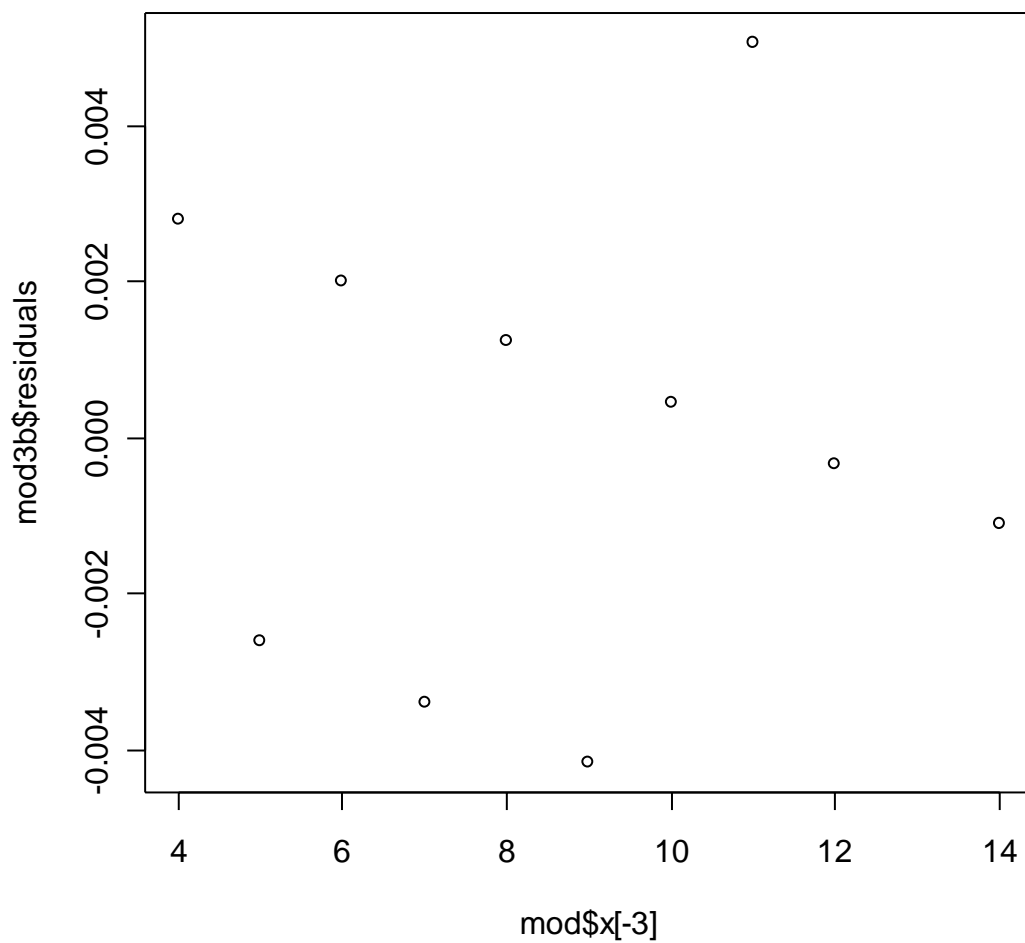
```
> plot(mod$x,mod1.lm$residuals)
```



#variance is not constant, increasing with x1.

```
> mod3b<-lm(mod$y3[-3]~mod$x[-3])
```

```
> plot(mod$x[-3],mod3b$residuals)
```



#the residuals are not independent, because there is pattern in the above Residuals VS X3 plot.

5.23

(a)

```
> data<-read.table("CH05PR04.txt",head=T)
```

```
> X<-cbind(1,data$X)
```

```
> Y<-data$Y
```

```
> coeff<-solve(t(X)%*%X)%*%(t(X)%*%Y)
```

```
> coeff
```

```
      [,1]
```

```
[1,]  9.940
```

```
[2,] -0.245
```

```
> res<-Y-X%*%coeff
```

```
> res
```

```
      [,1]
```

```
[1,] -0.18
```

```
[2,]  0.04
```

```

[3,] 0.26
[4,] 0.08
[5,] -0.20
> J<-matrix(rep(1,times=25),ncol=5)
> SSR<-t(coef)%*%t(X)%*%Y-2*t(Y)%*%J%*%Y
> SSR
      [,1]
[1,] 9.604
> SSE<-t(Y-X)%*%coef)%*%(Y-X)%*%coef)
> SSE
      [,1]
[1,] 0.148
> MSE<-SSE/(5-2)
> MSE<-MSE[1,1]
> B<-MSE*solve(t(X)%*%X)
> B
      [,1] [,2]
[1,] 0.009866667 0.000000000
[2,] 0.000000000 0.000308333
#the estimated variance-covariance matrix of b is the above B matrix.

```

```

> Xh<-c(1,-6)
> Yh<-Xh)%*%coef
> Yh
      [,1]
[1,] 11.41
#the point estimate of  $E(Y_h) = 11.41$  while  $X_h = -6$ 

```

```

> BYh<-t(Xh)%*%B)%*%Xh
> BYh
      [,1]
[1,] 0.02096667
#the estimated variance of  $\hat{Y}_h$  is 0.02096667 while  $X_h = -6$ 

```

(b)
#the sum of the X levels is zero so that transpose of X times X is diagonal, making the computation easy and faster.

```

(c)
> H<-X)%*%solve(t(X)%*%X)%*%t(X)
> H
      [,1] [,2] [,3] [,4] [,5]
[1,] 0.6 0.4 0.2 0.0 -0.2
[2,] 0.4 0.3 0.2 0.1 0.0

```

```
[3,] 0.2 0.2 0.2 0.2 0.2
[4,] 0.0 0.1 0.2 0.3 0.4
[5,] -0.2 0.0 0.2 0.4 0.6
```

(d)

```
> I<-matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),ncol=5)
```

```
> S<-MSE*(I-H)
```

```
> S
```

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.019733333 -0.019733333 -0.009866667 0.000000000 0.009866667
[2,] -0.019733333 0.034533333 -0.009866667 -0.004933333 0.000000000
[3,] -0.009866667 -0.009866667 0.039466667 -0.009866667 -0.009866667
[4,] 0.000000000 -0.004933333 -0.009866667 0.034533333 -0.019733333
[5,] 0.009866667 0.000000000 -0.009866667 -0.019733333 0.019733333
```

#the above matrix is $s^2\{e\}$