Econometric Homework 2

0613404 陸恭葦

Part 1

2-1

(a.)

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
3	4	2	4	2	4
2	2	1	1	0	0
1	3	0	0	1	0
−1	1	−2	4	−1	2
0	0	−1	1	−2	2
$\sum x_i = 5$	$\sum y_i = 10$	$\sum(x_i-\bar{x}) = 0$	$\sum(x_i-\bar{x})^2 = 10$	$\sum(y_i-\bar{y}) = 0$	$\sum(x_i-\bar{x})(y_i-\bar{y}) = 8$

$\bar{x}=1$, $\bar{y}=2$

(b.)

$b_2 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{8}{10} = 0.8$

$b_1 = \bar{y} - b_2\bar{x} = 2 - 0.8\times 1 = 1.2$

b_2 represent the slope of a simple regression model

b_1 represent intercept of a simple regression model

(c.) $\Sigma(x_i-\bar{x})^2 = \Sigma x_i^2 - N\bar{x}^2$

$10 = \Sigma x_i^2 - 5\times 1$

$\Rightarrow \Sigma x_i^2 = 15$

$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = \Sigma x_i y_i - N\bar{x}\bar{y}$

$8 = \Sigma x_i y_i - 5\times 1\times 2$

$\Rightarrow \Sigma x_i y_i = 18$

(d.)

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
3	4	3.6	0.4	0.16	1.2
2	2	2.8	−0.8	0.64	−1.6
1	3	2	1	1	1
−1	1	0.4	0.6	0.36	−0.6
0	0	1.2	−1.2	1.44	0
$\sum x_i = 5$	$\sum y_i = 10$	$\sum \hat{y}_i = 10$	$\sum \hat{e}_i = 0$	$\sum \hat{e}_i^2 = 3.6$	$\sum x_i\hat{e}_i = 0$

$\hat{y}_i = 1.2 + 0.8x_i$

median of $x = 1$

50th percentile of $x = 1$

$s_y^2 = \sum_{i=1}^{N}(y_i-\bar{y})^2/N-1 = 10/5-1 = 10/4 = 2.5$

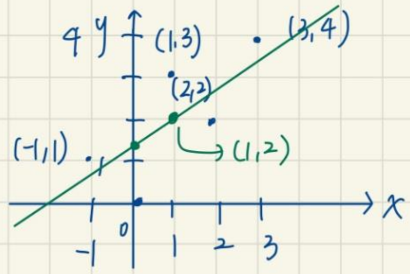
$s_x^2 = \sum_{i=1}^{N}(x_i-\bar{x})^2/N-1 = 10/5-1 = 2.5$

$s_{xy} = \sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})/N-1 = 8/5-1 = 2$

$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{2}{\sqrt{2.5\times 2.5}} = \frac{2}{2.5} = 0.8$

$CV_x = 100\left(\frac{s_x}{\bar{x}}\right) = 100\left(\frac{\sqrt{2.5}}{1}\right) = 50\sqrt{10}$

(e.)



(f.) The fitted line pass $(\bar{x}, \bar{y}) = (1, 2)$

(g.) Given $b_1 = 1.2$ $b_2 = 0.8$ and $\bar{y} = b_1 + b_2\bar{x}$, we have $\bar{y} = 2 = b_1 + b_2\bar{x} = 1.2 + 0.8 \times 1$

(h.) $\bar{\hat{y}} = \frac{\sum \hat{y}_i}{N} = \frac{10}{5} = 2 = \bar{y}$

(i) $\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{3.6}{3} = 1.2$

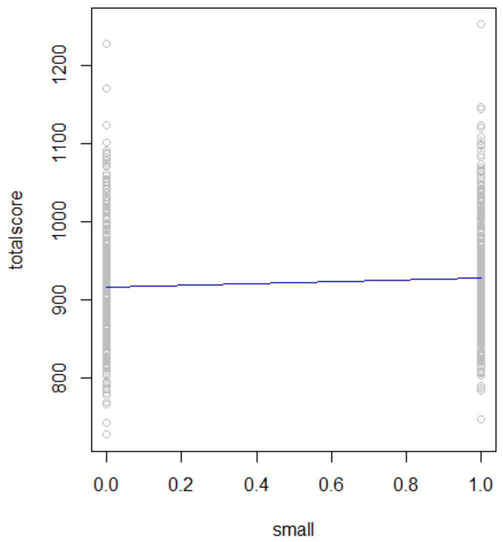
(j) $\widehat{var}(b_2) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} = \frac{1.2}{10} = 0.12$

2-22

(a) Because there are 425 students belong neither small nor regular class (small=0 & regular=0), we remove them then use a linear regression model $TOTALSCORE = \beta_1 + \beta_2 SMALL_i + e_i$ to fit the data. The result shows that the R-squared is small and residual's variance is large (from -189.44 to 324.38). The linear regression model can't fit the data well. After plotting the scatter plot and fitted line. We find that the data only distribute in where small=0 (regular=1) or small=1 (regular=0), so it doesn't make sense to use the linear fitted line in this data.

```
Coefficients:
(Intercept)        small
     916.44        12.18
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-189.44  -50.44   -9.44   43.06  324.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  916.442      3.675 249.396   <2e-16 ***
small         12.175      5.369   2.268   0.0236 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.59 on 773 degrees of freedom
Multiple R-squared:  0.006608,	Adjusted R-squared:  0.005323
F-statistic: 5.142 on 1 and 773 DF,  p-value: 0.02363
```

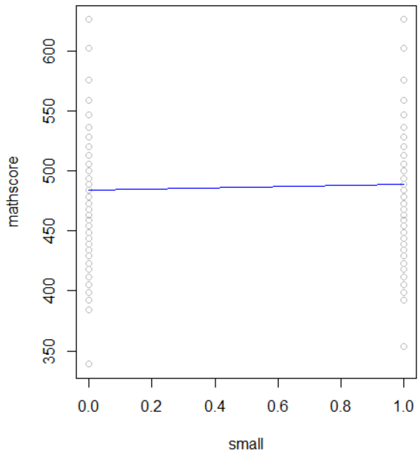


Rcode
total <-which(star5_small$small==1\|star5_small$regular==1)
newstar<-star5_small[total,]
newstar
mod1 <- lm(totalscore~small,data = newstar)
mod1
summary(mod1)
plot(newstar$small, newstar$totalscore,xlab = 'small',ylab = 'totalscore', col="grey")
lines(fitted(mod1)~newstar$small, col="blue")

(b) The result of the linear regression model $MATHSCORE = \beta_1 + \beta_2 SMALL_i + e_i$ shows the similar result. The R-squared is small and residual's variance is large. Because the distribution of SMALL is the same, the fitted line can't fit the data well.

```
Coefficients:
(Intercept)        small
    483.777        5.251
```

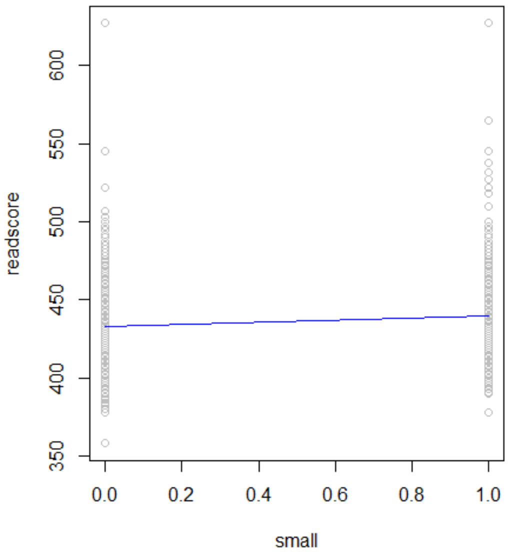
```
Call:
lm(formula = mathscore ~ small, data = newstar)

Residuals:
     Min       1Q   Median       3Q      Max
-144.777  -35.028   -5.028   29.223  142.223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  483.777      2.449 197.545   <2e-16 ***
small          5.251      3.578   1.467    0.143
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 49.71 on 773 degrees of freedom
Multiple R-squared:  0.002778,	Adjusted R-squared:  0.00148
F-statistic: 2.153 on 1 and 773 DF,  p-value: 0.1427
```



The result of the linear regression model $READSCORE = \beta_1 + \beta_2 SMALL_i + e_i$ shows the similar result. The R-squared is small and residual's variance is large. Because the distribution of SMALL is the same, the fitted line can't fit the data well.

```
Coefficients:
(Intercept)        small
    432.665        6.924
```

```
Call:
lm(formula = readscore ~ small, data = newstar)

Residuals:
    Min      1Q  Median      3Q     Max
-74.665 -21.590  -2.665  16.410 194.335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  432.665      1.488 290.760  < 2e-16 ***
small          6.924      2.174   3.185  0.00151 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.2 on 773 degrees of freedom
Multiple R-squared:  0.01295,    Adjusted R-squared:  0.01167
F-statistic: 10.14 on 1 and 773 DF,  p-value: 0.001507
```



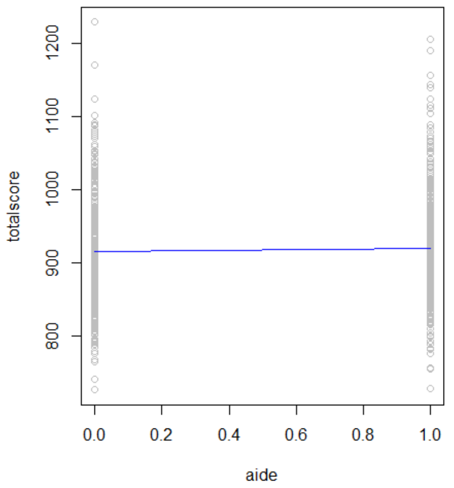
Rcode

```
mod2 <- lm(mathscore~small,data = newstar)
mod2
summary(mod2)
plot(newstar$small, newstar$mathscore,xlab = 'small',ylab = 'mathscore',
col="grey")
lines(fitted(mod2)~newstar$small, col="blue")
mod3 <- lm(readscore~small,data = newstar)
mod3
summary(mod3)
plot(newstar$small, newstar$readscore,xlab = 'small',ylab = 'readscore',
col="grey")
lines(fitted(mod3)~newstar$small, col="blue")
```

(c) We use a linear regression model $TOTALSCORE = \beta_1 + \beta_2 ASIDE_i + e_i$ to fit the data. The result shows that the R-squared is small and residual's variance is large. The linear regression model can't fit the data well. After plotting the scatter plot and fitted line. We find that the data only distribute in where aide=0 (regular=1) or aside=1 (regular=0), so it doesn't make sense to use the linear fitted line in this data.

```
Coefficients:
(Intercept)         aide
    916.442        4.306
```

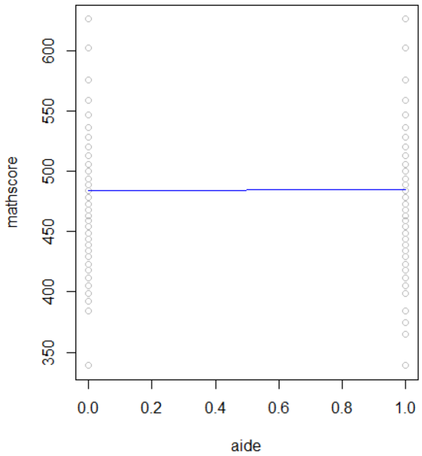
```
Call:
lm(formula = totalscore ~ aide, data = newstar2)

Residuals:
     Min       1Q   Median       3Q      Max
-191.748  -50.442   -5.442   40.558  312.558

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  916.442      3.559 257.529   <2e-16 ***
aide           4.306      4.994   0.862    0.389
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72.23 on 835 degrees of freedom
Multiple R-squared:  0.0008898, Adjusted R-squared:  -0.0003068
```



Rcode
total2 <-which(star5_small$aide==1\|star5_small$regular==1)
newstar2<-star5_small[total2,]
mod4 <- lm(totalscore~aide,data =newstar2)
mod4
summary(mod4)
plot(newstar2$aide, newstar2$totalscore,xlab = 'aide',ylab = 'totalscore',
col="grey")
lines(fitted(mod4)~newstar2$aide, col="blue")

(d) The result of the linear regression model $MATHSCORE = \beta_1 + \beta_2 SMALL_i + e_i$ shows the similar result. The R-squared is small and residual's variance is large. Because the distribution of SMALL is the same, the fitted line can't fit the data well.

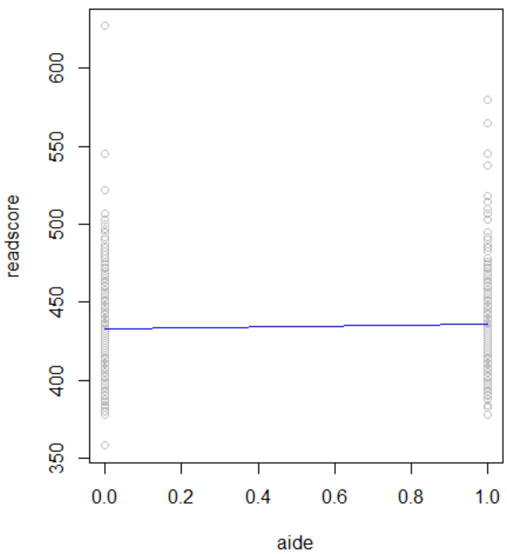
Coefficients: (Intercept) aide 483.777 1.435	Call: lm(formula = mathscore ~ aide, data = newstar2) Residuals: Min 1Q Median 3Q Max -146.212 -31.212 -5.777 29.223 142.223 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 483.777 2.355 205.464 <2e-16 *** aide 1.435 3.304 0.434 0.664 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 47.79 on 835 degrees of freedom Multiple R-squared: 0.0002258, Adjusted R-squared: -0.0009715 F-statistic: 0.1886 on 1 and 835 DF, p-value: 0.6642



The result of the linear regression model $MATHSCORE = \beta_1 + \beta_2 SMALL_i + e_i$ shows the similar result. The R-squared is small and residual's variance is large. Because the distribution of SMALL is the same, the fitted line can't fit the data well.

```
Coefficients:
(Intercept)         aide
    432.665        2.871
```

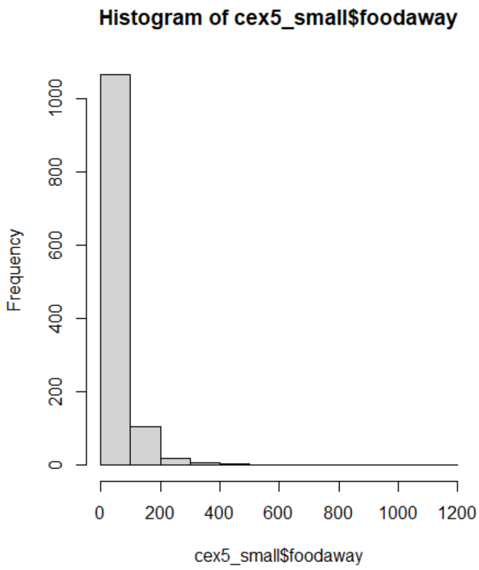
```
Call:
lm(formula = readscore ~ aide, data = newstar2)

Residuals:
    Min      1Q  Median      3Q     Max
-74.665 -21.536  -2.536  15.464 194.335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  432.665      1.460 296.341   <2e-16 ***
aide           2.871      2.049   1.401    0.161
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.64 on 835 degrees of freedom
Multiple R-squared:  0.002347,	Adjusted R-squared:  0.001152
F-statistic: 1.964 on 1 and 835 DF,  p-value: 0.1615
```



Rcode
mod5 <- lm(mathscore~aide,data =newstar2)
mod5
summary(mod5)
plot(newstar2$aide, newstar2$mathscore,xlab = 'aide',ylab = 'mathscore', col="grey")
lines(fitted(mod5)~newstar2$aide, col="blue")
mod6 <- lm(readscore~aide,data =newstar2)
mod6
summary(mod6)
plot(newstar2$aide, newstar2$readscore,xlab = 'aide',ylab = 'readscore', col="grey")
lines(fitted(mod6)~newstar2$aide, col="blue")

2-25

(a) The histogram of foodaway in data called cex5_small



And statistic summary:

The mean is 49.27 and median is 32.55. The 25th percentiles is 12.04 and 75th percentiles is 67.50

```
summary(cex5_small$foodaway)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.00   12.04   32.55   49.27   67.50 1179.00 
```

Rcode
hist(cex5_small$foodaway)
summary(cex5_small$foodaway)

(b) The mean and median values of *FOODAWAY* for households including a member with an advanced degree is 48.15 and 73.15494

```
> median(adv)
[1] 48.15
> mean(adv)
[1] 73.15494
```

The mean and median values of *FOODAWAY* for households including a member with an college degree is 36.11 and 48.59718

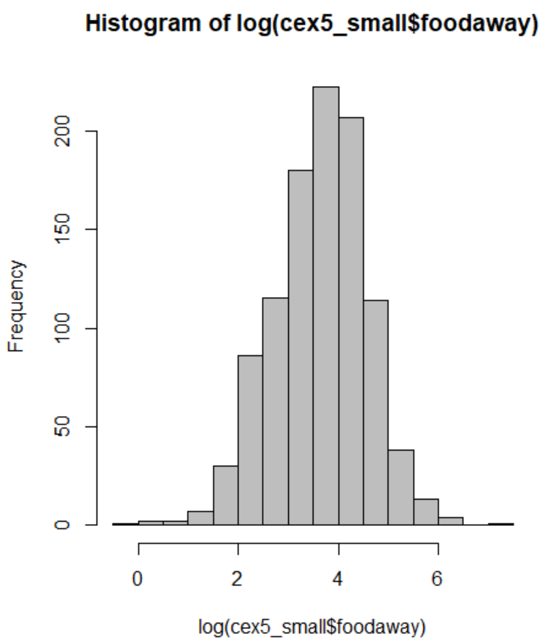
```
> median(col)
[1] 36.11
> mean(col)
[1] 48.59718
```

The mean and median values of *FOODAWAY* for households including a member with an advanced or college degree is 38.52 and 58.6792

```
> median(adco)
[1] 38.52
> mean(adco)
[1] 58.6792
```

Rcode
adv = cex5_small$foodaway[cex5_small$advanced==1]
median(adv)
mean(adv)
col = cex5_small$foodaway[cex5_small$college==1]
median(col)
mean(col)
adco=cex5_small$foodaway[cex5_small$advanced==1|cex5_small$college==1]
median(adco)
mean(adco)

(c) The histogram of ln(*FOODAWAY*) is:



And the statistic summary is :

```
summary(log(cex5_small$foodaway))
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -Inf   2.488   3.483    -Inf   4.212   7.072
```

The difference between *FOODAWAY* and ln(*FOODAWAY*) is the distribution of the former is skew to right and the letter is more like a normal and have some negative values.

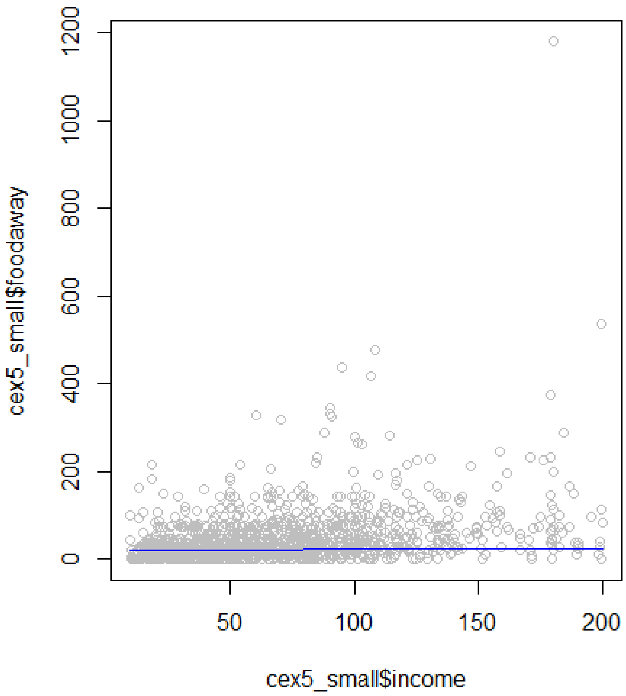
Rcode
hist(log(cex5_small$foodaway), col='grey') summary(log(cex5_small$foodaway))

(d) The result of the linear regression model $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME_i + e_i$ shows that the β_2 is 0.0005375. We estimate that if the income goes up by $100, expected ln(*FOODAWAY*) (food away from home expenditure per month per person past quarter) will increase approximately by $0.0005375.

```
Coefficients:
(Intercept)       income
  3.0704902    0.0005375
```

Rcode
ordat <- cex5_small[order(cex5_small$income),] log_foodaway<-log(cex5_small$foodaway) log_foodaway

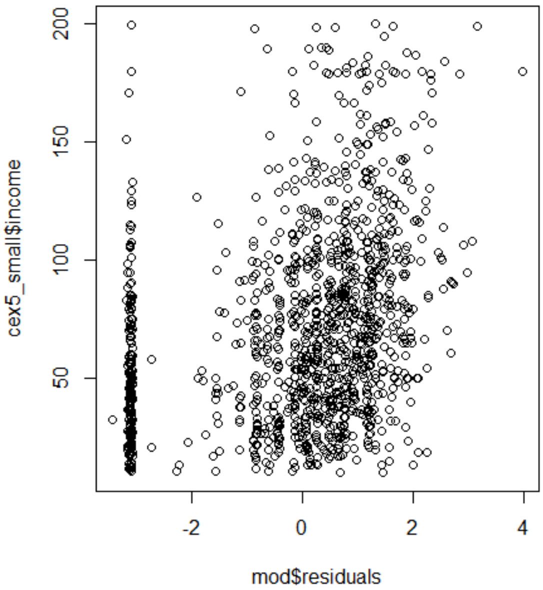
```
sum(is.infinite(log_foodaway))
log_foodaway[is.infinite(log_foodaway)]<-0
sum(is.infinite(log_foodaway))
mod <- lm(log_foodaway~income,data = ordat)
mod
```



(e)

Rcode
plot(cex5_small$income, log(cex5_small$foodaway), col="grey")
plot(cex5_small$income, cex5_small$foodaway, col="grey")
lines(exp(fitted(mod))~ordat$income, col="blue", main="Log-linear Model")

(f) After calculating the least squares residuals from the estimation. We plot of residuals v.s income. And find that there are some residual is negative (about -3). We think these residual happened because we replace the negative ln(*FOODAWAY*) with 0, so their residual value is not like others that distribution is like normal.



Rcode
attributes(mod) mod$residuals plot(mod$residuals,cex5_small$income)

Part 2

$$y_i = \beta x_i + e_i \qquad e_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$S(\beta) = \sum_{i=1}^{N} (y_i - \beta x_i)^2 = \sum_{i=1}^{N} (y_i^2 - 2\beta y_i x_i + \beta^2 x_i^2)$$

(a.) $$\frac{\partial S(\beta)}{\partial \beta} = 0 \qquad \sum_{i=1}^{N} -2y_i x_i + \sum_{i=1}^{N} 2\beta x_i^2 = 0 \qquad \sum_{i=1}^{N} \cancel{2}\beta x_i^2 = \sum_{i=1}^{N} \cancel{2} y_i x_i \qquad \hat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

Let $w_i = \frac{x_i}{\sum_{j=1}^{n} x_i^2}$, $\hat{\beta} = \sum_{i=1}^{n} w_i y_i$

(b.) $$E(\hat{\beta}) = E\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i E(y_i) = \sum_{i=1}^{n} w_i (x_i \beta) = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{j=1}^{n} x_i^2}\right) x_i \beta$$

$$= \sum_{i=1}^{n} \frac{x_i^2}{\sum_{j=1}^{n} x_j^2} \beta = \beta$$

$$Var(\hat{\beta}) = Var\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i^2 \, Var(y_i) = \sum_{i=1}^{n} w_i^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i^2}{\left(\sum_{j=1}^{n} x_j^2\right)^2}\right) = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{j=1}^{n} x_j^2\right)^2} = \frac{\sigma^2}{\sum_{j=1}^{n} x_j^2}$$

$$\rightarrow \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{j=1}^{n} x_j^2}\right)$$