4.4 The general manager of a large engineering firm wants to know whether the experience of technical artists influences their work quality. A random sample of 50 artists is selected. Using years of work experience (*EXPER*) and a performance rating (*RATING*, on a 100-point scale), two models are estimated by least squares. The estimates and standard errors are as follows:

Model 1:

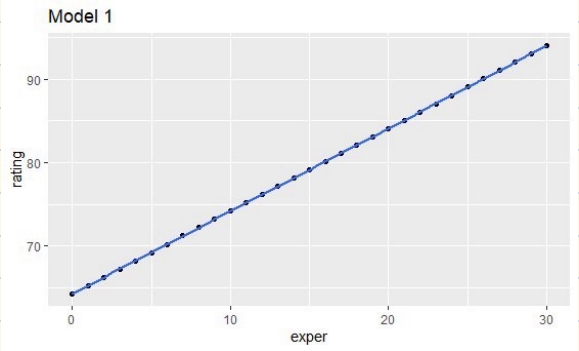
$$\widehat{RATING} = 64.289 + 0.990EXPER \quad N = 50 \quad R^2 = 0.3793$$
$$\text{(se)} \qquad (2.422) \quad (0.183)$$

Model 2:

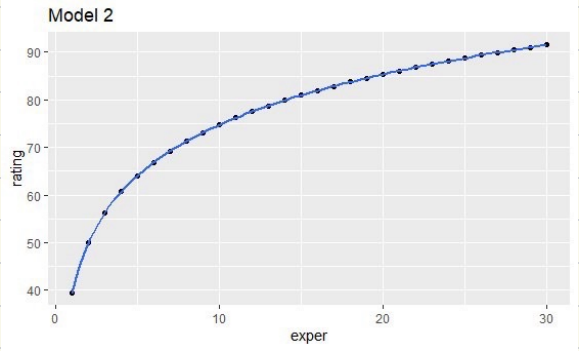
$$\widehat{RATING} = 39.464 + 15.312\ln(EXPER) \quad N = 46 \quad R^2 = 0.6414$$
$$\text{(se)} \qquad (4.198) \quad (1.727)$$

a. Sketch the fitted values from Model 1 for *EXPER* = 0 to 30 years.
b. Sketch the fitted values from Model 2 against *EXPER* = 1 to 30 years. Explain why the four artists with no experience are not used in the estimation of Model 2.
c. Using Model 1, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.
d. Using Model 2, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.
e. Which of the two models fits the data better? Estimation of Model 1 using just the technical artists with some experience yields $R^2 = 0.4858$.
f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning? Explain.

a.



b.



有可能原資料 EXPER=0，取 ln 後新的值趨近負無限大，造成資料筆數減少

c.

(i) EXPER = 10

$$\frac{\partial(\widehat{RATING})}{\partial EXPER} = 0.99$$

(ii) EXPER = 20

$$\frac{\partial \widehat{RATING}}{\partial EXPER} = 0.99$$

d.

(i) EXPER = 10

$$\frac{\partial(\widehat{RATING})}{\partial EXPER} = 15.312 \times \frac{1}{EXPER} = \frac{15.312}{10} = 1.5312$$

(ii) EXPER = 20

$$\frac{\partial(\widehat{RATING})}{\partial EXPER} = \frac{15.312}{20} = 0.7656$$

(e)

Model 2 fits the data better because R-squared is 0.6414, higher than Model 1's R-squared of 0.3793 (or 0.4858 when excluding artists with no experience). This suggests Model 2 explains more variation in the performance ratings.

(f)

Model 2 is more reasonable based on economic reasoning. It reflects diminishing marginal returns to experience, which aligns with the idea that early years of experience lead to larger improvements, while additional experience has less impact over time. Model 1 assumes a constant marginal effect, which is less realistic.

4.28 The file *wa-wheat.dat* contains observations on wheat yield in Western Australian shires. There are 48 annual observations for the years 1950–1997. For the Northampton shire, consider the following four equations:

$$① \quad YIELD_t = \beta_0 + \beta_1 TIME + e_t$$
$$② \quad YIELD_t = \alpha_0 + \alpha_1 \ln(TIME) + e_t$$
$$③ \quad YIELD_t = \gamma_0 + \gamma_1 TIME^2 + e_t$$
$$④ \quad \ln(YIELD_t) = \phi_0 + \phi_1 TIME + e_t$$

a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.

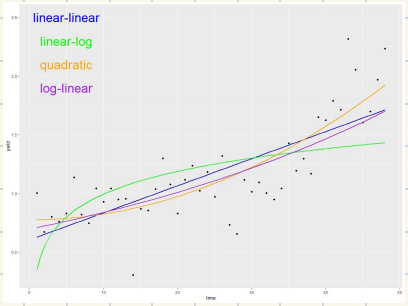
b. Interpret the coefficient of the time-related variable in your chosen specification.

c. Using your chosen specification, identify any unusual observations, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*.

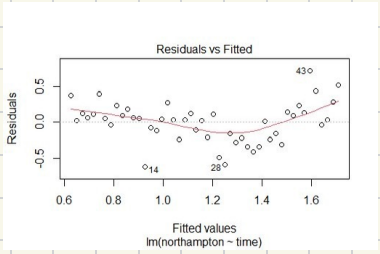
d. Using your chosen specification, use the observations up to 1996 to estimate the model. Construct a 95% prediction interval for *YIELD* in 1997. Does your interval contain the true value?

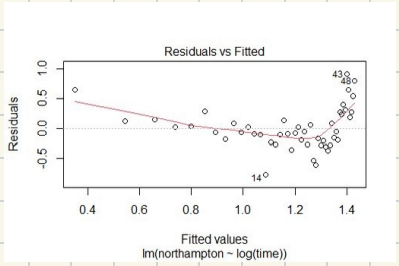
a.

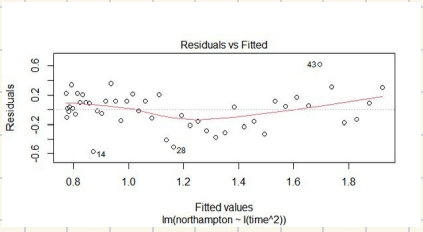
(i)

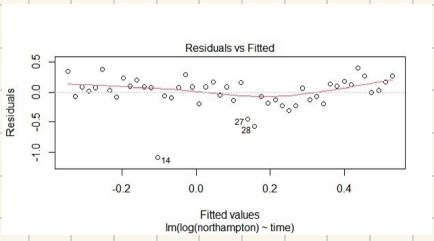


(ii)

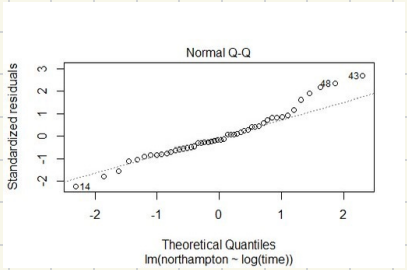
① 

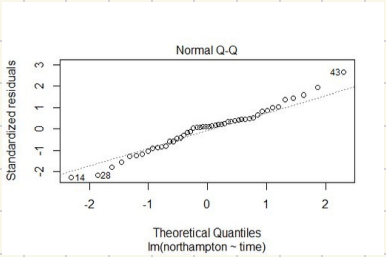
② 

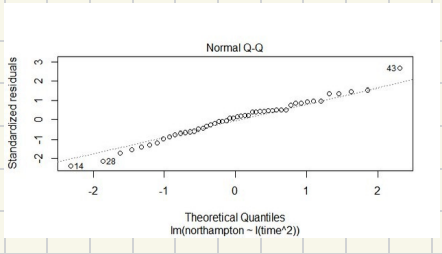
③ 

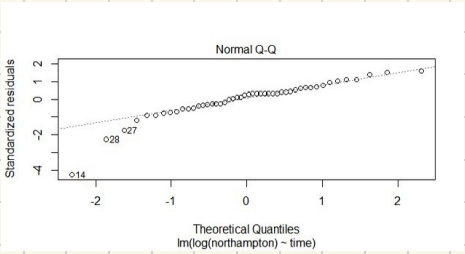
④ 

(iii)

① 

② 

③ 

④ 

(iii)

Jarque Bera Test p-value

① $R^2 = 0.5978$ SSE = 3.58 don't reject H_0

② $R^2 = 0.3386$ SSE = 5.61 don't reject H_0

③ $R^2 = 0.689$ SSE = 3.11 don't reject H_0

④ $R^2 = 0.5090$ SSE = 2.64 reject H_0

模型④雖然SSE最小，但誤差項不符合常態，模型③ $R^2 = 0.689$ 最接近1，SSE第2小，殘差項也符合常態
從(i)的圖也可以看出模型3的估計線更能貼合真實值

b.

$\widehat{YIELD} = 0.7737 + 0.0005\ TIME$

當 TIME 增加 1 個單位，YIELD 會增加 0.0005 (hecture)，也就是每年平均產量比去年增加 0.0005 (hecture)

C.

LEVERAGE

```
> which(leverage_values > threshold_leverage)  # 顯示高槓桿點的索引
45 46 47 48
45 46 47 48
```

DFBETAS

```
> which(abs(dfbetas_values) > 2/sqrt(n), arr.ind=TRUE)
   row col
6    6   1
14  14   1
14  14   2
43  43   2
44  44   2
48  48   2
```

DIFFITS

```
> which(abs(dffits_values) > threshold_dffits)  # 顯示影響值較大的點
14 43 48
14 43 48
```

d.

95% predict interval: [1.3619, 2.3823]

真實值為 1.9691 包含在 interval 中

4.29 Consider a model for household expenditure as a function of household income using the 2013 data from the Consumer Expenditure Survey, *cex5_small*. The data file *cex5* contains more observations. Our attention is restricted to three-person households, consisting of a husband, a wife, plus one other. In this exercise, we examine expenditures on a staple item, food. In this extended example, you are asked to compare the linear, log-log, and linear-log specifications.

a. Calculate summary statistics for the variables: *FOOD* and *INCOME*. Report for each the sample mean, median, minimum, maximum, and standard deviation. Construct histograms for both variables. Locate the variable mean and median on each histogram. Are the histograms symmetrical and "bell-shaped" curves? Is the sample mean larger than the median, or vice versa? Carry out the Jarque–Bera test for the normality of each variable.

b. Estimate the linear relationship $FOOD = \beta_1 + \beta_2 INCOME + e$. Create a scatter plot *FOOD* versus *INCOME* and include the fitted least squares line. Construct a 95% interval estimate for β_2. Have we estimated the effect of changing income on average *FOOD* relatively precisely, or not?

c. Obtain the least squares residuals from the regression in (b) and plot them against *INCOME*. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. Is it more important for the variables *FOOD* and *INCOME* to be normally distributed, or that the random error e be normally distributed? Explain your reasoning.

d. Calculate both a point estimate and a 95% interval estimate of the elasticity of food expenditure with respect to income at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Are the estimated elasticities similar or dissimilar? Do the interval estimates overlap or not? As *INCOME* increases should the income elasticity for food increase or decrease, based on Economics principles?

e. For expenditures on food, estimate the log-log relationship $\ln(FOOD) = \gamma_1 + \gamma_2 \ln(INCOME) + e$. Create a scatter plot for $\ln(FOOD)$ versus $\ln(INCOME)$ and include the fitted least squares line. Compare this to the plot in (b). Is the relationship more or less well-defined for the log-log model relative to the linear specification? Calculate the generalized R^2 for the log-log model and compare it to the R^2 from the linear model. Which of the models seems to fit the data better?

f. Construct a point and 95% interval estimate of the elasticity for the log-log model. Is the elasticity of food expenditure from the log-log model similar to that in part (d), or dissimilar? Provide statistical evidence for your claim.

g. Obtain the least squares residuals from the log-log model and plot them against $\ln(INCOME)$. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

h. For expenditures on food, estimate the linear-log relationship $FOOD = \alpha_1 + \alpha_2 \ln(INCOME) + e$. Create a scatter plot for *FOOD* versus $\ln(INCOME)$ and include the fitted least squares line. Compare this to the plots in (b) and (e). Is this relationship more well-defined compared to the others? Compare the R^2 values. Which of the models seems to fit the data better?

i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide statistical evidence for your claim.

j. Obtain the least squares residuals from the linear-log model and plot them against $\ln(INCOME)$. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

k. Based on this exercise, do you prefer the linear relationship model, or the log-log model or the linear-log model? Explain your reasoning.

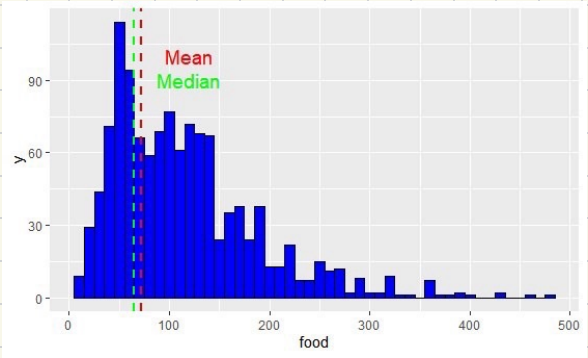
a.

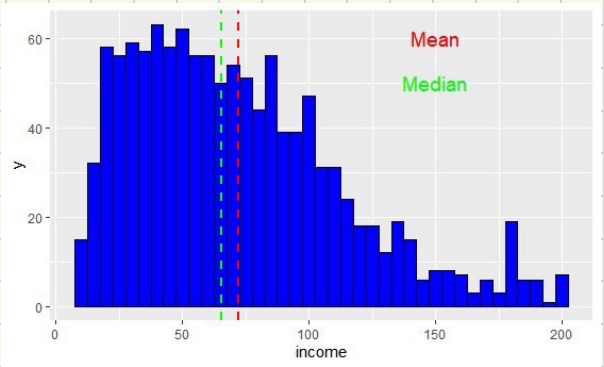
food	
Min.	: 9.63
1st Qu.	: 57.78
Median	: 99.80
Mean	:114.44
3rd Qu.	:145.00
Max.	:476.67

income	
Min.	: 10.00
1st Qu.	: 40.00
Median	: 65.29
Mean	: 72.14
3rd Qu.	: 96.79
Max.	:200.00

sd = 72.6575

sd = 41.6523





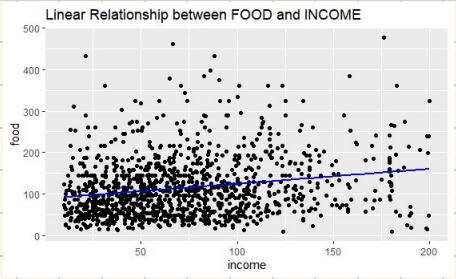

```
> jarque.bera.test(cex5_small$food)

        Jarque Bera Test

data:  cex5_small$food
X-squared = 648.65, df = 2, p-value < 2.2e-16
```

```
> jarque.bera.test(cex5_small$income)

        Jarque Bera Test

data:  cex5_small$income
X-squared = 148.21, df = 2, p-value < 2.2e-16
```

Both distributions are positively skewed with means greater than medians. They are not bell shaped or symmetrical. Both of their p-value are less than 5%, so we reject the null hypothesis of normality for each variable.

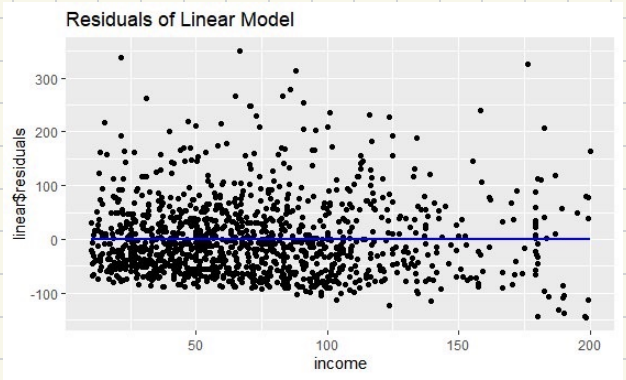
b.

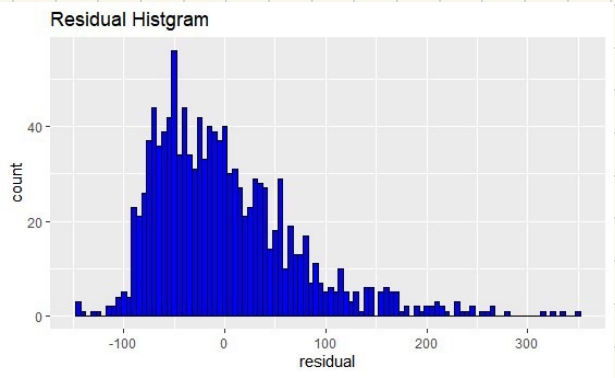
$\widehat{Food} = 88.5665 + 0.3587\ INCOME$



β_2 95% interval : [0.2619, 0.4555]

C.

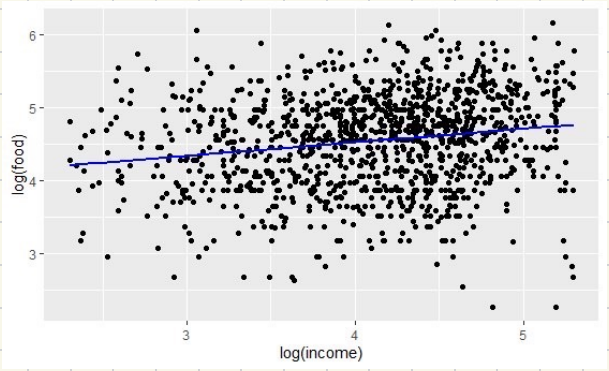


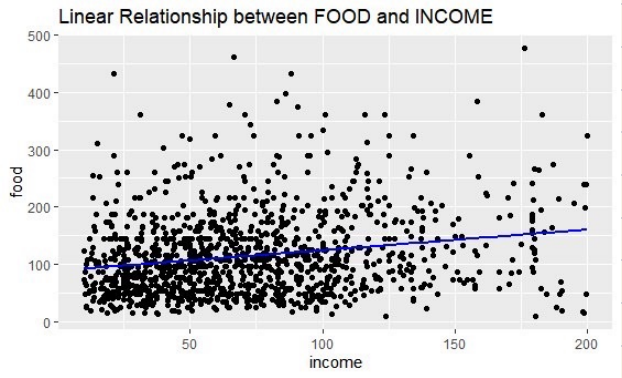



```
> jarque.bera.test(cex5_small$residual)

        Jarque Bera Test

data:  cex5_small$residual
X-squared = 624.19, df = 2, p-value < 2.2e-16
```

The random error e be normally distributed is more important than the variables Food and Income to be normal

e.





d.

$$\text{elasticity} = \frac{dy}{dx}\frac{X}{y}$$

$$= \beta_2 \frac{\text{Income}}{\text{Food}}$$

$$= \hat{\beta}_2 \frac{\text{Income}}{\hat{\beta}_1 + \hat{\beta}_2 \text{Income}}$$

Income = 19

$$\text{elasticity} = \hat{\beta}_2 \times \frac{19}{\hat{y}}$$

$$SE(\text{elasticity}) = \sqrt{\frac{19^2}{\hat{y}^2} Var(\hat{\beta}_2)} = \frac{19}{\hat{y}} SE(\hat{\beta}_2)$$

95% elasticity interval $= \hat{\beta}_2 \frac{19}{\hat{y}} \pm \frac{19}{\hat{y}} SE(\hat{\beta}_2)\, t_{\alpha/2}$

$$= \frac{19}{\hat{y}} \left[\hat{\beta}_2 \pm SE(\hat{\beta}_2)\, t_{\alpha/2} \right]$$

$$= [0.0522, 0.0907]$$

Income = 65

95% elasticity interval = [0.1522, 0.2646]

Income = 160

95% elasticity interval = [0.2871, 0.4993]

```
> summary(log_log)

Call:
lm(formula = log(food) ~ log(income), data = cex5_small)

Residuals:
     Min       1Q   Median       3Q      Max
-2.48175 -0.45497  0.06151  0.46063  1.72315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.77893    0.12035  31.400   <2e-16 ***
log(income)  0.18631    0.02903   6.417    2e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6418 on 1198 degrees of freedom
Multiple R-squared:  0.03323,   Adjusted R-squared:  0.03242
F-statistic: 41.18 on 1 and 1198 DF,  p-value: 1.999e-10
```

$R^2 = 0.03323$

```
> summary(linear)

Call:
lm(formula = food ~ income, data = cex5_small)

Residuals:
    Min      1Q  Median      3Q     Max
-145.37  -51.48  -13.52   35.50  349.81

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.56650    4.10819  21.559  < 2e-16 ***
income       0.35869    0.04932   7.272 6.36e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.13 on 1198 degrees of freedom
Multiple R-squared:  0.04228,   Adjusted R-squared:  0.04148
F-statistic: 52.89 on 1 and 1198 DF,  p-value: 6.357e-13
```

$R^2 = 0.04228$

The points surrounding the linear model are denser compared to the regression line of the log-log model, and the R-squared of the linear model is higher than that of the log-log model. Therefore, the linear model is a better model compared to the log-log model.

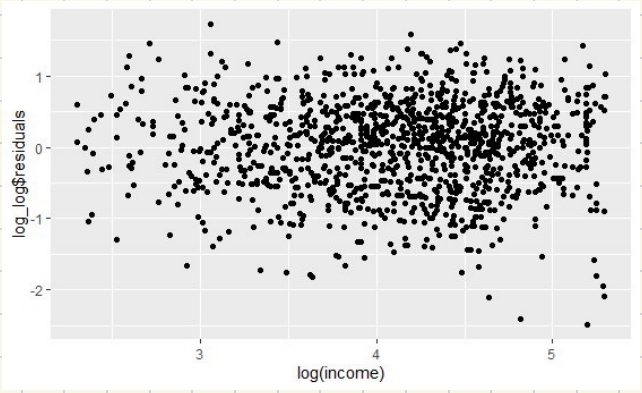
f.

elasticity = β_2

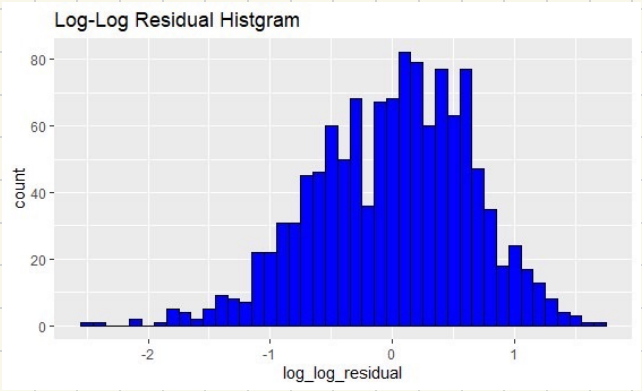
95% log-log model's elasticity interval

[0.2619, 0.4555]

g.



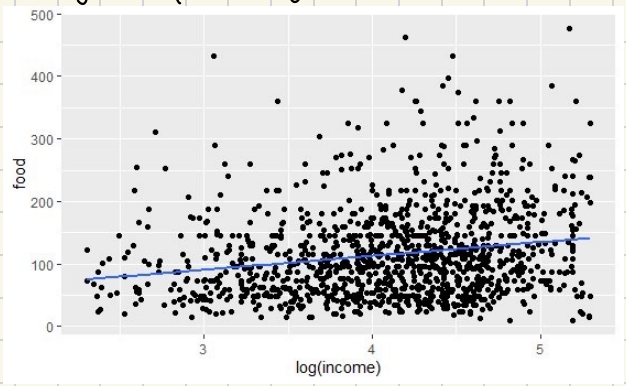
The residual points appear to be random.



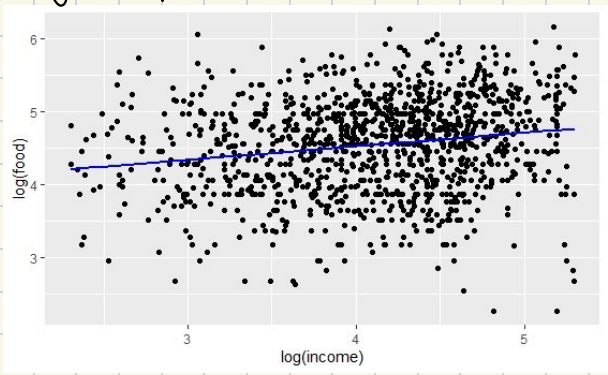

```
> jarque.bera.test(cex5_small$log_log_residual)

        Jarque Bera Test

data:  cex5_small$log_log_residual
X-squared = 25.85, df = 2, p-value = 2.436e-06
```

The histogram shows a left-skewed bell shape, and the test is a statistically significant, leading to the rejection of the null hypothesis. Therefore, the residuals do not follow a normal distribution.

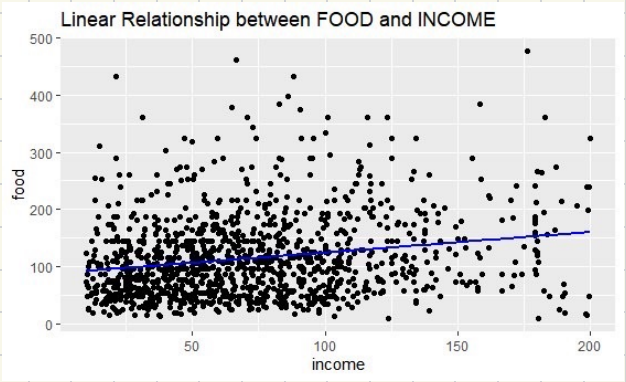
h. linear-log $R^2 = 0.038$



log-log $R^2 = 0.03323$



Linear $R^2 = 0.04228$



The figure is much like that for the linear model, and not as well defined as that for log-log model. The $R^2 = 0.038$, which is smaller than that of the linear model, and smaller than the generalized R^2 from the log-log model.

(i)

linear-log elasticity : $\frac{\beta_2}{y}$

$$SE\left(\frac{\beta_2}{y}\right) = \sqrt{Var\left(\frac{\beta_2}{y}\right)} = \sqrt{\frac{1}{y^2} Var(\beta_2)} = \frac{1}{y} SE(\beta_2)$$

$$\frac{\beta_2}{y} \pm \frac{1}{y} SE(\beta_2) t_{\alpha/2} = \frac{1}{y}\left[\beta_2 \pm SE(\beta_2) t_{\alpha/2}\right]$$

INCOME = 19 $\hat{FOOD} = 88.8979$ [0.1784, 0.3208]

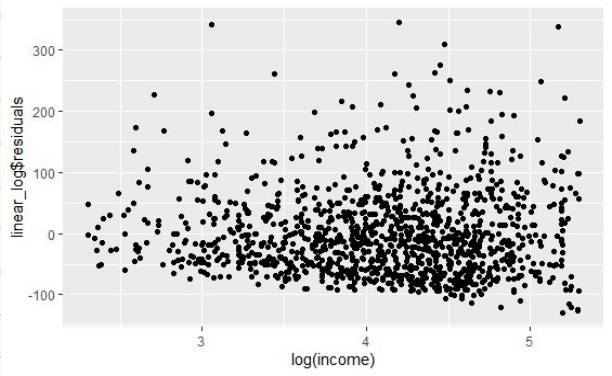
INCOME = 65 $\hat{FOOD} = 116.1872$ [0.1365, 0.2454]

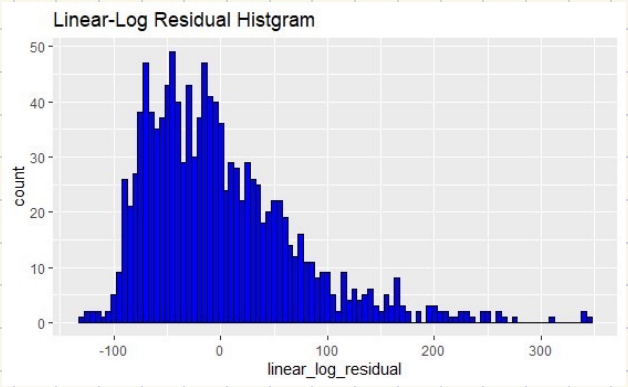
INCOME = 160 $\hat{FOOD} = 136.1733$ [0.1165, 0.2094]

INCOME

19	[0.0522, 0.0907]		[0.1784, 0.3208]
65	[0.1522, 0.2646]	[0.2619, 0.4555]	[0.1365, 0.2454]
160	[0.2871, 0.4993]		[0.1185, 0.2094]

(j)






```
        Jarque Bera Test

data:  cex5_small$linear_log_residual
X-squared = 628.07, df = 2, p-value < 2.2e-16
```

The residual scatter shows positive skewness at each income level and overall. The p-value is less than 5%, so we reject the normality of the model errors. The data scatter suggests a slight "spray" pattern.

k.

The linear model is counter-intuitive with increasing income elasticity. The linear-log model certainly statisfies economic reasoning, but the residual pattern is not an ideal random scatter. The log-log model implies that the income elasticity is constant for all income levels, which is not impossible to imagine, and the residual scatter is the most random, and the residuals are the least non-normal, based on skewness and kurtosis. On these grounds the log-log model seems like a good choice.