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:

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

Model 2:

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

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.

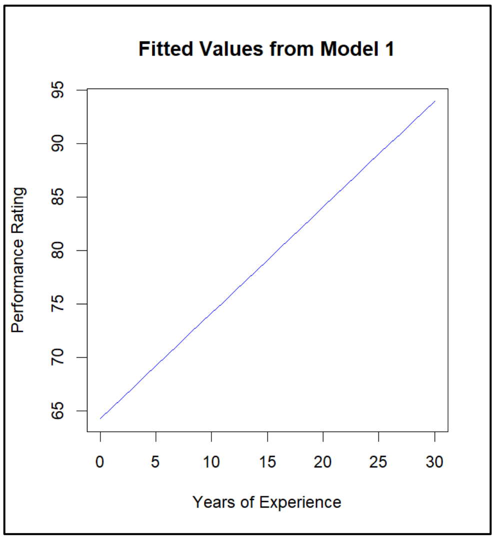
c. Using Model 1, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.

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.

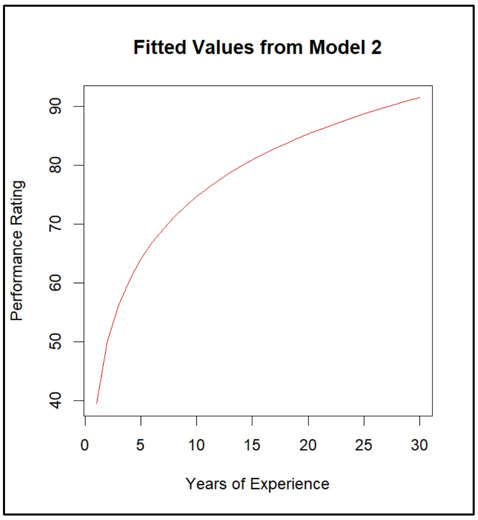
e. Which of the two models fits the data better? Estimation of Model 1 using just the technical artists with some experience yields $R^2 = 0.4858$.

f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning? Explain.

(a)



(b)



(c)

```
# 設定不同的工作經驗年數
years_10 <- 10
years_20 <- 20

# 計算邊際效應，對模型 1 來說是固定的
marginal_effect <- 0.990

# 顯示結果
marginal_effect_10 <- marginal_effect
marginal_effect_20 <- marginal_effect

marginal_effect_10
1] 0.99
marginal_effect_20
1] 0.99
```

(d)

```
> # 模型 2 的邊際效應公式
> marginal_effect_model2 <- function(EXPER) {
+     return(15.312 / EXPER)
+ }
>
> # 計算 10 年和 20 年經驗的邊際效應
> marginal_effect_10_model2 <- marginal_effect_model2(10)
> marginal_effect_20_model2 <- marginal_effect_model2(20)
>
> # 顯示結果
> marginal_effect_10_model2
[1] 1.5312
> marginal_effect_20_model2
[1] 0.7656
```

(e)

兩個模型有相同的因變數（RATING），可以用 R^2 作為判斷模型的優劣標準。線性模型一其 R^2=0.3793。在對數模型二中，排除沒有經驗的樣本後觀測數較少，其 R^2=0.6414。統計上 R^2 越接近 1 越好，代表模型越有更好的解釋力，故數線性模型二能更好地擬合這些數據。

(f)

模型一假設每增加一年經驗，評分的提升幅度始終相同。而模型 2 則有經驗的報酬遞減現象，即經驗對評分的影響在初期較大，但隨著經驗的增加，影響會逐漸減少。代表經驗較少的工作者，年資增加時的評分增幅較大；而對於經驗豐富的工作者，年資增加的幅度較小。

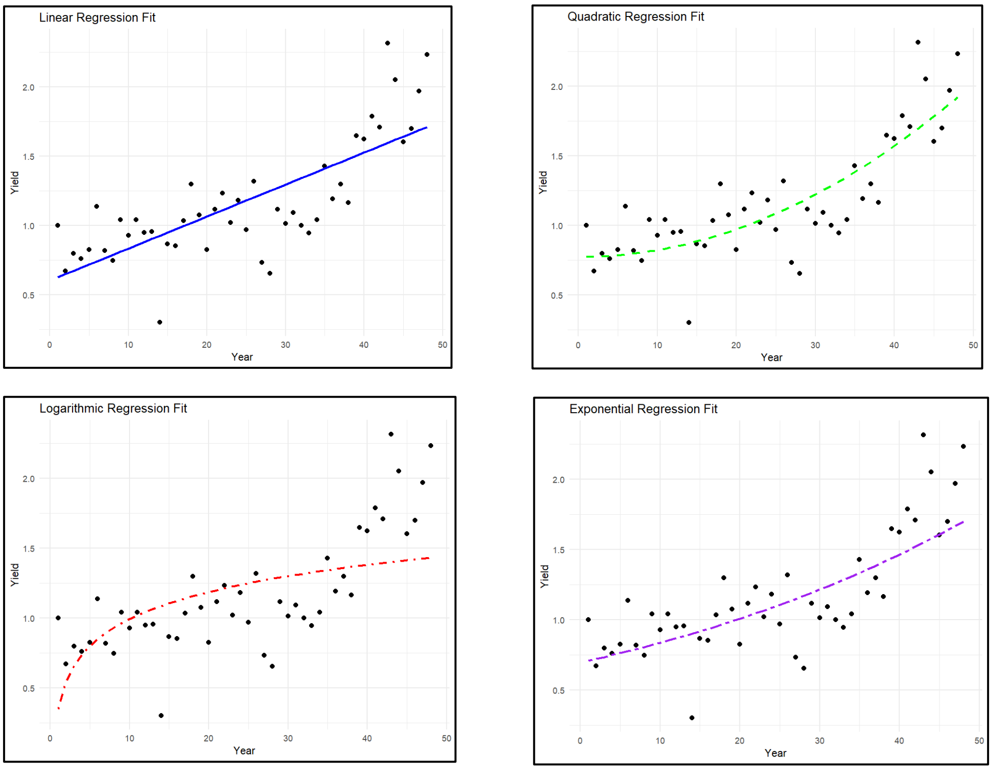
4.28 The file *wa-wheat.dat* contains observations on wheat yield in Western Australian shires. There are 48 annual observations for the years 1950–1997. For the Northampton shire, consider the following four equations:

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

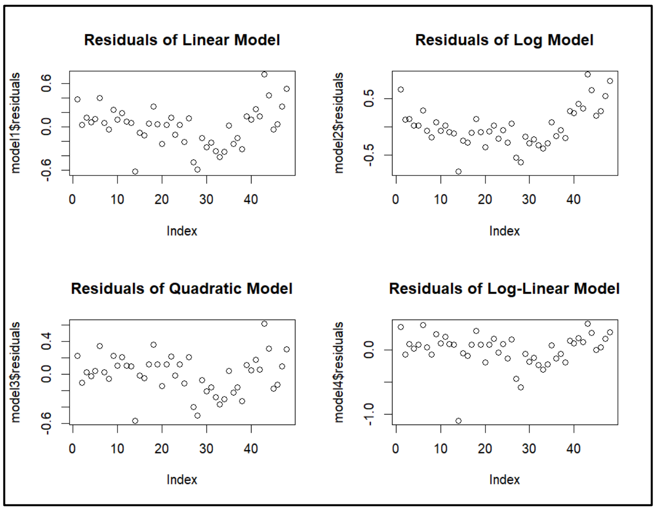
a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.
b. Interpret the coefficient of the time-related variable in your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



(iii) Error Normality Test & R^2

```
> # 4. 錯誤正態性檢驗
> shapiro.test(model1$residuals)

        Shapiro-Wilk normality test

data:  model1$residuals
W = 0.98236, p-value = 0.6792

> shapiro.test(model2$residuals)

        Shapiro-Wilk normality test

data:  model2$residuals
W = 0.96657, p-value = 0.1856

> shapiro.test(model3$residuals)

        Shapiro-Wilk normality test

data:  model3$residuals
W = 0.98589, p-value = 0.8266

> shapiro.test(model4$residuals)

        Shapiro-Wilk normality test

data:  model4$residuals
W = 0.86894, p-value = 7.205e-05
```

```
> # 5. 查看R^2
> summary(model1)$r.squared
[1] 0.5778369
> summary(model2)$r.squared
[1] 0.3385733
> summary(model3)$r.squared
[1] 0.6890101
> summary(model4)$r.squared
[1] 0.5073566
```

模型 3 (二次模型)：有最高的 R^2 值（0.6890），表示它對資料的擬合能力最強，並且 Error Normality Test 結果也顯示殘差分佈符合正態分佈，故模型三能最有效地解釋資料。

(b)

模型三 γ_1=0.0004986 隨著時間的平方增加，小麥產量會以加速的方式增長，即隨著時間推移，增長的速率會逐漸加快。

```
> # 模型1 (線性模型)
> summary(model1)$coefficients
              Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 0.60324512 0.081858457 7.369368 2.551676e-09
time        0.02307792 0.002908408 7.934899 3.689432e-10
>
> # 模型2 (對數模型)
> summary(model2)$coefficients
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.3509620 0.17589498 1.995293 5.195365e-02
log(time)   0.2790085 0.05749804 4.852487 1.439587e-05
>
> # 模型3 (二次模型)
> summary(model3)$coefficients
                Estimate   Std. Error  t value     Pr(>|t|)
(Intercept) 0.7736655220 5.221813e-02 14.81603 3.953882e-19
I(time^2)   0.0004986181 4.939119e-05 10.09528 3.007857e-13
>
> # 模型4 (對數線性模型)
> summary(model4)$coefficients
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) -0.36393758 0.076191522 -4.776615 1.852666e-05
time         0.01863235 0.002707063  6.882864 1.365839e-08
```

(c)

```
> print(influence_measures)
   Observation Studentized_Residuals   Leverage      DFBETAS      DFFITS
1            1            0.97117127 0.04743473  0.216718802  0.21671884
2            2           -0.43892315 0.04723338 -0.097727849 -0.09772816
3            3            0.09154376 0.04689948  0.020306565  0.02030689
4            4           -0.09362102 0.04643560 -0.020658634 -0.02065970
5            5            0.17150978 0.04584531  0.037589949  0.03759473
6            6            1.48946942 0.04513318  0.323736684  0.32382335
7            7            0.09207526 0.04430484  0.019814787  0.01982480
8            8           -0.25121369 0.04336691 -0.053440095 -0.05348720
9            9            0.97031031 0.04232704  0.203696093  0.20399098
10          10            0.43928373 0.04119390  0.090847178  0.09105340
11          11            0.88308253 0.03997718  0.179588800  0.18020490
12          12            0.43133925 0.03868759  0.086097858  0.08653117
13          13            0.40663499 0.03733687  0.079509348  0.08008229
14          14           -2.56068246 0.03593775 -0.489450177 -0.49440017
15          15           -0.07921998 0.03450401 -0.014769753 -0.01497595
16          16           -0.20039139 0.03305043 -0.036357010 -0.03704808
17          17            0.49312413 0.03159284  0.086845864  0.08906797
18          18            1.55776314 0.03014805  0.265587279  0.27464917
19          19            0.51140018 0.02873391  0.084160844  0.08796079
20          20           -0.61544215 0.02736930 -0.097452600 -0.10323931
21          21            0.51116364 0.02607410  0.077606439  0.08363762
22          22            0.92505667 0.02486923  0.134136097  0.14772978
23          23           -0.06616263 0.02377660 -0.009122960 -0.01032555
24          24            0.50898647 0.02281918  0.066410121  0.07778015
25          25           -0.48508765 0.02202093 -0.059553461 -0.07279025
26          26            0.87138263 0.02140683  0.100005566  0.12887955
27          27           -1.74863798 0.02100290 -0.186175532 -0.25612316
28          28           -2.24684727 0.02083617 -0.219908713 -0.32775913
29          29           -0.31870520 0.02093468 -0.028358757 -0.04660328
30          30           -0.87713750 0.02132750 -0.069984024 -0.12948485
31          31           -0.66971694 0.02204473 -0.047073157 -0.10055048
32          32           -1.20147489 0.02311746 -0.072666378 -0.18482622
33          33           -1.58783937 0.02457783 -0.079964118 -0.25204730
34          34           -1.31340954 0.02645899 -0.052431496 -0.21652582
35          35            0.17862729 0.02879511  0.005207561  0.03075755
36          36           -0.96321184 0.03162137 -0.017387057 -0.17405621
37          37           -0.67396104 0.03497398 -0.004450857 -0.12830329
38          38           -1.40993533 0.03889017  0.007330872 -0.28361722
39          39            0.48980475 0.04340819 -0.008509056  0.10433872
```

```
40          40            0.21784595 0.04856730 -0.006520055  0.04921896
41          41            0.75037258 0.05440780 -0.032182354  0.17999300
42          42            0.24372124 0.06097100 -0.013713793  0.06210342
43          43            2.88944743 0.06829921 -0.202525494  0.78231995
44          44            1.37882863 0.07643579 -0.116349648  0.39666614
45          45           -0.77948519 0.08542511  0.077302871 -0.23822701
46          46           -0.56948934 0.09531255  0.065201030 -0.18484656
47          47            0.41115999 0.10614453 -0.053605470  0.14168569
48          48            1.38846474 0.11796846 -0.203926321  0.50778020
```

(d)

在 95% 信賴水準下，1997 年估計值的信賴區間為[1.412563, 2.432401]，真實值為 2.2318。

```
> pred_i
       fit      lwr      upr
1 1.922482 1.412563 2.432401
```

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?

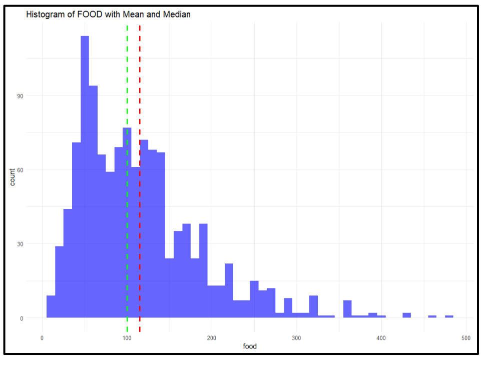
h. For expenditures on food, estimate the linear-log relationship $FOOD = \alpha_1 + \alpha_2 \ln(INCOME) + e$. Create a scatter plot for *FOOD* versus ln(*INCOME*) and include the fitted least squares line. Compare this to the plots in (b) and (e). Is this relationship more well-defined compared to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

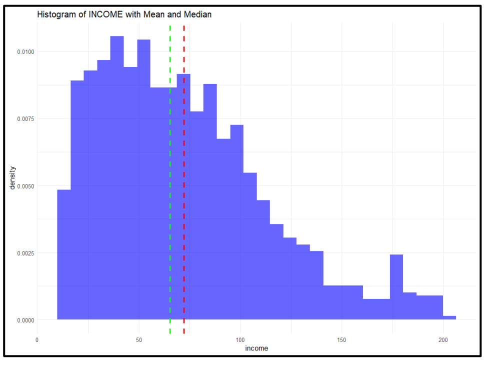
i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide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

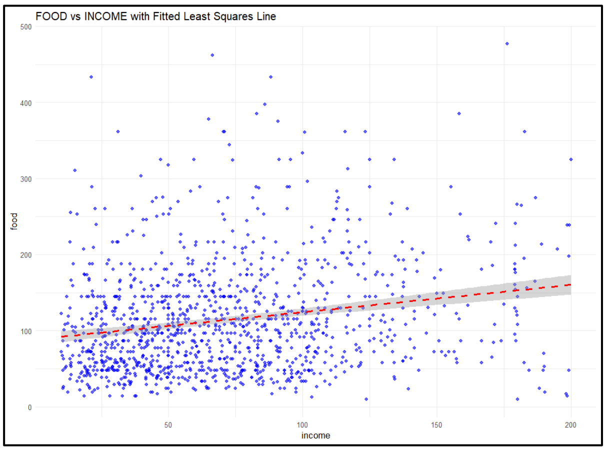




```
> # 進行Jarque-Bera正態性檢驗
> jarque.bera.test(cex5_small$food)

        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
```

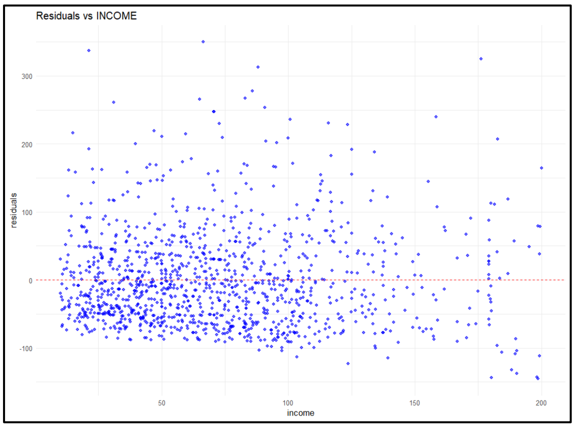
(b)

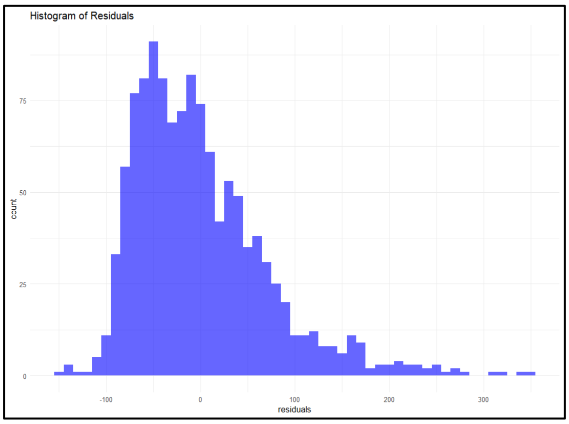
線性回歸其 $beta_2$ 在 95%信賴水準下的區間為(0.2619,0.4556)




```
> # 4. 構建95%置信區間估計
> confint(model, level = 0.95)  # 計算beta2的95%置信區間
                 2.5 %     97.5 %
(Intercept) 80.5064570 96.626543
income       0.2619215  0.455452
```

(c)殘差分佈圖沒有明顯的系統趨勢，其直方圖右偏。Jarque-Bera 統計量為 624.186，遠大於 5%的臨界值 5.99，故殘差並不符合常態分佈。





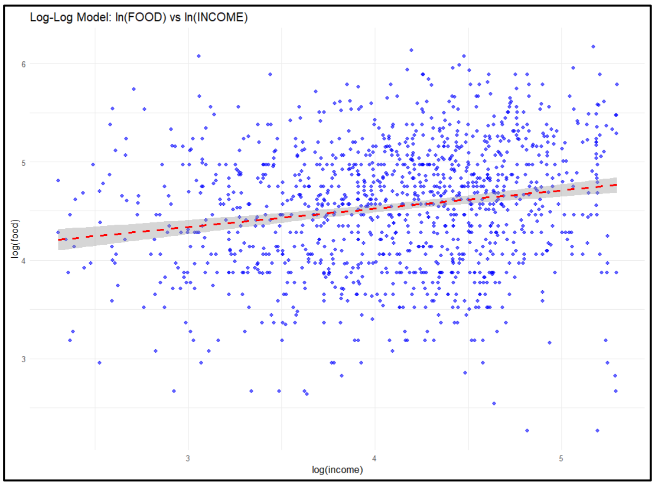

```
> # 3. 進行Jarque-Bera檢驗
> jarque.bera.test(residuals)

        Jarque Bera Test

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

(d)

```
b1 + b2*INCOME 和標準誤結果：
> print(data.frame(INCOME = income_values,
+                  `b1 + b2*INCOME` = b1_b2_income,
+                  `Standard Error` = se_b1_b2_income))
  INCOME b1...b2.INCOME Standard.Error
1     19       95.38155       3.329666
2     65      111.88114       2.083476
3    160      145.95638       4.795158
>
> cat("\n彈性和區間估計：\n")

彈性和區間估計：
> # 印出彈性估計、區間估計及彈性標準誤
> print(data.frame(INCOME = income_values,
+                  Elasticity = elasticities,
+                  `se(Elasticity)` = se_elasticities,
+                  LB = elasticity_intervals[, 1],
+                  UB = elasticity_intervals[, 2]))
  INCOME Elasticity se.Elasticity.         LB         UB
1     19 0.07145038      0.6632694 0.05217475 0.09072601
2     65 0.20838756      1.2104450 0.15216951 0.26460562
3    160 0.39319883      5.2565380 0.28712305 0.49927462
```

(e)

log-log 模型的 R^2 為 0.033，略小於線型模型的 0.042。




```
> # 顯示結果
> cat("Linear Model R^2: ", linear_r_squared, "\n")
Linear Model R^2:  0.0422812
> cat("Log-Log Model R^2: ", log_log_r_squared, "\n")
Log-Log Model R^2:  0.03322915
> cat("Log-Log Model Generalized R^2: ", generalized_r_squared, "\n")
Log-Log Model Generalized R^2:  0.03322915
```

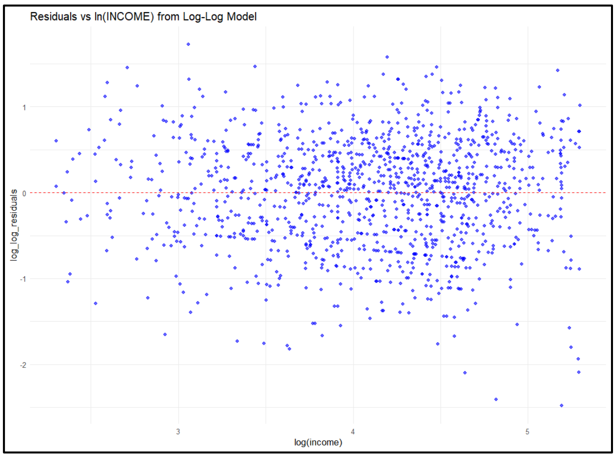
(f)

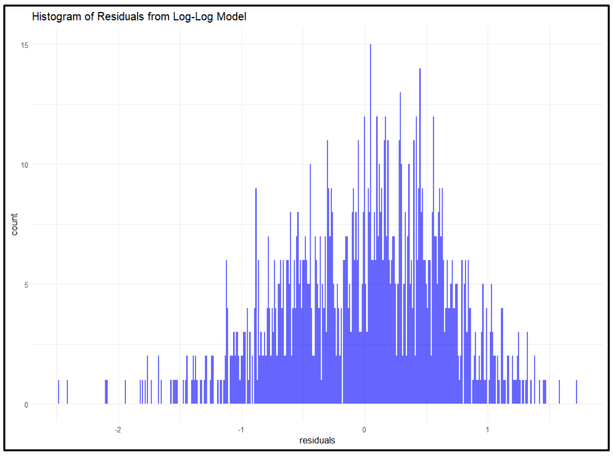
95%信賴水準下，log-log 模型的彈性區間為(0.1293,0.2433)。

```
> # 4. 打印結果
> cat("Point Estimate of Elasticity (log-log model): ", elasticity_point_estimate,
"\n")
Point Estimate of Elasticity (log-log model):  0.1863054
> cat("95% Confidence Interval for Elasticity: ", conf_intervals_gamma2, "\n")
95% Confidence Interval for Elasticity:  0.1293432 0.2432675
```

(g)

Jarque Bera 統計量為 25.85，大於臨界值 5.99，故拒絕 log-log 模型殘差為常態分佈的假設。

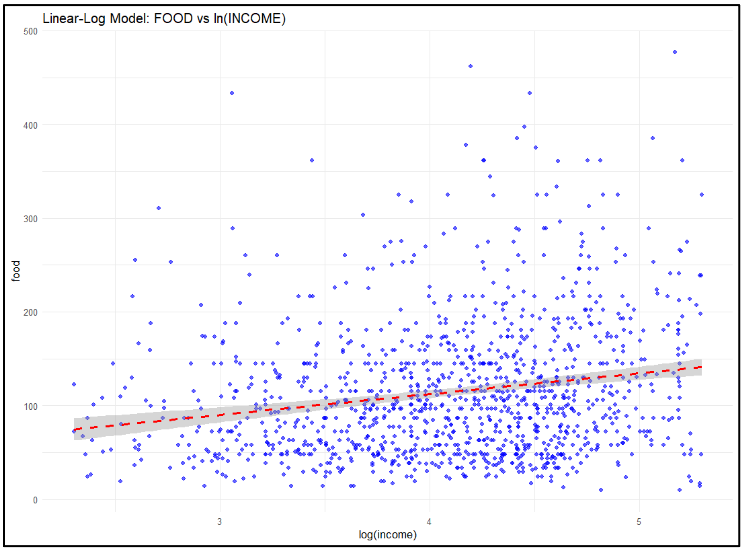





```
> # 4. 進行Jarque-Bera檢驗
> library(tseries)
> jarque.bera.test(log_log_residuals)

        Jarque Bera Test

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

(h)

linear-log 模型的為 R^2 為 0.038，略大於 log-log 模型，但小於線性模型。



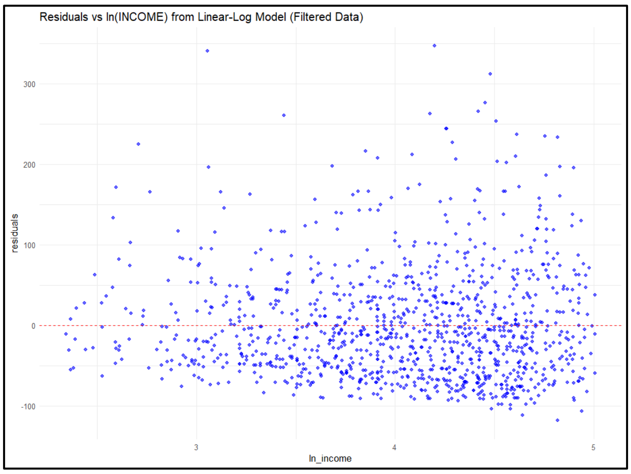

```
> # 5. 打印比較結果
> cat("Linear-Log Model R^2: ", linear_log_r_squared, "\n")
Linear-Log Model R^2:  0.03799984
> cat("Log-Log Model R^2: ", log_log_r_squared, "\n")
Log-Log Model R^2:  0.03322915
> cat("Linear Model R^2: ", linear_r_squared, "\n")
Linear Model R^2:  0.0422812
```

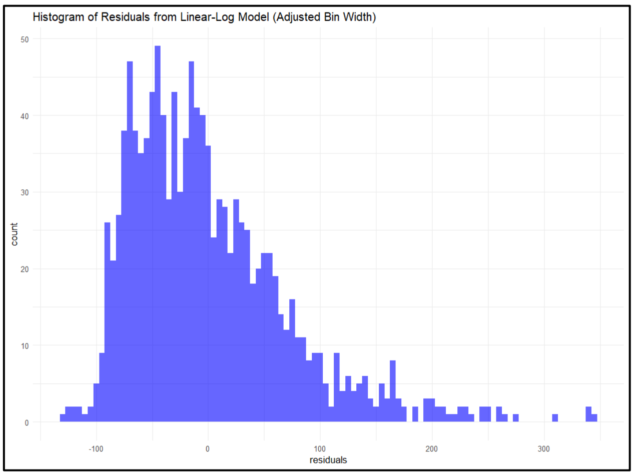
(i)

Jarque-Bera 統計量為 628.07，遠大於臨界值 5.99，故可以拒絕 linear-log 模型的殘差為常態分佈的假設。

```
> # 輸出結果
> print(elasticity_results)
  INCOME Predicted_FOOD Elasticity  Lower_CI  Upper_CI
1     19       88.89788  0.2495828 0.1784009 0.3207648
2     65      116.18722  0.1909624 0.1364992 0.2454256
3    160      136.17332  0.1629349 0.1164652 0.2094046
```

(j)





(k)

線性模型是不太合理的，因為收入增加時，支出增長不太可能保持成比例上升。

Linear-Log Model 雖然滿足收入的增長對食品支出的影響隨著收入的增加而變化的經濟推理，但殘差並不完全符合理想的隨機散佈。

Log-Log Model 假設無論收入是多高，收入對食品支出的影響（即彈性）保持恆定，並且根據殘差的散佈來看，它的誤差最為隨機，與理想的常態分佈最為接近。