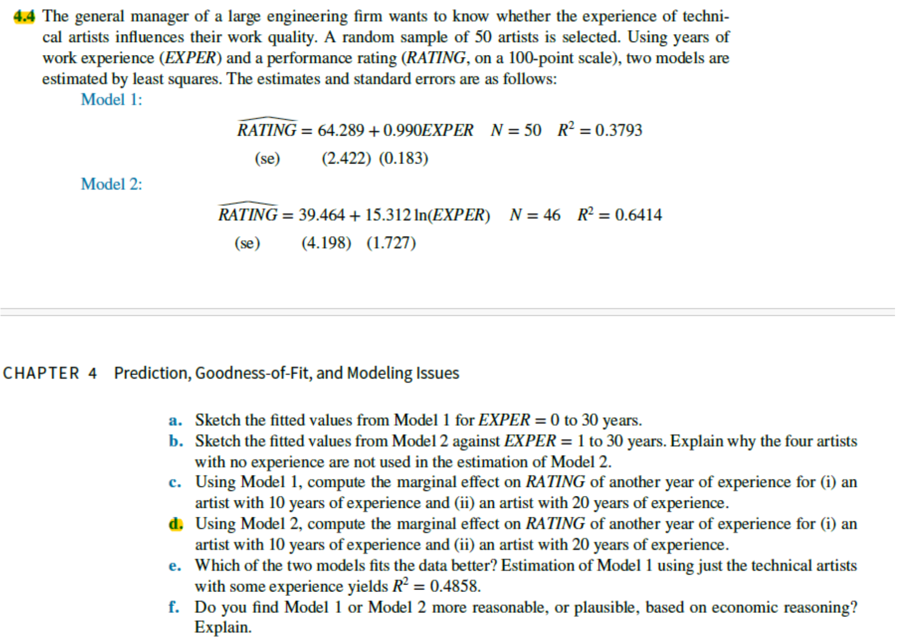


4.4 The general manager of a large engineering firm wants to know whether the experience of technical artists influences their work quality. A random sample of 50 artists is selected. Using years of work experience (*EXPER*) and a performance rating (*RATING*, on a 100-point scale), two models are estimated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

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

CHAPTER 4 Prediction, Goodness-of-Fit, and Modeling Issues

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.

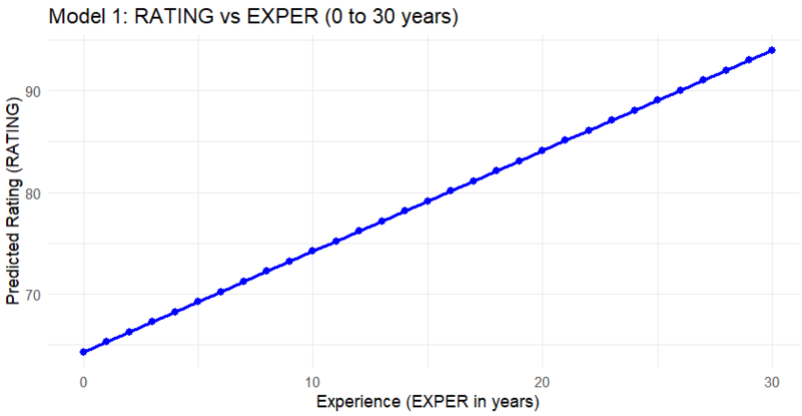
c. Using Model 1, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.

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.

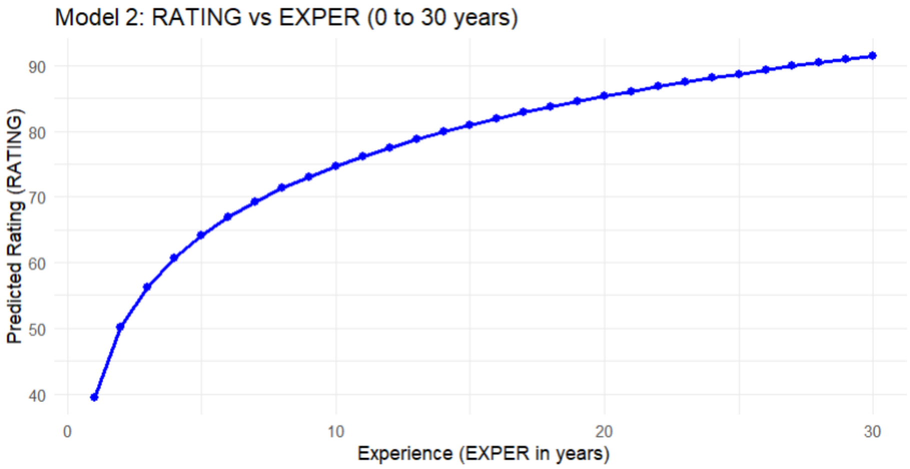
e. Which of the two models fits the data better? Estimation of Model 1 using just the technical artists with some experience yields $R^2 = 0.4858$.

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

a.



b.

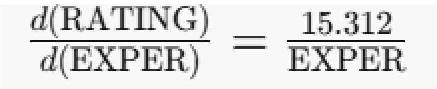


Model 2 的方程使用了 ln，而當 EXPER = 0 時，ln(0)是未定義的，無法計算 RATING。因此，這四位沒有經驗（EXPER = 0）的藝術家無法被納入 Model 2 的估計。

c.

在 Model 1 中，邊際效應表示每增加 1 年經驗，RATING 增加的量。因為這是一個線性模型，無論 EXPER 是 10 或 20，邊際效應都是常數 0.990。

d.



$$\frac{d(\text{RATING})}{d(\text{EXPER})} = \frac{15.312}{\text{EXPER}}$$

當 EXPER = 10 時，邊際效應 = 1.5312。這意味著當經驗年數為 10 年時，每增加 1 年經驗，評分增加 1.5312 分。
當 EXPER = 20 時，邊際效應 = 0.7656。這意味著當經驗年數為 20 年時，每增加 1 年經驗，評分增加 0.7656 分。
符合邊際效應遞減

e.

R^2 表示模型解釋的變異量占總變異量的比例，範圍在 0 到 1 之間。R^2 越高，說明模型對數據的擬合度越好。故 model 2 擬合度較高
如果 Model 1 在僅包括有經驗的藝術家時 R^2 = 0.4858 ，這比原始的 0.3793 高，說明去除無經驗的數據後，Model 1 的擬合度有所提升，但仍小於 model 2。

f.

Model 2 的對數形式假設經驗的回報遞減，這與經濟學中的回報遞減原理一致。在職業生涯早期，經驗的增加通常帶來顯著的技能提升和績效改善。例如，一個技術藝術家從 0 年經驗到 1 年經驗，可能學會了基礎技能，評分提升較大。隨著經驗積累，技能提升的空間變小，評分的增長速度會減緩。例如，一個有 20 年經驗的藝術家再增加 1 年經驗，可能只是在細節上有所改進，評分提升幅度較小。Model 2 能更好地捕捉這種非線性關係。

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.

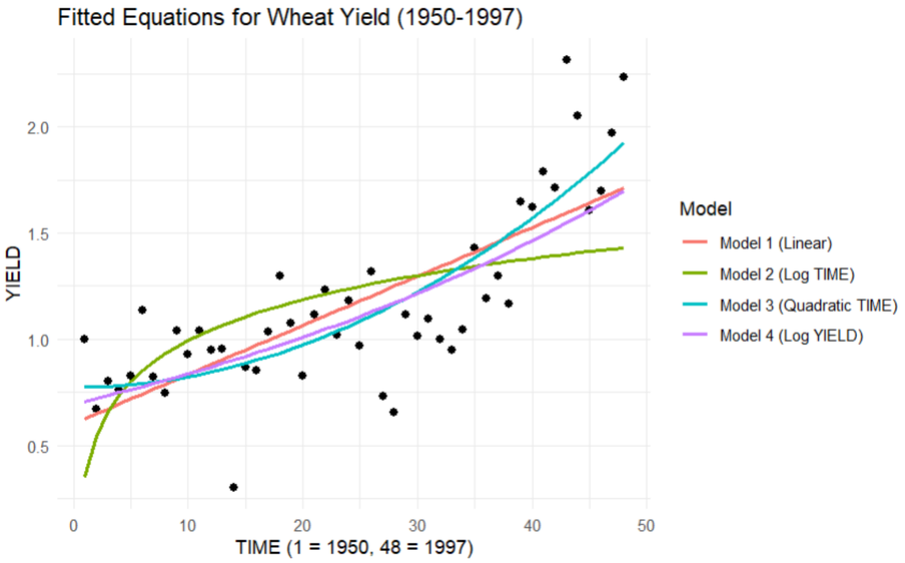
b. Interpret the coefficient of the time-related variable in your chosen specification.

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

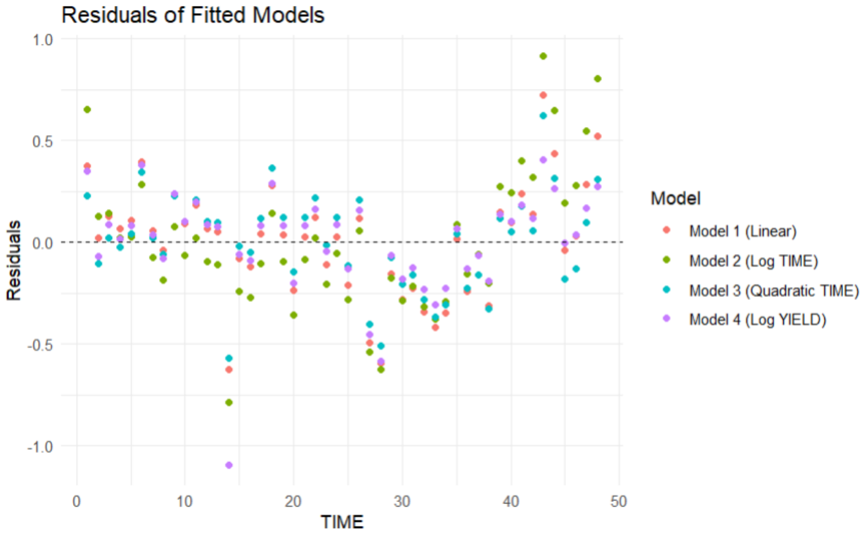
d. Using your chosen specification, use the observations up to 1996 to estimate the model. Construct a 95% prediction interval for *YIELD* in 1997. Does your interval contain the true value?

a.

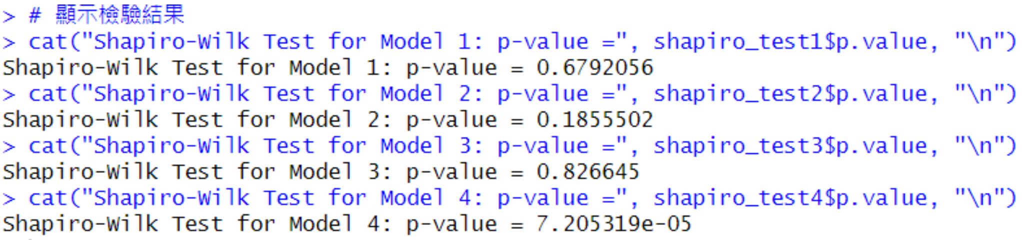
(i)



(ii)

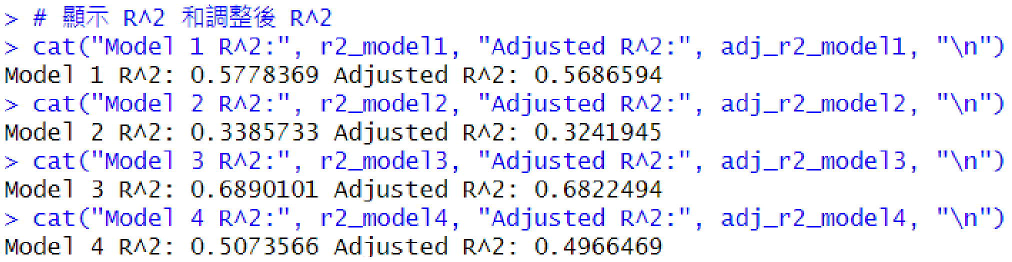


(iii)



```
> # 顯示檢驗結果
> cat("Shapiro-Wilk Test for Model 1: p-value =", shapiro_test1$p.value, "\n")
Shapiro-Wilk Test for Model 1: p-value = 0.6792056
> cat("Shapiro-Wilk Test for Model 2: p-value =", shapiro_test2$p.value, "\n")
Shapiro-Wilk Test for Model 2: p-value = 0.1855502
> cat("Shapiro-Wilk Test for Model 3: p-value =", shapiro_test3$p.value, "\n")
Shapiro-Wilk Test for Model 3: p-value = 0.826645
> cat("Shapiro-Wilk Test for Model 4: p-value =", shapiro_test4$p.value, "\n")
Shapiro-Wilk Test for Model 4: p-value = 7.205319e-05
```

(iv)



```
> # 顯示 R^2 和調整後 R^2
> cat("Model 1 R^2:", r2_model1, "Adjusted R^2:", adj_r2_model1, "\n")
Model 1 R^2: 0.5778369 Adjusted R^2: 0.5686594
> cat("Model 2 R^2:", r2_model2, "Adjusted R^2:", adj_r2_model2, "\n")
Model 2 R^2: 0.3385733 Adjusted R^2: 0.3241945
> cat("Model 3 R^2:", r2_model3, "Adjusted R^2:", adj_r2_model3, "\n")
Model 3 R^2: 0.6890101 Adjusted R^2: 0.6822494
> cat("Model 4 R^2:", r2_model4, "Adjusted R^2:", adj_r2_model4, "\n")
Model 4 R^2: 0.5073566 Adjusted R^2: 0.4966469
```

綜合考慮， Model 3 是更可取的模型。

原因如下，在圖形上只有 model 3 與 model 4 能捕捉到後段的資料點上升趨勢， 而在殘差正態性檢驗中 model 3 殘差符合正態分佈（p-value = 0.826645 > 0.05)，統計推斷可靠。相反的，model 4 的 p 值小於 0.05，顯示其殘差不符合正態分布，影響統計推斷

b.

在 Model 3 中，當 TIME=0 時，預測的 YIELD 為 0.7737 噸/公頃。

與時間相關的變量 TIME^2 的係數 γ1=0.0004986，表明時間的平方對小麥產量有顯著的正向影響，YIELD 隨時間呈拋物線增長，增長速度隨時間加快。這可能反映了技術進步在 1950-1997 年間的累積效應，特別是在後期對產量的推動作用更強。

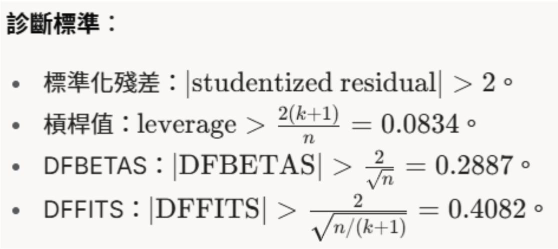
c.

Studentized Residuals（標準化殘差）：用來識別異常點。

Leverage（槓桿值）：用來識別高槓桿點。

DFBETAS：用來識別對特定係數估計有顯著影響的觀測值（針對截距和 TIME^2 係數）。

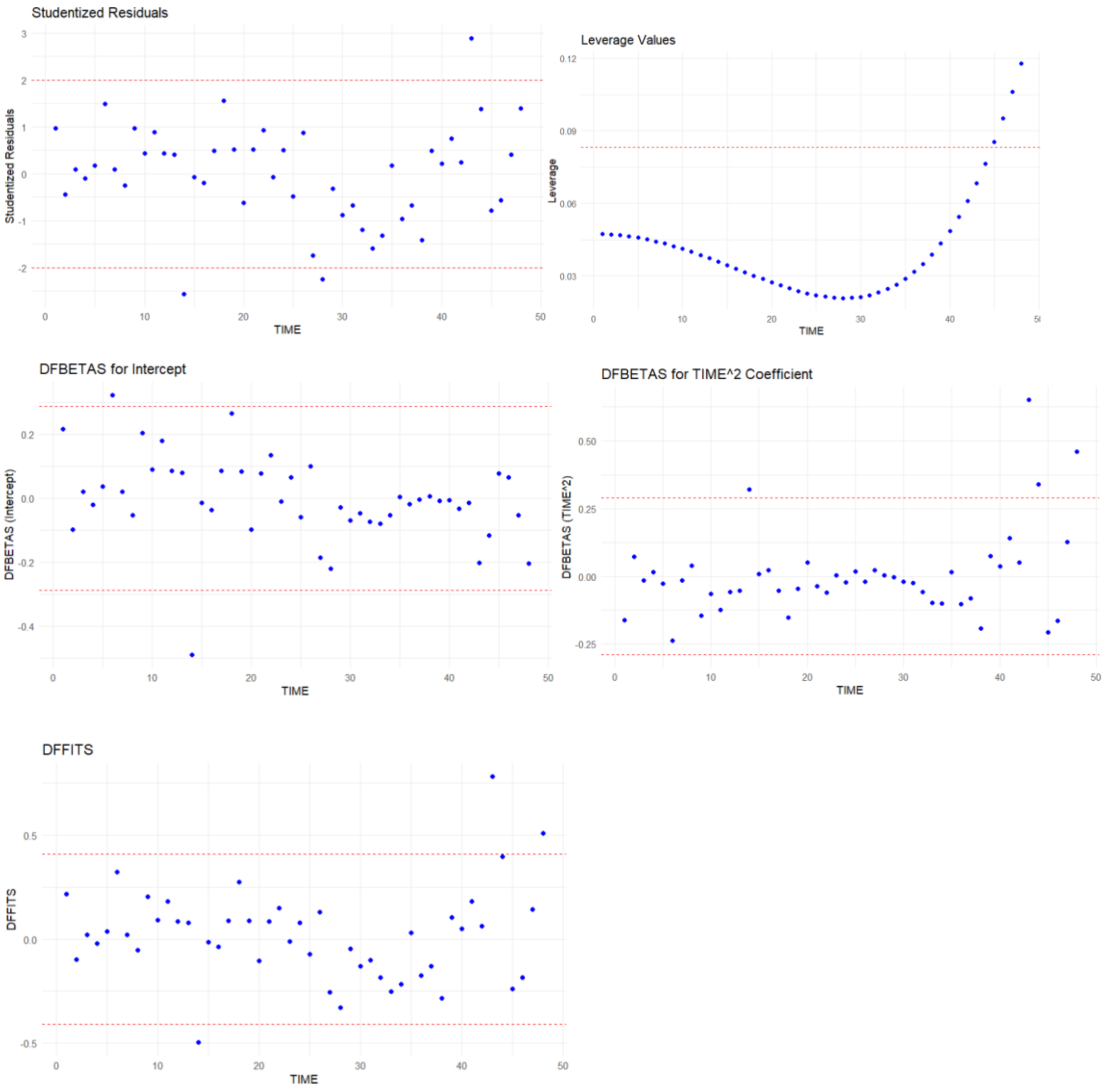
DFFITS：用來識別對預測值有顯著影響的觀測值。



診斷標準：

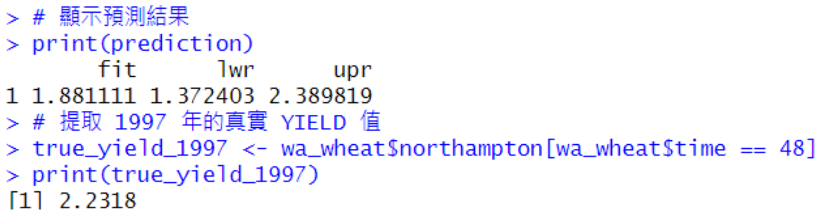
- 標準化殘差：$|\text{studentized residual}| > 2$。
- 槓桿值：leverage $> \frac{2(k+1)}{n} = 0.0834$。
- DFBETAS：$|\text{DFBETAS}| > \frac{2}{\sqrt{n}} = 0.2887$。
- DFFITS：$|\text{DFFITS}| > \frac{2}{\sqrt{n/(k+1)}} = 0.4082$。

若符合上述條件則為異常值。



由上述圖形可得知主要異常值都是出現在 time 接近 50 的數據中，model 較無法捕捉數據後段的趨勢。

d.



```
> # 顯示預測結果
> print(prediction)
       fit      lwr      upr
1 1.881111 1.372403 2.389819
> # 提取 1997 年的真實 YIELD 值
> true_yield_1997 <- wa_wheat$northampton[wa_wheat$time == 48]
> print(true_yield_1997)
[1] 2.2318
```

真實值 2.2318 介於我們的 95% 信心水準預測中。

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.

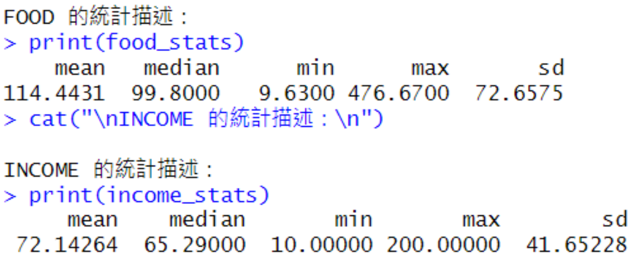
a. Calculate summary statistics for the variables: *FOOD* and *INCOME*. Report for each the sample mean, median, minimum, maximum, and standard deviation. Construct histograms for both variables. Locate the variable mean and median on each histogram. Are the histograms symmetrical and "bell-shaped" curves? 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.

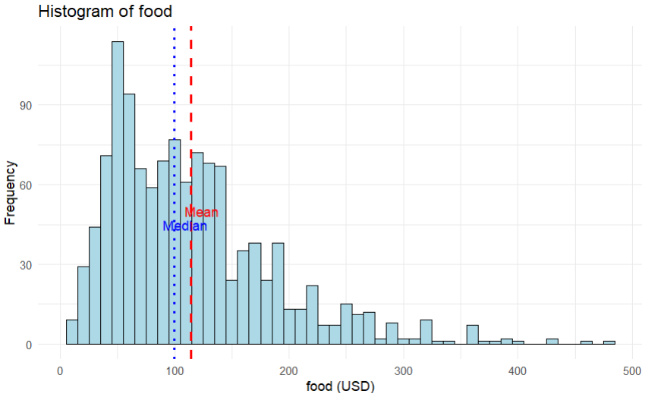
d. Calculate both a point estimate and a 95% interval estimate of the elasticity of food expenditure with respect to income at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Are the estimated elasticities similar or dissimilar? Do the interval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

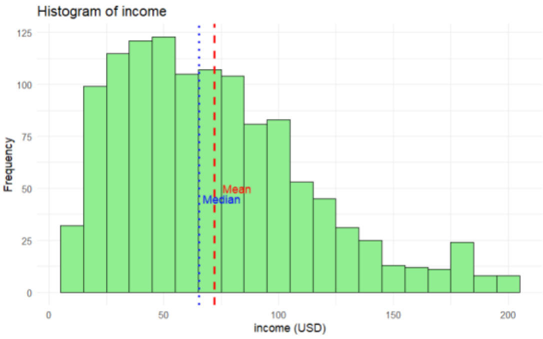
a.



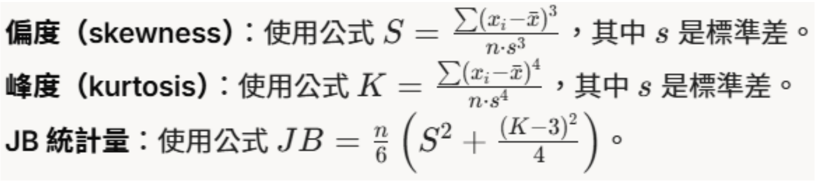
```
FOOD 的統計描述：
> print(food_stats)
    mean   median      min      max       sd
114.4431  99.8000   9.6300 476.6700  72.6575
> cat("\nINCOME 的統計描述：\n")

INCOME 的統計描述：
> print(income_stats)
     mean    median       min       max        sd
 72.14264  65.29000  10.00000 200.00000  41.65228
```





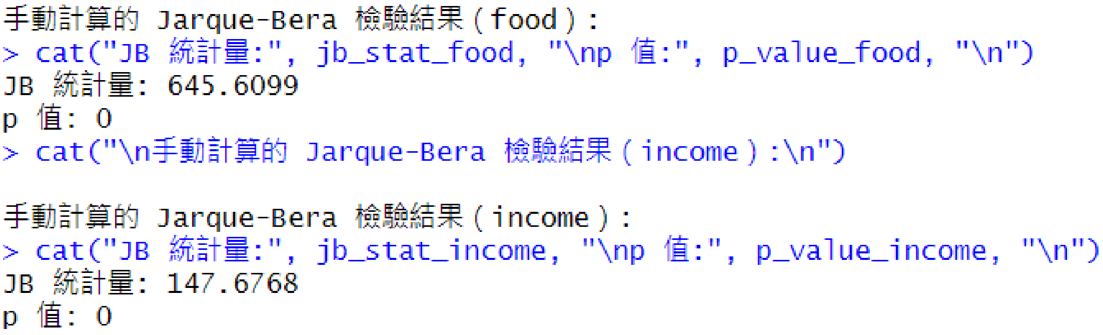
圖形可得知兩者皆為右偏而非鐘型，其平均數都大於中位數



偏度（skewness）：使用公式 $S = \frac{\sum(x_i-\bar{x})^3}{n\cdot s^3}$，其中 s 是標準差。

峰度（kurtosis）：使用公式 $K = \frac{\sum(x_i-\bar{x})^4}{n\cdot s^4}$，其中 s 是標準差。

JB 統計量：使用公式 $JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$。

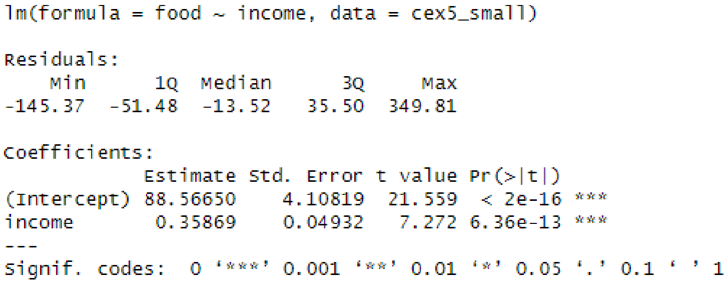


```
手動計算的 Jarque-Bera 檢驗結果（food）:
> cat("JB 統計量:", jb_stat_food, "\np 值:", p_value_food, "\n")
JB 統計量: 645.6099
p 值: 0
> cat("\n手動計算的 Jarque-Bera 檢驗結果（income）:\n")

手動計算的 Jarque-Bera 檢驗結果（income）:
> cat("JB 統計量:", jb_stat_income, "\np 值:", p_value_income, "\n")
JB 統計量: 147.6768
p 值: 0
```

由 jb 統計量與 p 值皆可得知兩者皆不符合正態分佈。

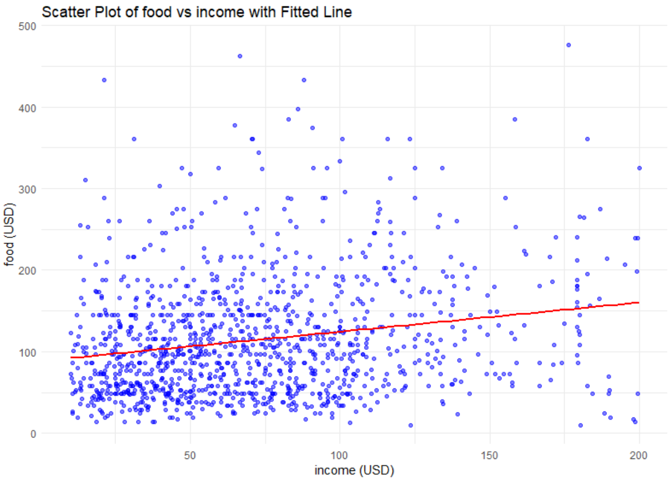
b.

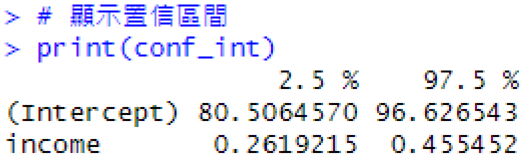


```
lm(formula = food ~ income, data = cex5_small)

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
```





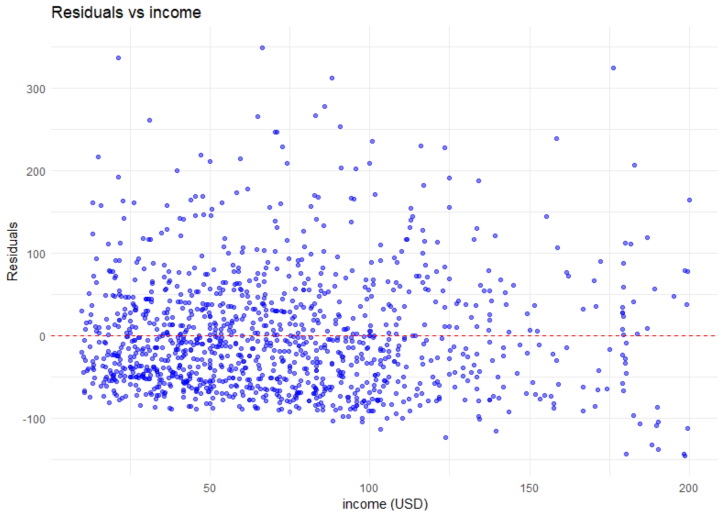
```
> # 顯示置信區間
> print(conf_int)
                 2.5 %     97.5 %
(Intercept) 80.5064570 96.626543
income       0.2619215  0.455452
```

(beta 2 為下面那行)

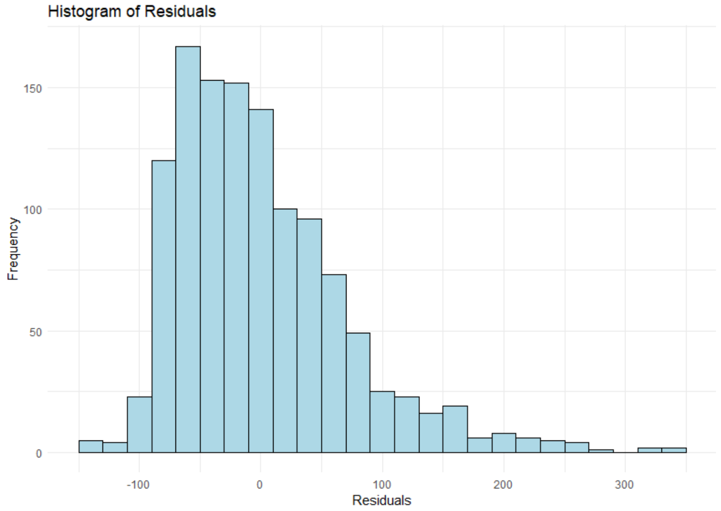
beta 2 係數為 0.35869，而信賴區間介於 0.2619-0.4554，寬度較寬，估計的不確定性較高。SE(β^2)=0.04932，相對於 β₂ 的估計值 0.35869，標準誤不

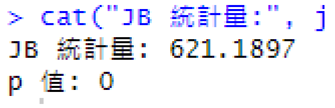
算太小（約為估計值的 13.75%)，表明估計的變異性中等，精確性一般。R^2 為 0.04 ，模型解釋力有限，估計的精確性會受到影響。因此，我們**不能完全精確地估計** income 變化對平均 food 的影響。

c.



負值的殘差主要落在-100 之內，而正的有許多大於 100，介於 100-300 之間。





```
> cat("JB 統計量:", j
JB 統計量: 621.1897
p 值: 0
```

隨機誤差項 ϵ 的正態分佈是線性回歸模型的核心假設，直接影響參數估計的統計推斷（例如置信區間和 p 值)。相比之下，food 和 income 的正態分佈並非必要條件，它們的非正態性主要影響模型的擬合效果

d.

income=19

預測 food^=88.56650+0.35869x19=88.56650+6.81511=95.38161

點彈性： E=0.35869×1995.38161=0.35869×0.19922≈0.07146

區間彈性：

- 下限：Elower=0.2619215×1995.38161≈0.2619215×0.19922≈0.05219
- 上限：Eupper=0.4554542×1995.38161≈0.4554542×0.19922≈0.09073

結果：點彈性 = 0.07146，95% 區間彈性 = 0.05219 , 0.09073

income=65

預測 food^=111.88135

E= 0.35869×0.58087≈0.20837

Elower= 0.2619215×0.58087≈0.15216

Eupper=0.4554542×0.58087≈0.26458

結果：點彈性 = 0.20837，95% 區間彈性 = 0.15216,0.264580.15216,0.264580.15216,0.26458

income=160

預測 food^= 145.95690

E= 0.39320

Elower= 0.28712

Eupper= 0.49928

結果：點彈性 = 0.39320，95% 區間彈性 = 0.28712,0.499280.28712,0.499280.28712,0.49928

點彈性隨著 income 增加而顯著增加（從 0.07146 到 0.20837 再到 0.39320）。 這表明估計彈性在不同 income 水平下**不相似**，彈性隨著收入增加而變大。

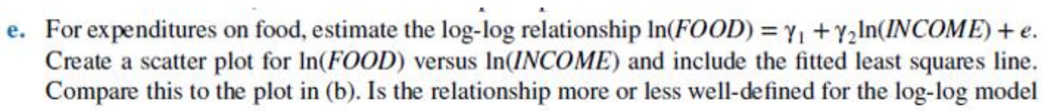
比較區間：

- `income = 19` 和 `income = 65`：0.05219,0.09073 和 0.15216,0.26458 不重疊。
- `income = 65` 和 `income = 160`：0.15216,0.26458 和 0.28712,0.49928 不重疊。
- `income = 19` 和 `income = 160`：更不可能重疊。

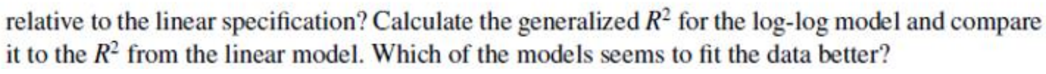
結論：三個 `income` 水平的 95% 區間估計都不重疊，表明彈性在不同收入水平下的差異在統計上顯著。

根據恩格爾定律，隨著收入增加，家庭在食品上的支出比例減少，食品支出的收入彈性應該**減少**。

本例結果：線性模型中，彈性隨著 income 增加而增加（從 0.07146 到 0.39320），這與恩格爾定律相反。

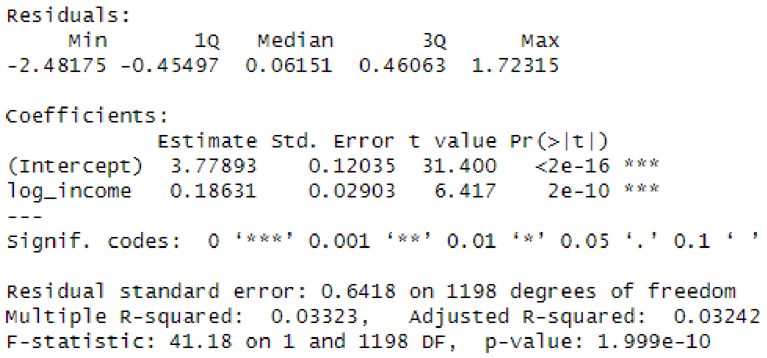


e. For expenditures on food, estimate the log-log relationship $\ln(FOOD) = \gamma_1 + \gamma_2 \ln(INCOME) + e$. Create a scatter plot for $\ln(FOOD)$ versus $\ln(INCOME)$ and include the fitted least squares line. Compare this to the plot in (b). Is the relationship more or less well-defined for the 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?

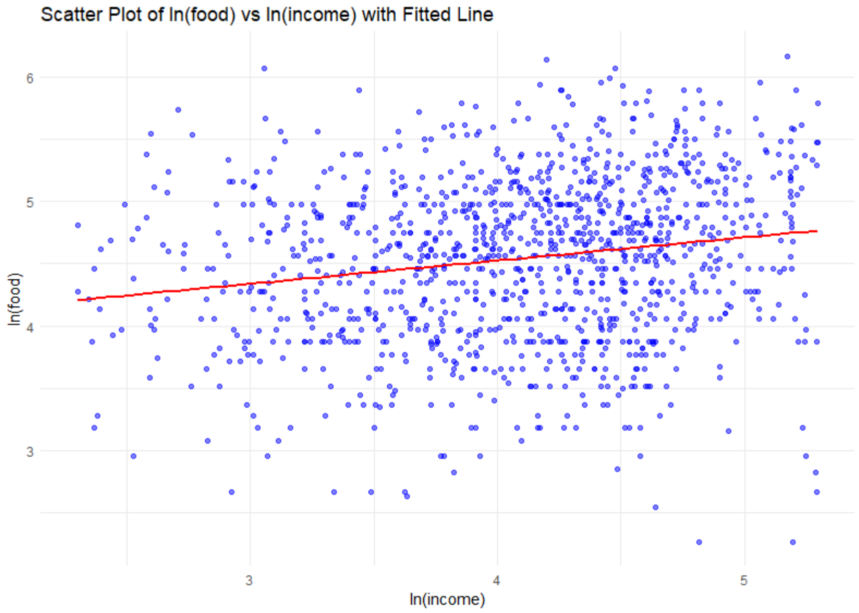
e.



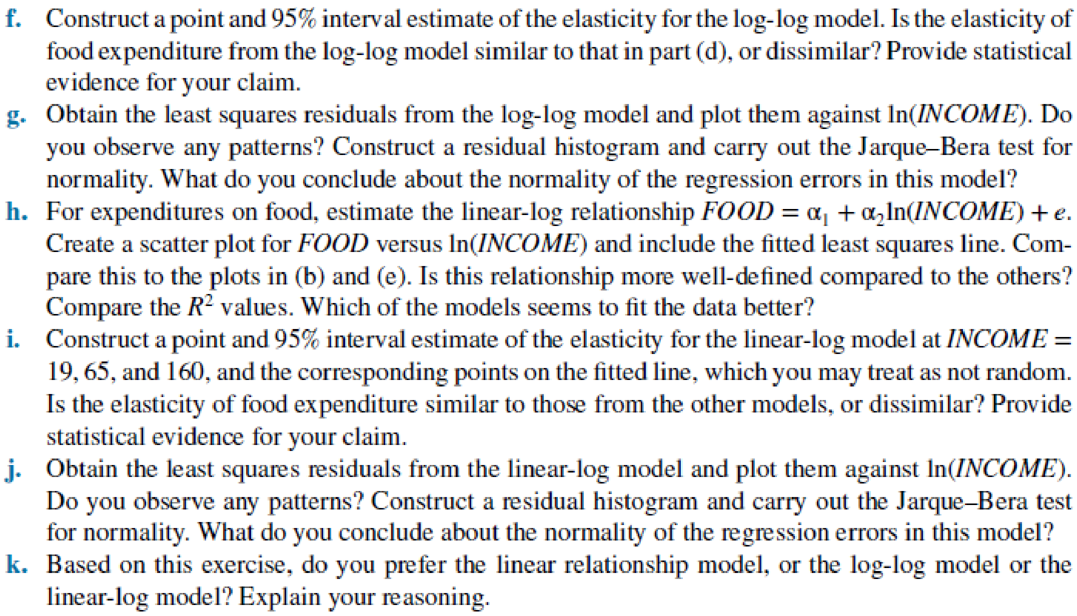
```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.48175 -0.45497  0.06151  0.46063  1.72315 

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
```



與 b 小題相比，此圖顯示更線性的關係，因為對數轉換減少了偏態。然而 r^2 值約為 0.03，仍為極小值，擬合程度兩者皆很低。



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.

f.

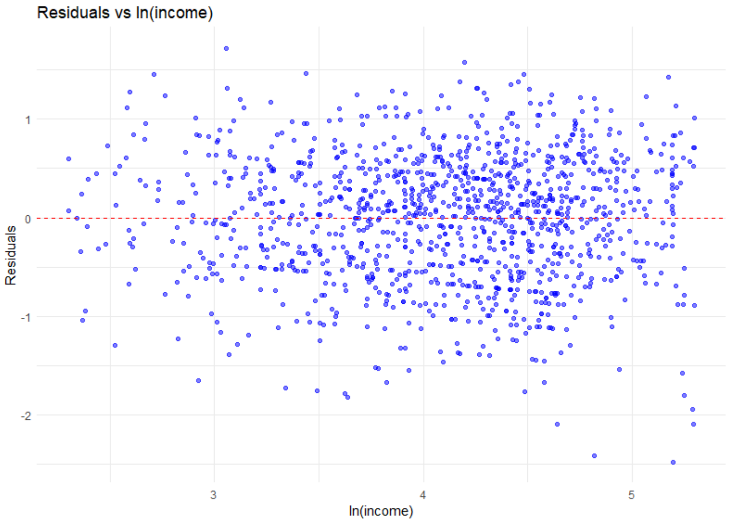
$$\ln(\hat{\text{food}}) = 3.77893 + 0.18631 \times \ln(\text{income})$$

$$E_I = \frac{\partial \ln(\text{food})}{\partial \ln(\text{income})} = \gamma_2$$

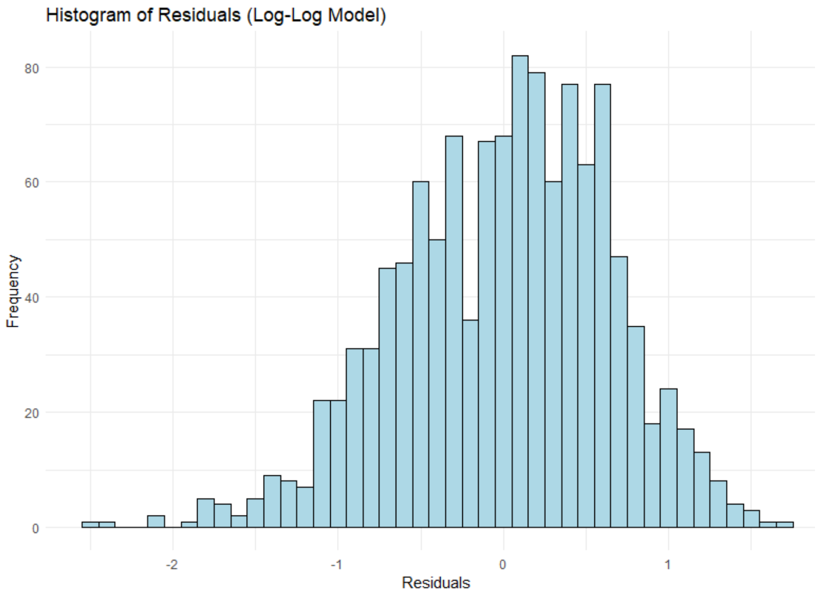
0.18631±1.96×0.02903 = [0.12941,0.24321]

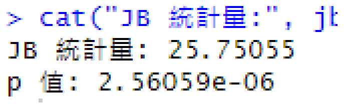
對數-對數模型的彈性（0.18631）與線性模型的彈性（0.07146 到 0.39320）不相似（dissimilar）。線性模型的彈性隨收入變化很大，而對數-對數模型的彈性為常數。

g.



對數-對數模型的殘差範圍（-2.48175 到 1.72315）顯著減小，異方差問題也減輕，表明對數轉換有效改善了模型的擬合效果。





```
> cat("JB 統計量:", jb
JB 統計量: 25.75055
p 值: 2.56059e-06
```

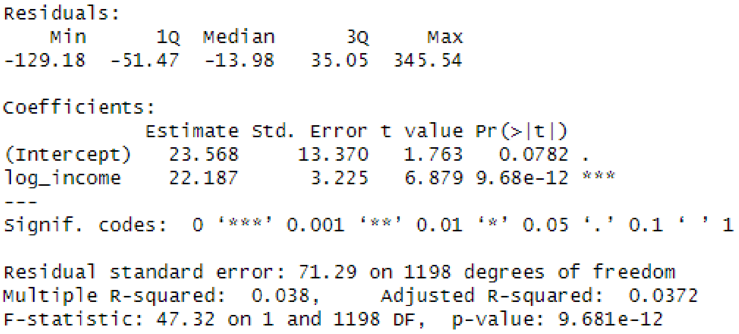
殘差直方圖：

- 殘差分佈左偏（中位數 0.06151 偏正值，左尾 -2.48175 比右尾 1.72315 更極端），不呈完美鐘形曲線，偏離正態分佈。

Jarque-Bera 檢驗：

- JB 統計量 25.75055，p 值 2.56059e-06 < 0.05，拒絕原假設，殘差不服從正態分佈。
- 這與直方圖的左偏和可能的尖峭峰度一致。

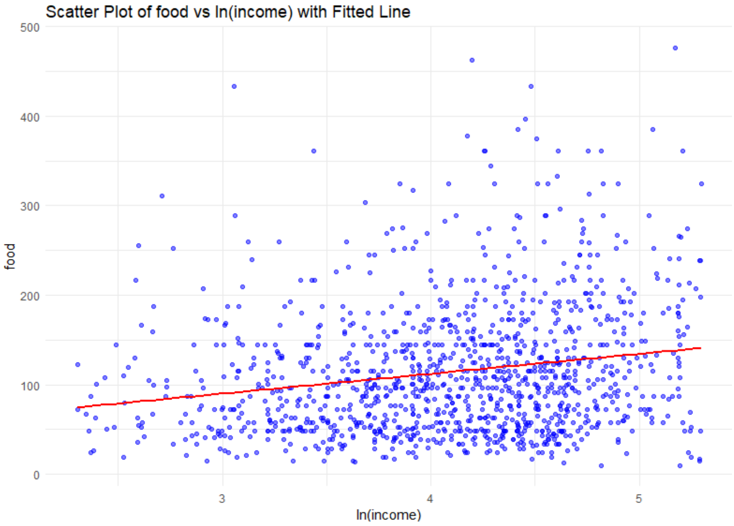
h.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-129.18  -51.47  -13.98   35.05  345.54 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   23.568     13.370   1.763   0.0782 .  
log_income    22.187      3.225   6.879 9.68e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.29 on 1198 degrees of freedom
Multiple R-squared:  0.038,	Adjusted R-squared:  0.0372 
F-statistic: 47.32 on 1 and 1198 DF,  p-value: 9.681e-12
```



散點圖顯示 FOOD 隨 ln(INCOME) 緩慢增長，但點分佈分散，擬合直線無法捕捉非線性趨勢，存在輕微異方差。

b 小題（線性模型）：關係最不清晰，非線性趨勢和強烈異方差。

e 小題（對數-對數模型）：關係最清晰， ln(FOOD) 和 ln(INCOME) 更接近線性。

h 小題（線性-對數模型）：關係介於 b 和 e 之間， FOOD 右偏導致非線性趨勢，清晰度不如 e 小題。

R^2 值為 B 小題最大，但仍為極小值