4.4

Model 1:

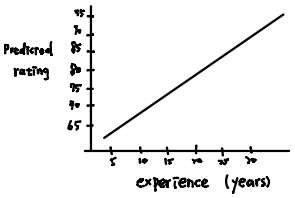
$$\widehat{RATING} = 64.289 + 0.990EXPER \quad N = 50 \quad R^2 = 0.3793$$
$$\text{(se)} \qquad (2.422) \quad (0.183)$$

Model 2:

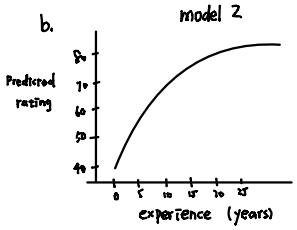
$$\widehat{RATING} = 39.464 + 15.312\ln(EXPER) \quad N = 46 \quad R^2 = 0.6414$$
$$\text{(se)} \qquad (4.198) \quad (1.727)$$

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. model 1



b. model 2



因 $\ln(0) = -\infty$
因此 4位 exper=0
不納入

c. $\frac{\partial Rating}{\partial Exper} = 0.99\Big|_{Exper=10}$ $\frac{\partial Rating}{\partial Exper} = 0.99\Big|_{Exper=20}$ ⇒ 無論 exper 多少，每增1 exper，Rating 增 0.99

d. 邊際影響為 $\frac{\partial Rating}{\partial Exper} \times \frac{1}{Exper} = 15.312 \times \frac{1}{Exper}$

exper = 10 ⇒ 邊際影響 $= 15.312 \times \frac{1}{10} = 1.5312$

exper = 20 ⇒ 邊際影響 $= 15.312 \times \frac{1}{20} = 0.7656$

e. model 1 的 $R^2 = 0.3793$
model 2 的 $R^2 = 0.6414$
⇒ R^2 越高越好
因此 model 2 更能擬合

f. model 1 為線性，但實際上，exper 的回報可能會隨時間遞減
而 model 2 用對數，更符合原理

4.28

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

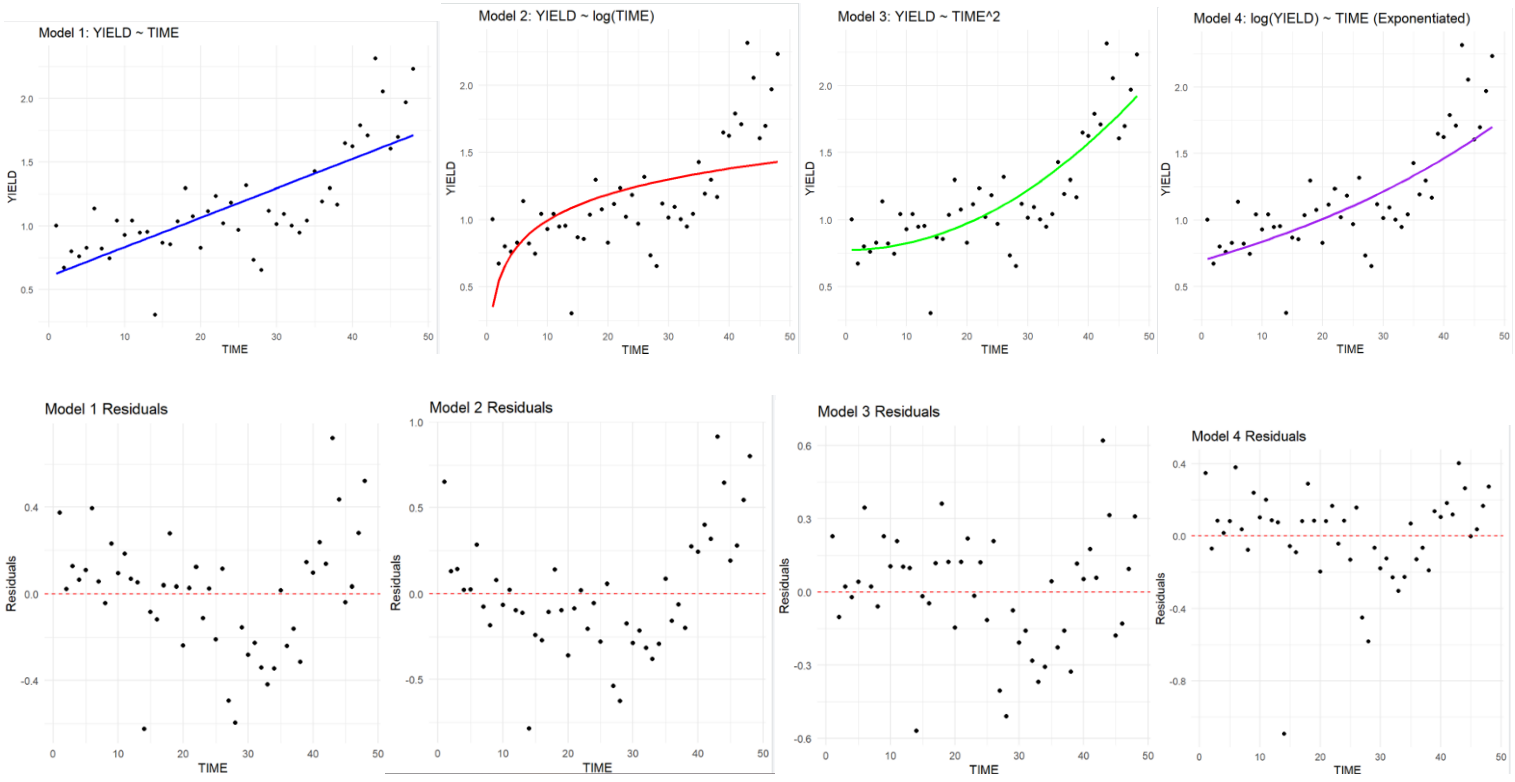
a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.

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.




```
Shapiro-Wilk Test for Normality of Residuals:
> cat("Model 1:", shapiro.test(resid(model1))$p.value, "\n")
Model 1: 0.6792056
> cat("Model 2:", shapiro.test(resid(model2))$p.value, "\n")
Model 2: 0.1855502
> cat("Model 3:", shapiro.test(resid(model3))$p.value, "\n")
Model 3: 0.826645
> cat("Model 4:", shapiro.test(resid(model4))$p.value, "\n")
Model 4: 7.205319e-05
```

```
R-squared Values:
> cat("Model 1:", summary(model1)$r.squared, "\n")
Model 1: 0.5778369
> cat("Model 2:", summary(model2)$r.squared, "\n")
Model 2: 0.3385733
> cat("Model 3:", summary(model3)$r.squared, "\n")
Model 3: 0.6890101
> cat("Model 4:", summary(model4)$r.squared, "\n")
Model 4: 0.5073566
```

除了 4 他們的 P 都>0.05，符合隨機性，而在 R^2 下 model3 最大因此 model3 比較合適

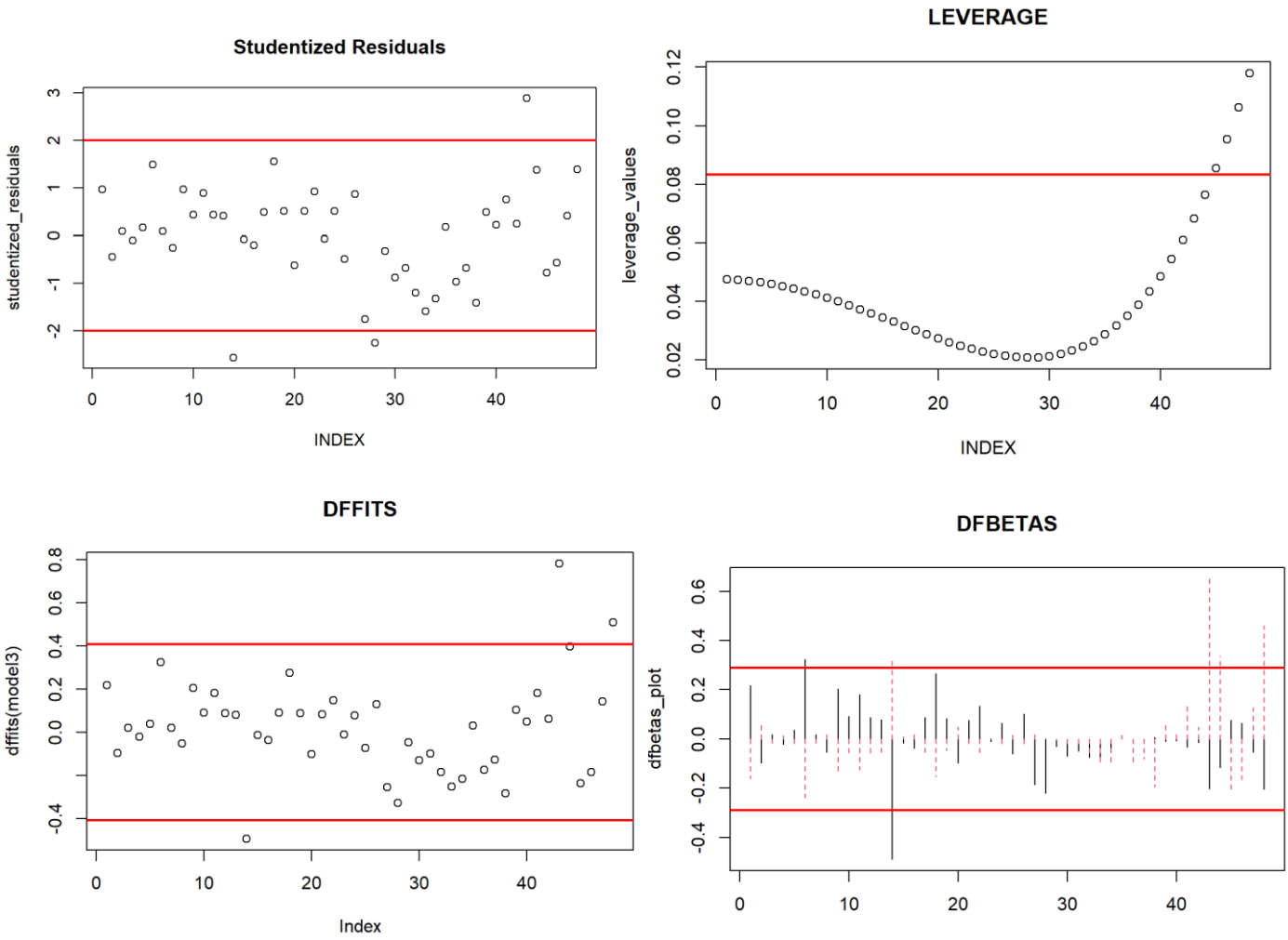
b.

```
> print(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
```

r1 顯示產量（yield）隨著時間以加速的速度增加。

r1 在統計上具有高度顯著性

c.



標準化殘差：INDEX = 47（1997 年）超過閾值。 **槓桿值**：INDEX = 47（1997 年）超過閾值。

DFFITS：INDEX = 47（1997 年）超過閾值。 **DFBETAS**：INDEX = 47（1997 年）超過閾值。

d.

```
95% Prediction Interval for YIELD in 1997:
> print(pred)
       fit      lwr      upr
1 1.881111 1.372403 2.389819
> true_yield_1997 <- northampton$YIELD[nor
> cat("True YIELD in 1997:", true_yield_19
True YIELD in 1997: 2.2318
```

利用 model3 所建構出來的模型其 95%信賴區間下預測 1997 年 yield 確實有包含真實數據

4.29

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?

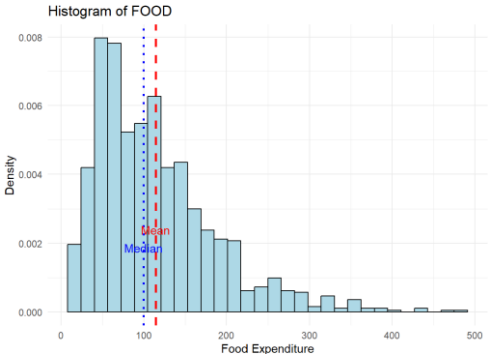
i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at *INCOME* = 19, 65, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide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

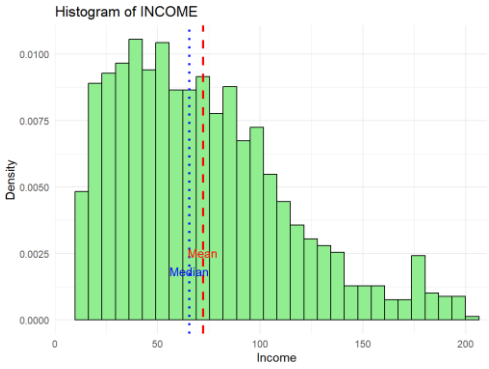
j. Obtain the least squares residuals from the linear-log model and plot them against $\ln(INCOME)$. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

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.

```
Summary Statistics for FOOD and INCOME:
> summary(cex)
      food             income
 Min.   :  9.63   Min.   : 10.00
 1st Qu.: 57.78   1st Qu.: 40.00
 Median : 99.80   Median : 65.29
 Mean   :114.44   Mean   : 72.14
 3rd Qu.:145.00   3rd Qu.: 96.79
 Max.   :476.67   Max.   :200.00
> cat("\nStandard Deviation for FOOD an

Standard Deviation for FOOD and INCOME:
> sapply(cex, sd)
    food   income
72.65750 41.65228
```





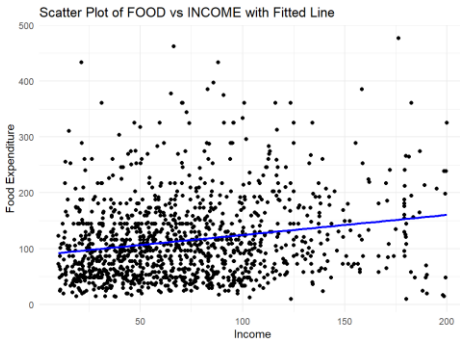

```
Jarque-Bera Test for FOOD:
> print(jarque.bera.test(cex

        Jarque Bera Test

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

```
Jarque-Bera Test for INCOME:
> print(jarque.bera.test(cex

        Jarque Bera Test

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

沒有呈現鐘形曲線 P-value 都小於 0.05 拒絕虛無假設，因此不服從常態分配

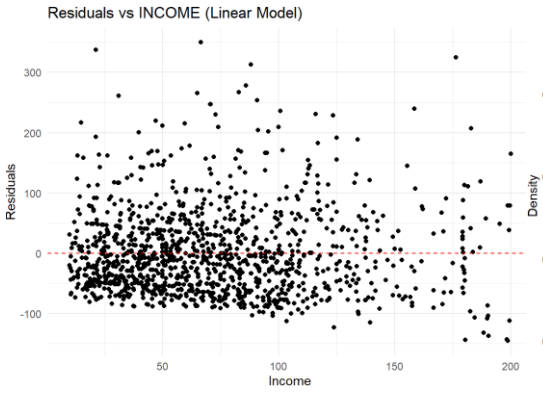
b.

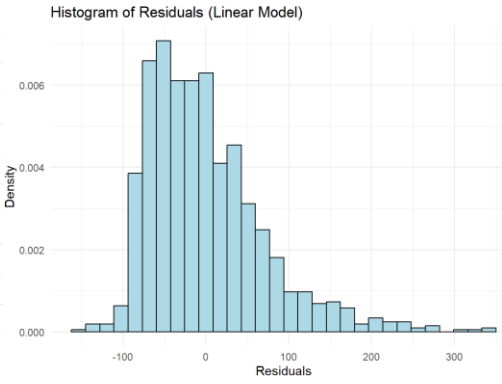
```
lm(formula = food ~ income, data = cex)

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

Coefficients:
            Estimate Std. Error
(Intercept) 88.56650    4.10819
income       0.35869    0.04932
            t value Pr(>|t|)
(Intercept)  21.559  < 2e-16 ***
income        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
```




```
95% Confidence Interval for beta_2:
> print(confint_beta2)
              2.5 %    97.5 %
income 0.2619215 0.455452
```

c.






```
Jarque-Bera Test for Residuals (Linear Model):
> print(jarque.bera.test(cex$resid_linear))

        Jarque Bera Test

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

對隨機誤差（殘差）服從常態分佈的要求比 **FOOD** 和 **INCOME** 服從常態分佈更重要。這是因為最小平方估計（Least Squares Estimation, LSE）假設誤差項服從常態分佈，以確保推論的有效性（例如信賴區間、假設檢定）。如果殘差不服從常態分佈，統計檢定的結果可能會不可靠。然而，**FOOD** 和 **INCOME** 本身不需要服從常態分佈，因為模型的重點在於它們之間的關係，而不是它們各自的邊際分佈。

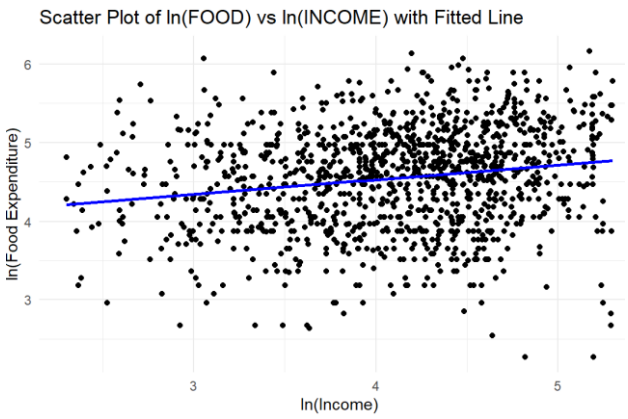
d.

	INCOME	b_1 + b_2*INCOME	ϵ	se(ϵ)	LB	UB
1	19	95.3815	0.0715	0.0098	0.0522	0.0907
2	65	111.8811	0.2084	0.0287	0.1522	0.2645
3	160	145.9564	0.3932	0.0541	0.2872	0.4992

根據恩格爾定律（Engel's Law），隨著收入增加，家庭用於食品的支出比例會下降，這意味著食品支出的所得彈性應該隨著收入的增加而降低。

然而，在線性模型中，彈性是 **INCOME/FOOD** 的比例。由於 **FOOD** 與 **INCOME** 呈線性關係，這可能導致所得彈性隨 **INCOME** 增加（對應的數值為 **0.07145038, 0.2083876, 0.3931988**）。這樣的結果可能與恩格爾定律不一致，因為理論上，隨著收入增加，食品支出的所得彈性應該下降，而不是上升。

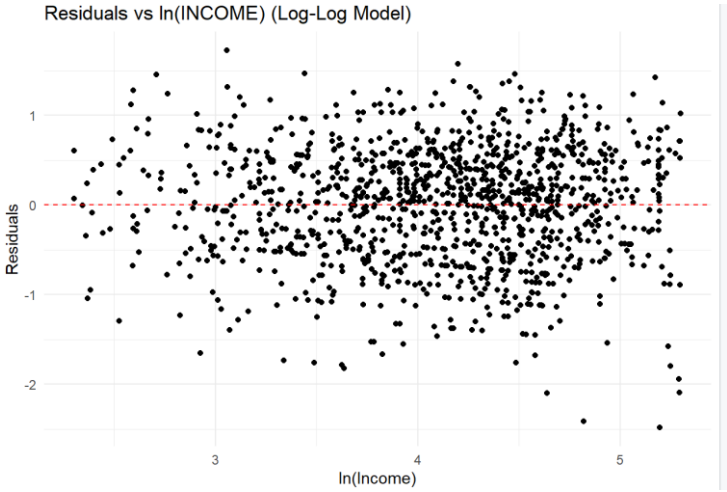
e.

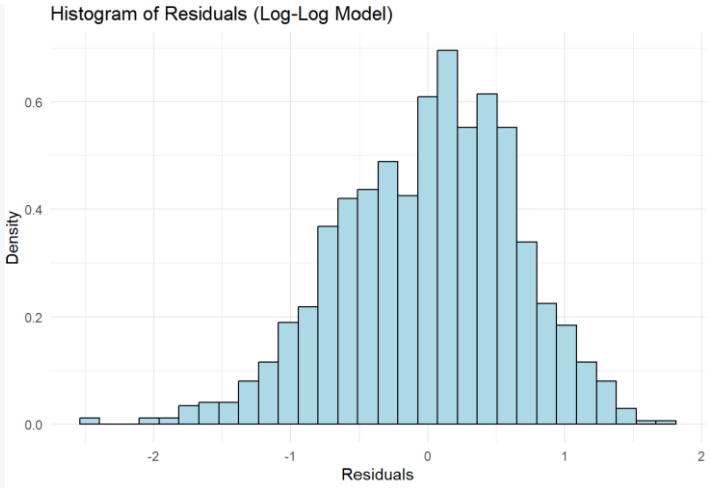


f.

```
95% Confidence Interval for Elasticity: ( 0.1293432 , 0.2432675 )
```

g.

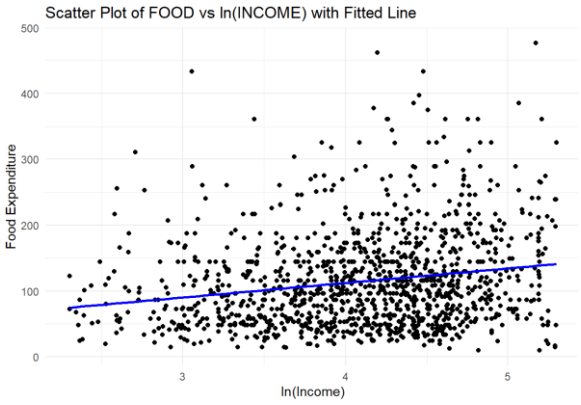





```
Jarque-Bera Test for Residuals (Log-Log Model):
> print(jarque.bera.test(cex$resid_loglog))

        Jarque Bera Test

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

Jarque-Bera 檢定的 p 值小於 0.05，表示殘差不服從常態分佈，這可能會影響統計推論的可靠性

h.



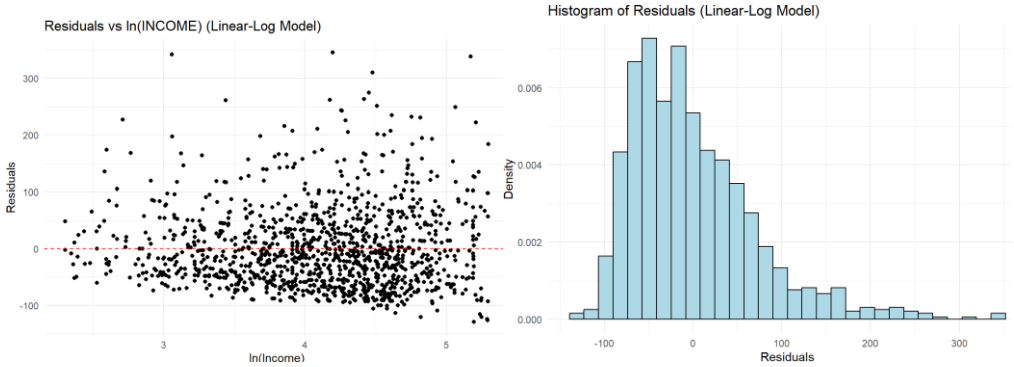

```
R-squared for Linear-Log Model:
> print(summary(model_linearlog
[1] 0.03799984
```

最高 R 平方（或對數-對數模型的廣義 R 平方）的模型更能貼合數據。

i.

	INCOME	α_1 + α_2*ln(INCOME)	ϵ	se(ϵ)	LB	UB
1	19	88.8979	0.2496	0.0363	0.1785	0.3207
2	65	116.1872	0.1910	0.0278	0.1366	0.2454
3	160	136.1733	0.1629	0.0237	0.1165	0.2094

j.




```
Jarque-Bera Test for Residuals (Linear-Log Model)
> print(jarque.bera.test(cex$resid_linearlog))

        Jarque Bera Test

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

Jarque-Bera 檢定的 p 值小於 0.05，殘差不服從常態分佈，這可能會影響統計推論的可靠性。

在解釋變異方面，對數-對數模型和線性-對數模型具有最高 R 平方，對數據的擬合效果最佳。此外，對數-對數模型和線性-對數模型更符合恩格爾定律，因為它們的彈性分別是恆定的（對數-對數模型）或隨著收入增加而降低（線性-對數模型），而線性模型的彈性則隨收入增加。基於這些標準，線性-對數模型通常是食品支出數據的較佳選擇，因為它在擬合度和經濟可解釋性之間取得了平衡。