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.

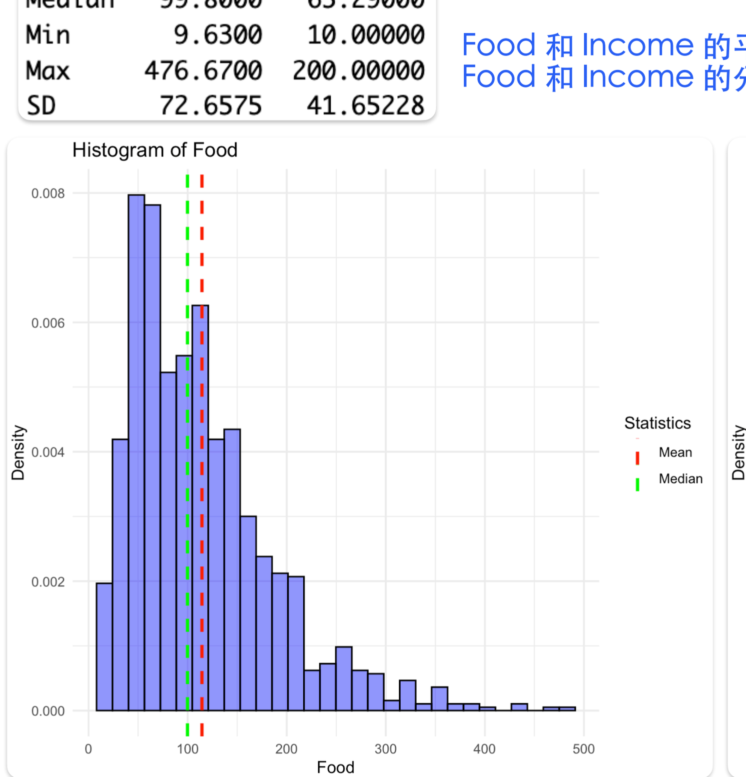
j. Obtain the least squares residuals from the linear-log model and plot them against ln(*INCOME*). Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

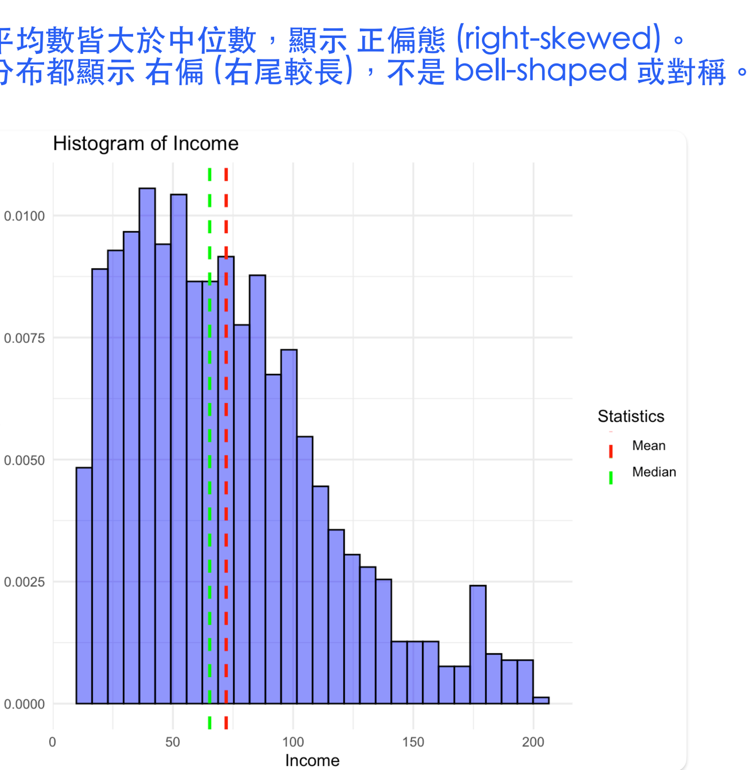
k. Based on this exercise, do you prefer the linear relationship model, or the log-log model or the linear-log model? Explain your reasoning.

a.

```
> print(stats_table)
             Food      Income
N       1200.0000  1200.00000
Mean     114.4431    72.14264
Median    99.8000    65.29000
Min        9.6300    10.00000
Max      476.6700   200.00000
SD        72.6575    41.65228
```

Food 和 Income 的平均數皆大於中位數，顯示 正偏態 (right-skewed)。
Food 和 Income 的分布都顯示 右偏 (右尾較長)，不是 bell-shaped 或對稱。





Jarque-Bera (JB) 常態性檢定
H0 ：資料符合常態分布　H1 ：資料不符合常態分布
臨界值 (5% 顯著水準)：$\chi^2(2)=5.99$
結論：因為 JB 統計量 (food: 648. 65, income: 148.21) 都遠大於 5.99 ，且 P 值 = 0.00 (小於 0.05)，拒絕 H0 ，表示 food 和 income 都不符合常態分布。

```
> 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.

```
> summary(lm_model)

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

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
```

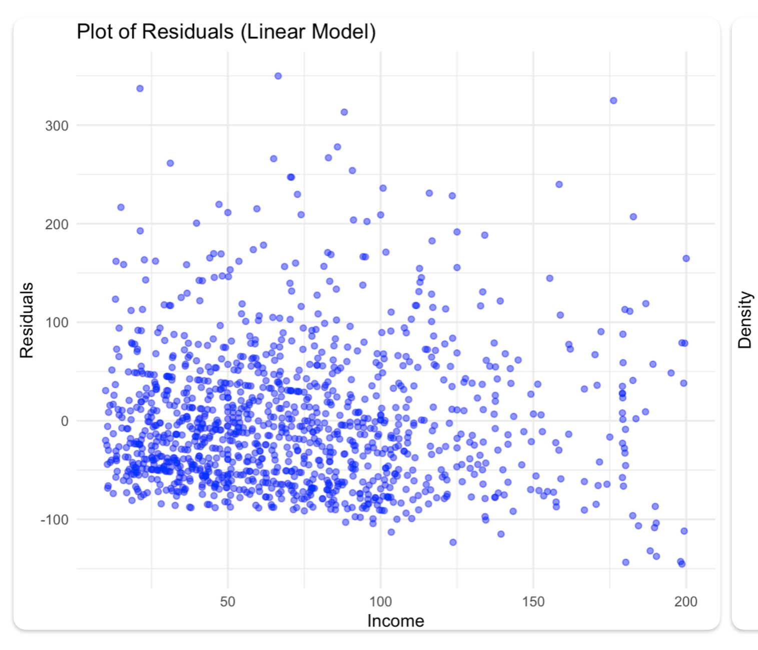
Linear Model:
FOOD = 88.5665 + 0.3587*INCOME

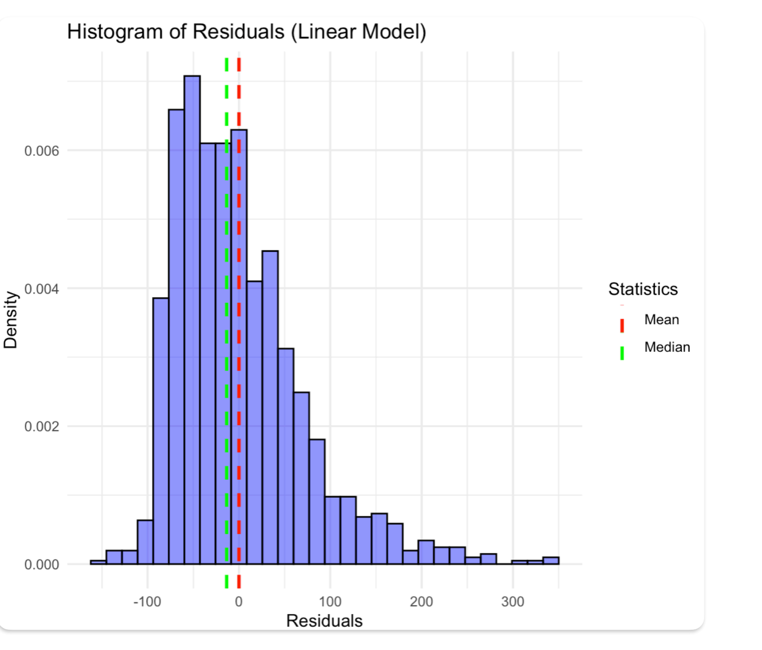
ß2 95% 信賴區間為 [0.2619, 0.4555]

從標準誤與信賴區間來看，Linear Model 對收入變動對食物支出的影響估計相對精確；但由於 R^2 很低，表示收入對食物支出的解釋力有限，可能還需要考慮其他影響因素，因此模型的整體預測能力仍然較弱。

```
> conf_interval <- confint(lm_model, level = 0.95)
> print(conf_interval)
                  2.5 %    97.5 %
(Intercept) 80.5064570 96.626543
income       0.2619215  0.455452
```

c.





由殘差散布圖可知，Linear Model 的殘差未呈現隨機分佈。
由殘差直方圖可知，Linear Model 的殘差呈現正偏態。
Skewness = 1.30 > 0 : 右偏 (Right-skewed)
Kurtosis = 4.40 > 3 : 高峰分佈 (Leptokurtic)

```
> cat("Skewness:", skewness_value_lm, "\n")
Skewness: 1.29554
> cat("Kurtosis:", kurtosis_value_lm, "\n")
Kurtosis: 5.402088
```

Jarque-Bera (JB) 常態性檢定
H0 ：資料符合常態分布　H1 ：資料不符合常態分布
臨界值 (5% 顯著水準)：$\chi^2(2)=5.99$
結論：因為 JB 統計量 (624.19) 遠大於 5.99，且 P 值 = 0.00 (小於 0.05)，拒絕 H0 ，表示 Linear Model 的殘差不符合常態分布。

```
	Jarque Bera Test

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

隨機誤差項 e的常態性比 FOOD 和 INCOME 的常態性更重要，因為 Assumptions of the Simple Linear Regression Model 中SR 6: Error Normality (optional) ei | x ~ N(0,σ²)。

d.

```
> print(elasticity_table_lm)
  INCOME  FOOD_hat          ε       se_ε ε_lower_bound ε_upper_bound
1     19  95.38155 0.07145038 0.00982475    0.05219423    0.09070654
2     65 111.88114 0.20838756 0.02865423    0.15222630    0.26454882
3    160 145.95638 0.39319883 0.05406661    0.28723022    0.49916745
```

- Linear Model 彈性估計比較：彈性 (ε) 隨著 INCOME 的增加而上升， 食品支出的收入彈性在不同收入水準下顯著不同。
- 彈性的信賴區間沒有重疊，表示不同收入群體的食品支出行為在統計上具有顯著差異。
- 從 經濟學預期 來看，食物為必需品，食物的 所得彈性 (income elasticity) 通常應該隨收入上升而下降或趨於穩定，但這次的結果顯示 彈性隨收入上升而增加，這與經濟學的一般預測不完全一致。

e.

```
> summary(log_log_model)  # 顯示回歸結果

Call:
lm(formula = ln_food ~ ln_income, data = data_log)

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 ***
ln_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
```

Log-Log Model:
ln(FOOD) = 3.7789 + 0.1863*ln(INCOME)

比較 Linear Model 與 Log-Log Model 散佈圖，看不出何者更符合數據分佈。

Log-Log Model 的 Generalized R²（0.0397）接近但仍低於 Linear Model 的 R²（0.0423）。這表示 Log-Log Model 的預測力與 Linear Model 差異不大，但仍略低。
在考慮 R ² 下，Linear Model 的解釋能力最好。

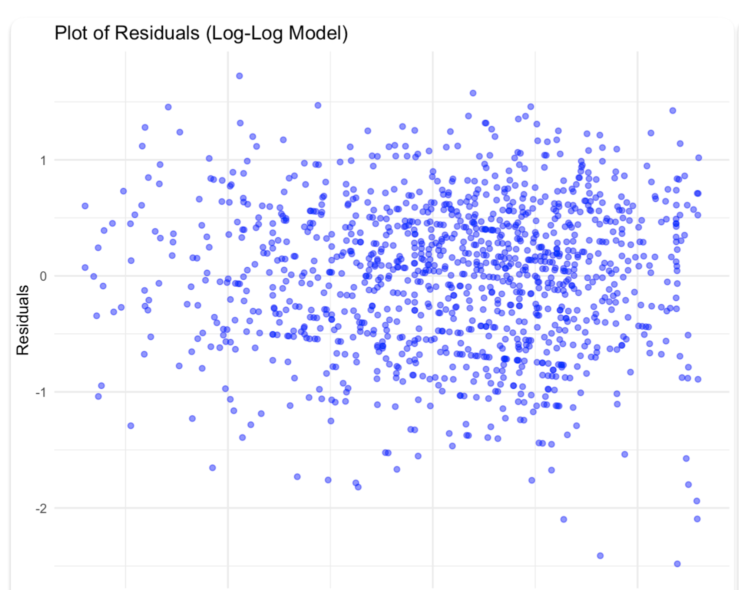
```
> print(r2_comparison_e)
          Model         R2 Generalized_R2
1  Linear Model 0.04228120             NA
2 Log-Log Model 0.03322915     0.03965161
```

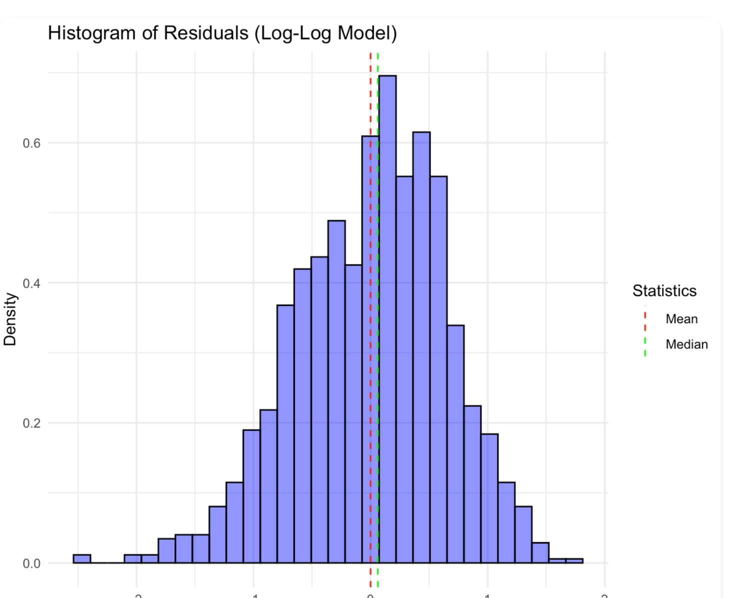
f.

```
> print(elasticity_table_loglog)
          Elasticity ε_Lower_Bound ε_Upper_Bound
ln_income     0.1863        0.1293        0.2433
```

- Log-Log Model 的彈性固定為 0.1863，其 95% 信賴區間為 [0.1293, 0.2433]。
- Log-Log Model 的彈性固定，但 Linear Model 的彈性會隨收入增加而上升。
  當 INCOME = 19，兩模型的彈性信賴區間不重疊，拒絕 H0。
  當 INCOME = 65，兩模型的彈性信賴區間重疊，不拒絕 H0。
  當 INCOME = 160，兩模型的彈性信賴區間不重疊，拒絕 H0。
- Log-Log Model 的彈性估計 只在中間收入區間與 Linear Model 一致，在低／高收入明顯不同，表示 Log-Log Model 的彈性估計在兩端與 Linear Model 在統計上顯著不同。

g.





由殘差散布圖可知，Log-Log Model 的殘差整體分佈較為均勻。
由殘差直方圖可知，Log-Log Model 的殘差呈現些微的負偏態。
Skewness = -0.36 < 0 : 左偏 (Left-skewed)
Kurtosis = 3.07 ≈ 3 : 接近常態分佈

```
> cat("Skewness:", skewness_value_loglog, "\n")
Skewness: -0.3577097
> cat("Kurtosis:", kurtosis_value_loglog, "\n")
Kurtosis: 3.071935
```

Jarque-Bera (JB) 常態性檢定
H0 ：資料符合常態分布　H1 ：資料不符合常態分布
臨界值 (5% 顯著水準)：$\chi^2(2)=5.99$
結論：因為 JB 統計量 (25.85) 遠大於 5.99，且 P 值 = 2.436e-06 (小於 0.05)，拒絕 H0 ，表示 Log-Log Model 的殘差不符合常態分布。

```
	Jarque Bera Test

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

Log-Log Model 的殘差比 Linear Model 更接近常態分布，但仍非常態分佈。

h.

```
> summary(linear_log_model)

Call:
lm(formula = food ~ ln_income, data = data_linear_log)

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 .  
ln_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
```

Linear-Log Model:
FOOD = 23.568 + 22.187*ln(INCOME)

比較三模型散佈圖， 看不出何者更符合數據分佈。

與 Log-Log Model 的 Generalized R²（0.0397）、Linear Model 的 R²（0.0423）相比，Linear-Log Model 的 R²（0.038）是三者中最小的。這表示在考慮模型擬合度的情況下， Linear-Log Model 的解釋能力是三者中最差的， Linear Model 的解釋能力是三者中最好的。

```
> print(r2_comparison_h)
                          Model         R2
1                  Linear Model 0.04228120
2 Log-Log Model(Generalized R²) 0.03965161
3              Linear-Log Model 0.03799984
```

i.

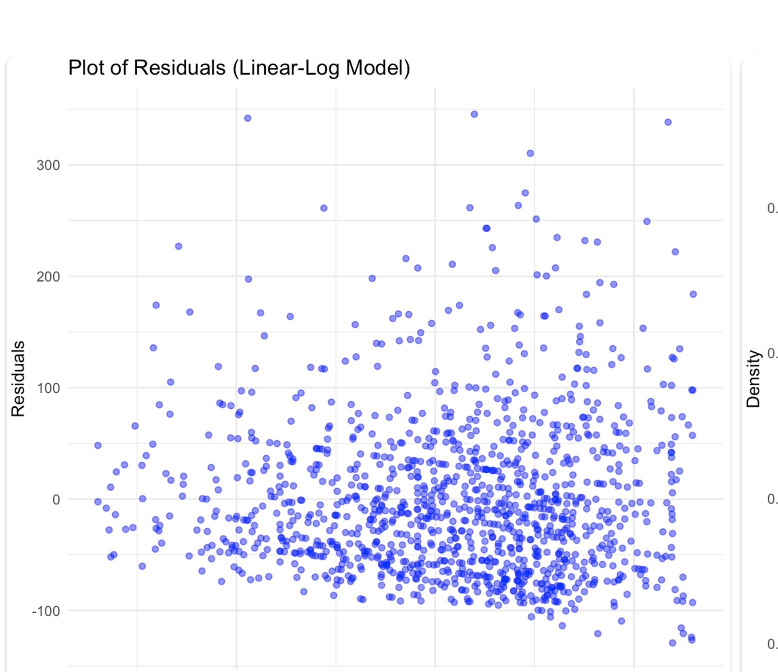
```
> print(elasticity_table_i)
  INCOME  FOOD_hat         ε       se_ε ε_lower_bound ε_upper_bound
1     19  88.89788 0.2495828 0.03628131     0.1784728     0.3206929
2     65 116.18722 0.1909624 0.02775978     0.1365542     0.2453705
3    160 136.17332 0.1629349 0.02368549     0.1165122     0.2093576
```

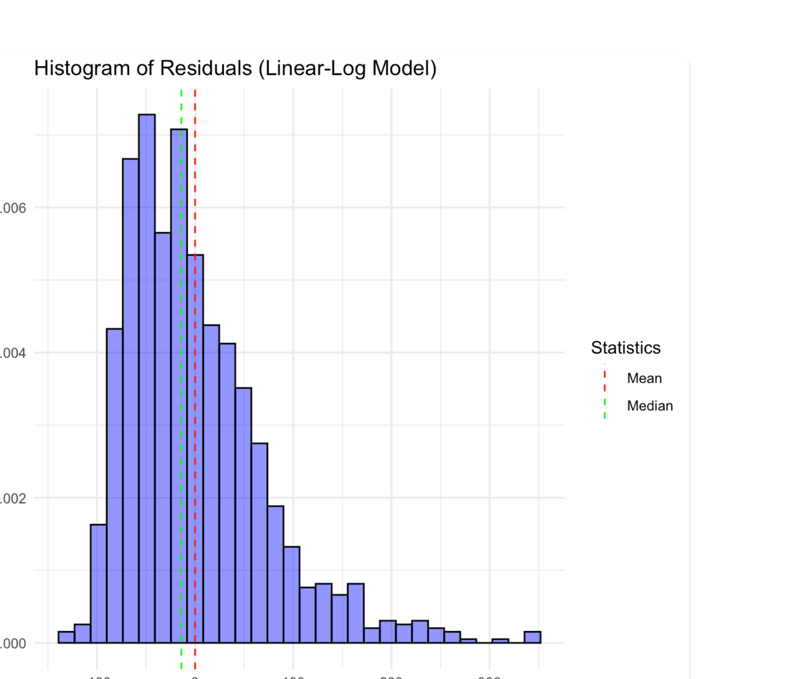
- Linear-Log Model 彈性估計比較：彈性 (ε) 隨著 INCOME 的增加而下降。
- 從 經濟學預期 來看，食物為必需品，食物的 所得彈性 (income elasticity) 通常應該 隨收入上升而下降或趨於穩定，但這次的結果顯示 彈性隨收入上升而下降，這與經濟學的一般預測一致。
- Linear Model 的彈性的彈性則隨收入增加而上升， Log-Log Model 的彈性固定，而 Linear-Log Model 的彈性隨收入增加而減小。

- Linear-Log Model vs Log-Log Model：在所有收入區間，兩模型的彈性信賴區間皆重疊，表示 Linear-Log Model 的彈性估計所有收入區間與 Log-Log Model 一致。

- Linear-Log Model vs Linear Model：
  當 INCOME = 19，兩模型的彈性信賴區間不重疊，拒絕 H0。
  當 INCOME = 65，兩模型的彈性信賴區間重疊，不拒絕 H0。
  當 INCOME = 160，兩模型的彈性信賴區間不重疊，拒絕 H0。
- Linear-Log Model 的彈性估計 只在中間收入區間與 Linear Model 一致，在低／高收入明顯不同，表示 Linear-Log Model 的彈性估計在兩端與 Linear Model 在統計上顯著不同。

j.





由殘差散布圖可知， Linear-Log Model 的殘差未呈現隨機分佈。
由殘差直方圖可知， Linear-Log Model 的殘差呈現正偏態。
Skewness = 1.31 > 0 : 右偏 (Right-skewed)
Kurtosis = 5.39 > 3 : 高峰分佈 (Leptokurtic)

```
> cat("Skewness:", skewness_value, "\n")
Skewness: 1.308532
> cat("Kurtosis:", kurtosis_value, "\n")
Kurtosis: 5.390057
```

Jarque-Bera (JB) 常態性檢定
H0 ：資料符合常態分布　H1 ：資料不符合常態分布
臨界值 (5% 顯著水準)：$\chi^2(2)=5.99$
結論：因為 JB 統計量 (628.07) 遠大於 5.99，且 P 值 < 2.2e-16 (小於 0.05)，拒絕 H0 ，表示 Linear-Log Model 的殘差不符合常態分布。

```
	Jarque Bera Test

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

k. 從 R² 值來看，三模型的擬合度相當，皆偏低。
Linear Model 估計的所得彈性會隨收入增加而上升，不符合經濟學預期。
Linear-Log Model 雖然符合經濟學理論，但殘差的分佈模式並非理想的隨機散佈。
Log-Log Model 假設所得彈性在所有收入水準下皆為固定值，可能限制了靈活性。
然而， Log-Log Mode l的殘差分佈最接近隨機分佈，且根據偏態（skewness）與峰度（kurtosis），其殘差的非常態性最小。
基於這些理由，對 Log-Log Model 似乎是較好的選擇。.