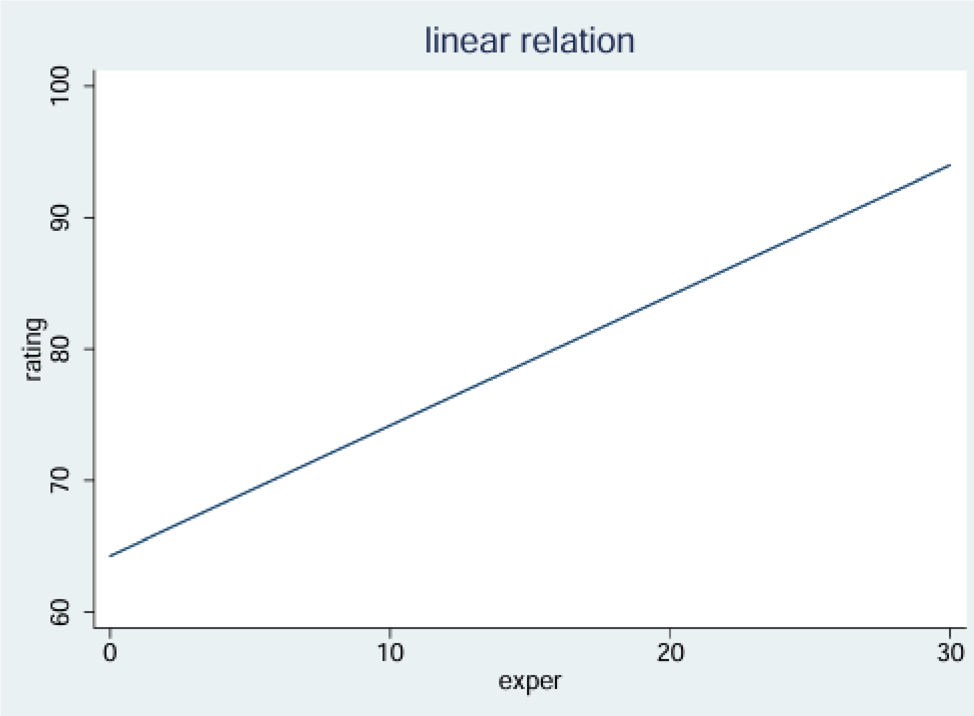
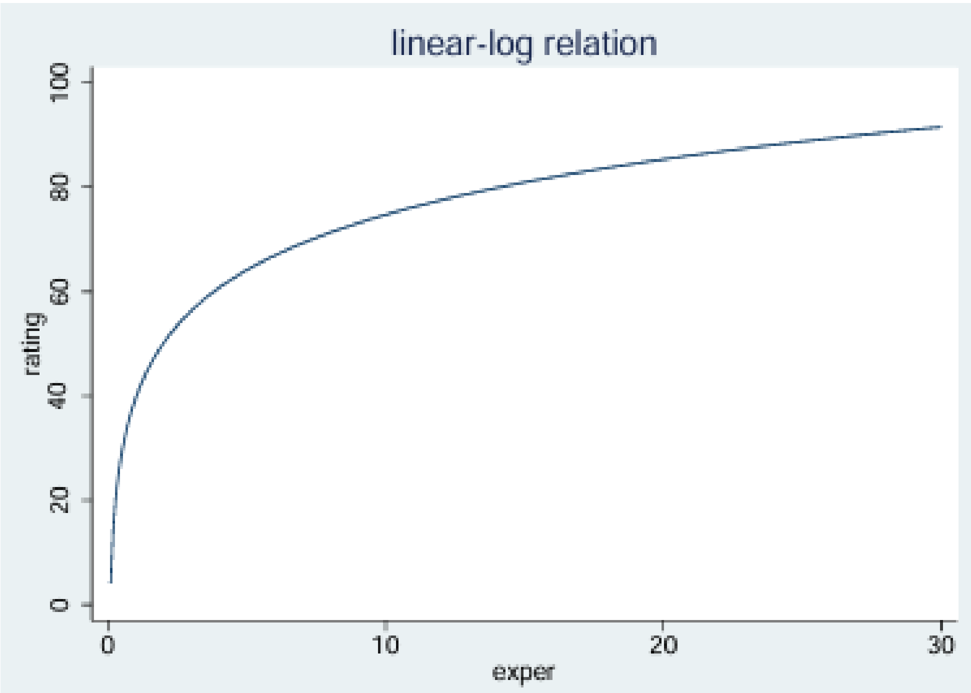
EXERCISE 4.4

(a) The graph for the linear model is



(b) The graph for the linear-log model is



By taking the log of experience, the four artists with no experience cannot be used in the estimation of Model 2 because ln(0) is undefined.

(c) Using the linear model, the marginal effect of experience on rating is a constant, 0.990. That is, each additional year of experience is estimated to increase rating by 0.99 points.

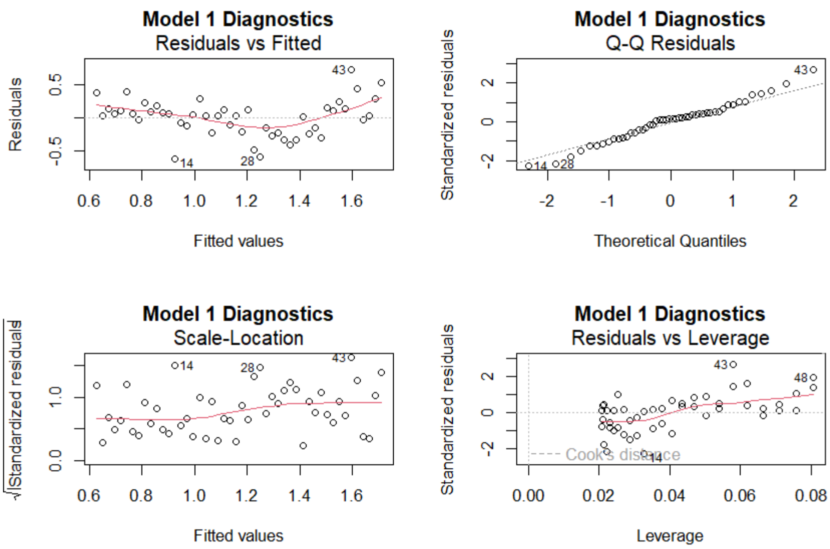
(d) For the linear-log model the marginal effect, or the slope of the relationship, is different at each point. For an artist with 10 years of experience the marginal effect of experience on rating is estimated to be 15.321/10 = 1.532 points. For an artist with 20 years of experience the marginal effect of experience on rating is estimated to be 15.321/20 = 0.766 points.

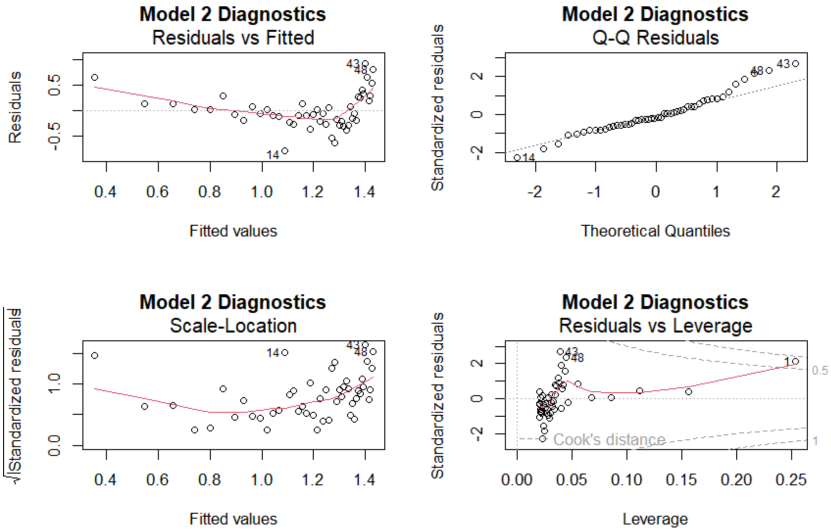
(e) Using R^2 as our guide for goodness of fit is valid here because each model has the same dependent variable. For the linear model $R^2 = 0.3793$. Using just the technical artists with experience yields $R^2 = 0.4858$ which is larger than that for the linear model. In the linear-log model there are fewer observations because some had no experience and $R^2 = 0.6414$. Thus, we would conclude that the linear-log model fits these data better.

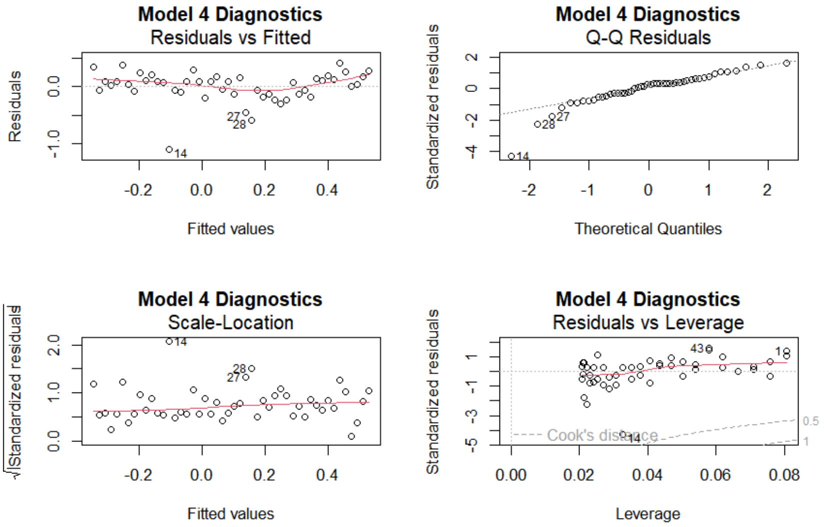
(f) Model two is more plausible. Model 1 suggests that each year of experience is associated with the same increase in rating. Model 2 suggests diminishing returns to experience, with the marginal effect of experience on rating larger for less experienced workers and less for more experienced workers.

EXERCISE 4.28

(a)







\

```
        Shapiro-Wilk normality test

data:  resid(model1)
W = 0.98236, p-value = 0.6792

        Shapiro-Wilk normality test

data:  resid(model2)
W = 0.96657, p-value = 0.1856
```

```
        Shapiro-Wilk normality test

data:  resid(model3)
W = 0.98589, p-value = 0.8266


        Shapiro-Wilk normality test

data:  resid(model4)
W = 0.86894, p-value = 7.205e-05
```

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

(b) 模型 4 中，time 的係數為 0.01863235 。解釋：當 time 增加 1 單位時，ln(northampton) 平均增加 0.01863235 單位，可近似解釋為 northampton yield 平均變化約 1.86 %（適用於變化較小的情形）。

(c)

```
> cat("Studentized residuals 超過門檻的觀測：", outlier_res, "\n")
Studentized residuals 超過門檻的觀測： 14 28
> cat("Leverage 超過門檻的觀測：", outlier_lev, "\n")
Leverage 超過門檻的觀測：
> cat("DFBETAS 超過門檻的觀測：", outlier_dfb, "\n")
DFBETAS 超過門檻的觀測： 1 6 14 43
> cat("DFFITS 超過門檻的觀測：", outlier_dffits, "\n")
DFFITS 超過門檻的觀測： 1 14
```

(d)

```
=== 1997 年預測 ===
> cat("預測的 95% prediction interval (northampton yield)：\n")
預測的 95% prediction interval (northampton yield)：
> print(pred)
       fit       lwr      upr
1 1.659434 0.9625655 2.860815
```

```
1997 年實際 yield 值： 2.2318
> if(actual_1997 >= pred[,"lwr"] && actual_1997 <= pred[,"upr"]){
+   cat("實際值落在 95% prediction interval 內。\n")
+ } else {
+   cat("實際值未落在 95% prediction interval 內。\n")
+ }
實際值落在 95% prediction interval 內。
```

EXERCISE 4.29

(a)

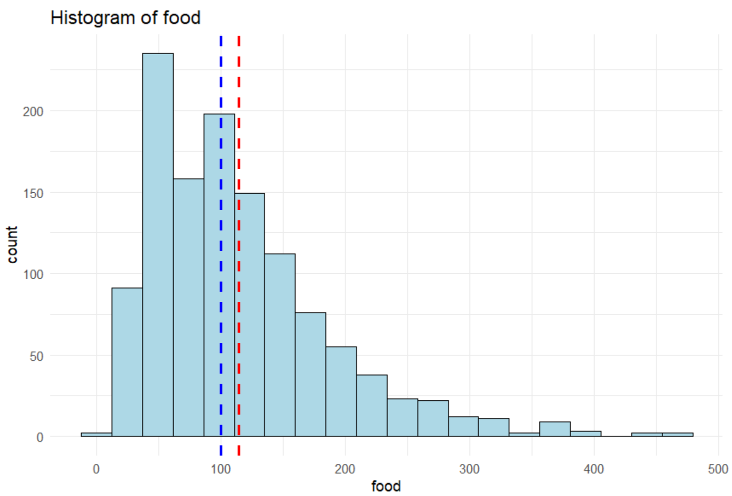
==== Summary Statistics ====

food:

Mean = 114.44 , Median = 99.8 , Min = 9.63 , Max = 476.67 , SD = 72.66

income:

Mean = 72.14 , Median = 65.29 , Min = 10 , Max = 200 , SD = 41.65



Jarque–Bera test for food:

Jarque-Bera Normality Test

data: cex5_small$food

JB = 648.65, p-value < 2.2e-16

alternative hypothesis: greater

Jarque–Bera test for income:

Jarque-Bera Normality Test

data: cex5_small$income

JB = 148.21, p-value < 2.2e-16

alternative hypothesis: greater

- **直方圖形狀：**
  從 ggplot2 繪製的直方圖來看，**food** 的分布並非完全對稱，雖然接近鐘型，但仍有輕微右偏；而 **income** 的分布則呈現較明顯的右偏態，未形成完整的「鐘型」曲線。

- **平均數與中位數：**
  對於 **food** 變數，樣本平均數略大於中位數；而對 **income** 而言，平均數遠大於中位數，這與收入分布右偏的特性相符。

- **Jarque–Bera 檢定結果：**
  進一步的 Jarque–Bera 測試也支持這些觀察，兩個變數均拒絕常態分布的假設，顯示其分布偏離理想的對稱鐘型曲線。

這些結果說明：

1. 直方圖並不完全呈現對稱、鐘型的常態分布；
2. 樣本平均數通常大於中位數，特別是 income，由於極端較大值使得平均數拉高。
3. Jarque – Bera **檢定背景：**
    1. Jarque – Bera (JB) 檢定是一種用於檢查資料是否符合常態分布的統計方法，根據樣本的**偏度** (skewness) 與**峰度** (kurtosis) 進行計算。

2. 檢定的 **虛無假設 (H0)**：資料服從常態分布。
3. **對立假設 (H1)**：資料不服從常態分布。

4. **檢定結果解讀：**
   1. JB = 648.65, p-value < 2.2e-16 (food)
   2. JB = 148.21, p-value < 2.2e-16 (income)
      這兩個 p-value 都極小（遠小於常見的顯著水準 0.05 或 0.01），表示有足夠證據拒絕「資料服從常態分布」的假設，也就是說，這兩組資料（food 與 income）**在統計上顯著地偏離常態分布**。
5. **代表意涵：**
   1. 檢定統計量 JB 值越大，代表偏度或峰度（或兩者）越明顯地偏離理想的常態分布；再加上 p-value 極小，因此可以推斷這些資料不符合常態分布。
   2. 「alternative hypothesis: greater」是 R 函式中對於檢定統計量大於某臨界值的默認表達方式，意即「資料顯著地不呈常態」。
6. **可能原因與影響：**
   1. 資料可能存在 **右偏 (right skewed)** 或 **厚尾 (heavy-tailed)** 等特徵，造成與常態分布差異顯著。
   2. 在後續分析中，如果要使用假設常態分布的統計方法（例如某些參數檢定或線性迴歸的殘差常態性檢查），需留意這些變數或殘差實際上並不常態，可能需要進行適當的轉換（如對數轉換）或改用更穩健的方法。

總結而言，Jarque–Bera 檢定結果顯示 **food** 與 **income** 的分布皆與常態分布有顯著差異，說明這些變數具備明顯的偏態或峰度特徵。

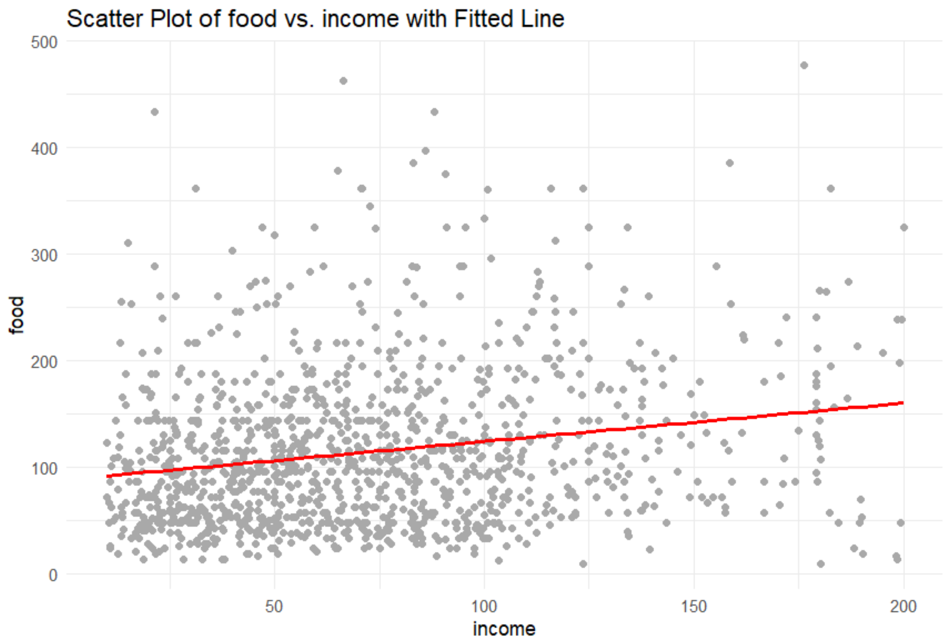
(b)

```
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
```

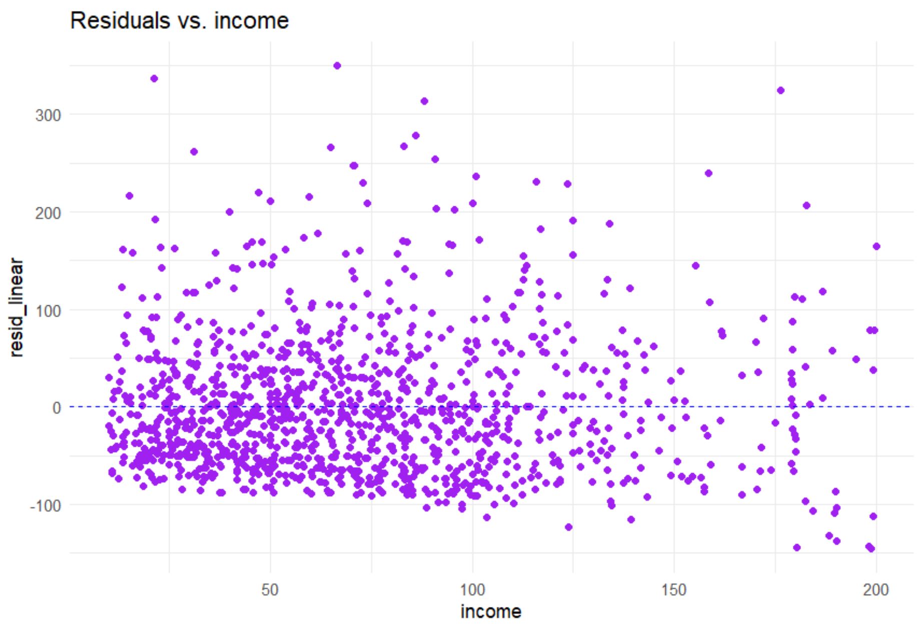


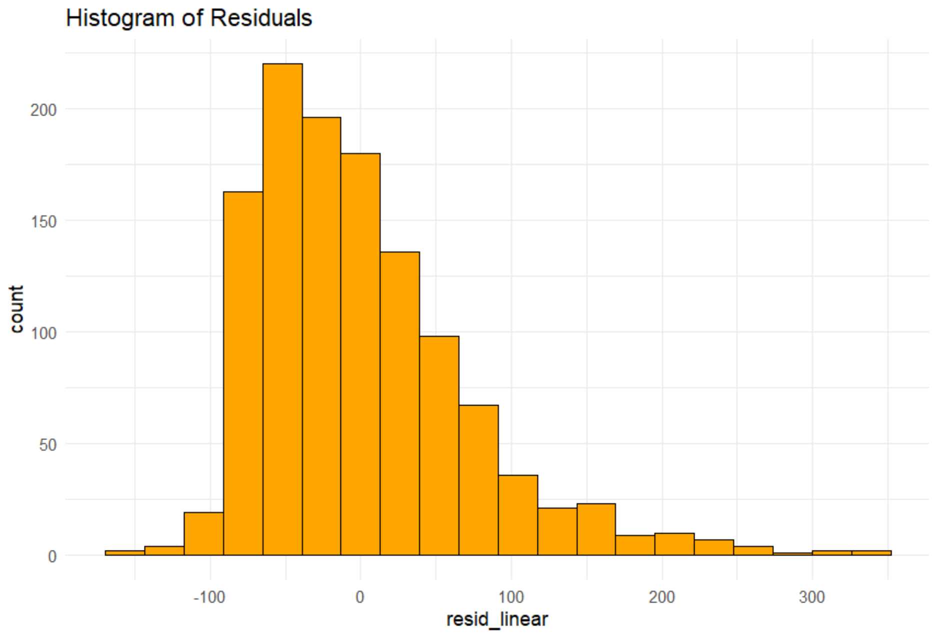
95% Confidence Interval for β2 (income coefficient):

	2.5 %	97.5 %
Income	0.2619215	0.455452

若 $\beta2$ 的信賴區間相對狹窄，則表示改變 income 對 food 的平均效應估計相對精確。

(c)





Jarque–Bera test for Residuals:

data: resid_linear

JB = 624.19, p-value < 2.2e-16

alternative hypothesis: greater

殘差圖與直方圖有助於檢查是否有系統性結構（如異質性或非線性），而 Jarque－Bera 檢定檢查誤差 e 的常態性。對於迴歸推論來說，關鍵在於隨機誤差需近似常態，而 food 與 income 本身不必要求常態分布。

In regression analysis, it is more critical that the random error term e be normally distributed than the variables (FOOD and INCOME) themselves. Here's why:

Normality of Errors:

The normality assumption for the error term underpins many of the inferential statistics in linear regression (such as t-tests and confidence intervals). If the errors are normally distributed, the sampling distributions of the estimators can be assumed to be normal (especially in small samples), which is essential for valid hypothesis testing and constructing confidence intervals.

Distribution of Variables:

The independent (and even the dependent) variables do not necessarily need to be normally distributed. OLS estimators remain unbiased and consistent regardless of the distribution of these variables, as long as the key assumptions (like linearity, independence, and homoscedasticity) are met. Non-normal variables may, however, indirectly affect the error distribution if the relationship is misspecified, but the primary focus for inference is on the error term.

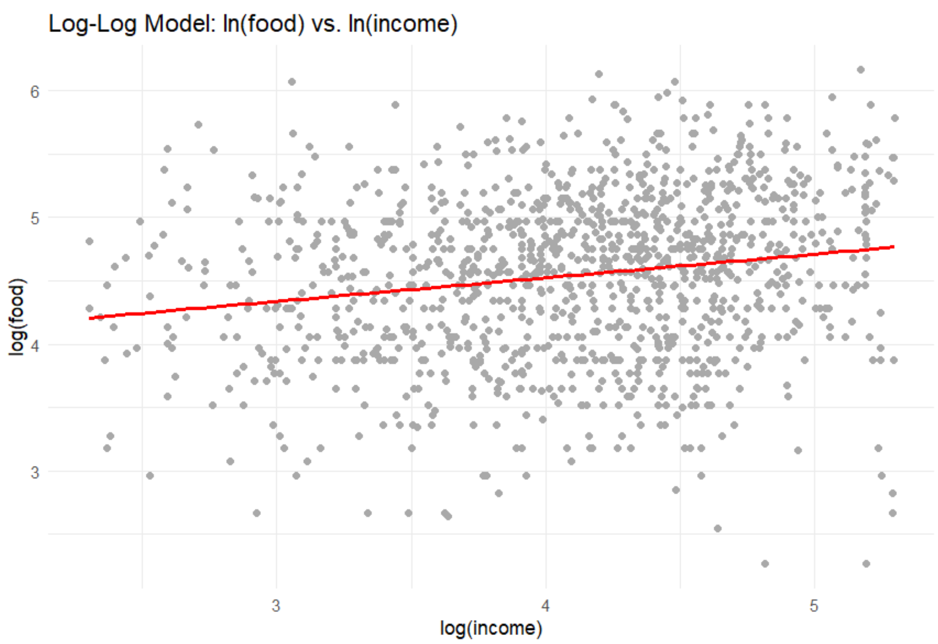
In summary, for the validity of regression inference, ensuring that the error term e is approximately normally distributed is more important than the normality of FOOD and INCOME themselves.

(d)

```
==== (d) Elasticity Estimates ====
> print(elasticity_df)
  income elasticity         se      lower     upper
1     19 0.07145038 0.01188812 0.04814966 0.0947511
2     65 0.20838756 0.02955888 0.15045216 0.2663230
3    160 0.39319883 0.04275257 0.30940380 0.4769939
```

計算結果顯示在 income = 19, 65, 160 時，food 支出對 income 的彈性估計值，以及各點估計值的 95% 信賴區間。根據經濟學原則，隨著 income 增加，一般 food 作為必需品其彈性會降低（即比例下降），而不同 income 水準下彈性值是否顯著不同，可藉由信賴區間是否重疊加以檢視。

(e)



==== (e) Model Fit Comparison ====

R^2 for linear model (food ~ income): 0.0422812

Generalized R^2 for log-log model (ln(food) ~ ln(income)): 0.03322915

Interpretation:

從散佈圖及 R^2 值比較，若 log-log 模型的 R^2 較高且資料點在轉換後呈現較明顯線性關係，則可認為 log-log 模型比線性模型更能清楚定義兩者之間的關係。

(f)

==== (f) Elasticity for Log-Log Model ====

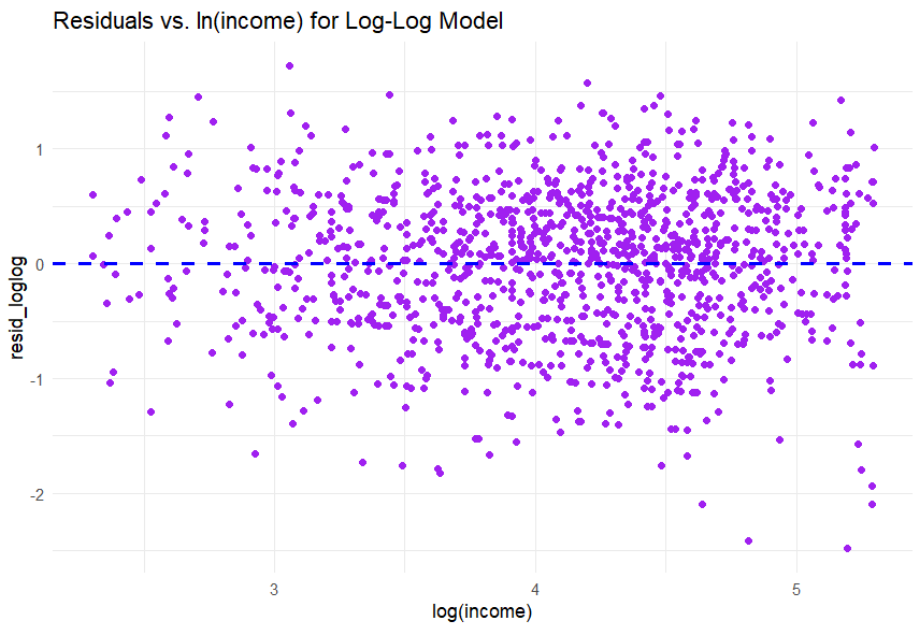
Point estimate for elasticity (γ2): 0.1863054

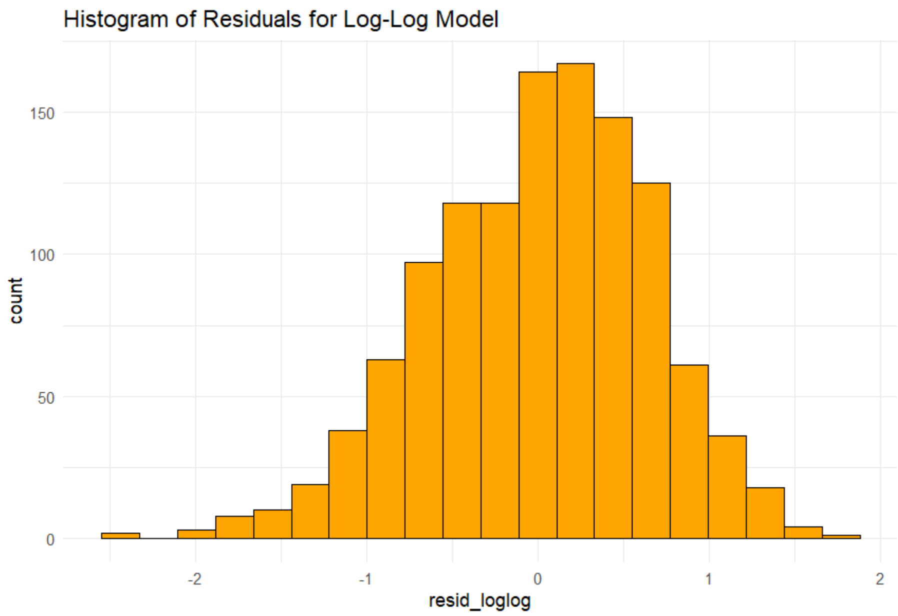
95% Confidence Interval for elasticity: 0.1293432 0.2432675

Interpretation:

由於 log-log 模型中彈性為常數，上述估計值直接代表 food 對 income 的彈性，可與 (d) 部分線性模型中依 income 水準變化的彈性進行比較。

(g)





==== (g) Jarque–Bera Test for Log-Log Residuals ====

data: resid_loglog

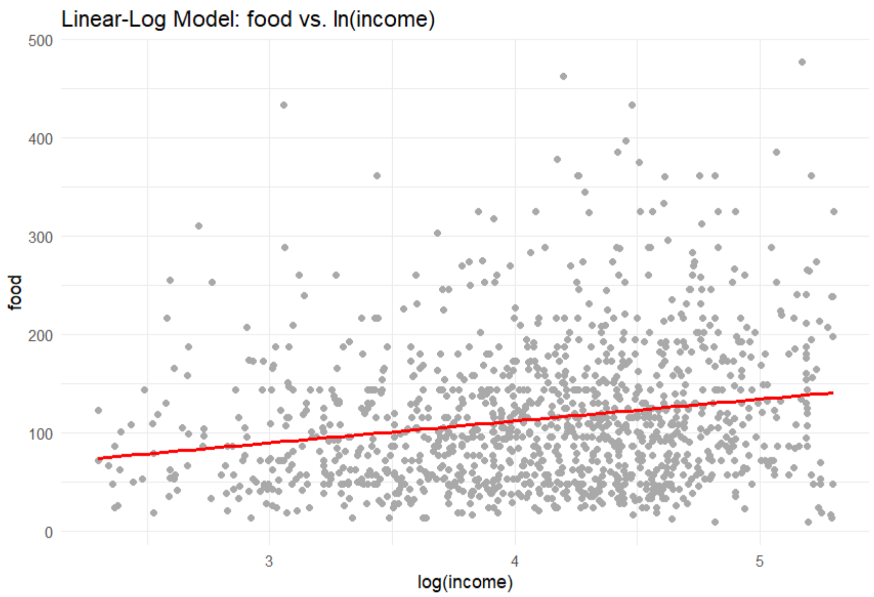
JB = 25.85, p-value = 2.436e-06

alternative hypothesis: greater

Interpretation:

若檢定結果的 p-value 遠小於 0.05，則可拒絕殘差呈常態分布的假設；但需注意，即使 p-value 顯著，也要觀察殘差散佈圖是否無系統性結構。

(h)



==== (h) Linear-Log Model Fit ====

R^2 for linear model (food ~ income): 0.0422812

R^2 for linear-log model (food ~ ln(income)): 0.03799984

Interpretation:

比較散佈圖及 R^2 值，若 linear-log 模型的資料點較緊密排列且 R^2 較高，則該模型對資料的解釋能力更佳。

(i)

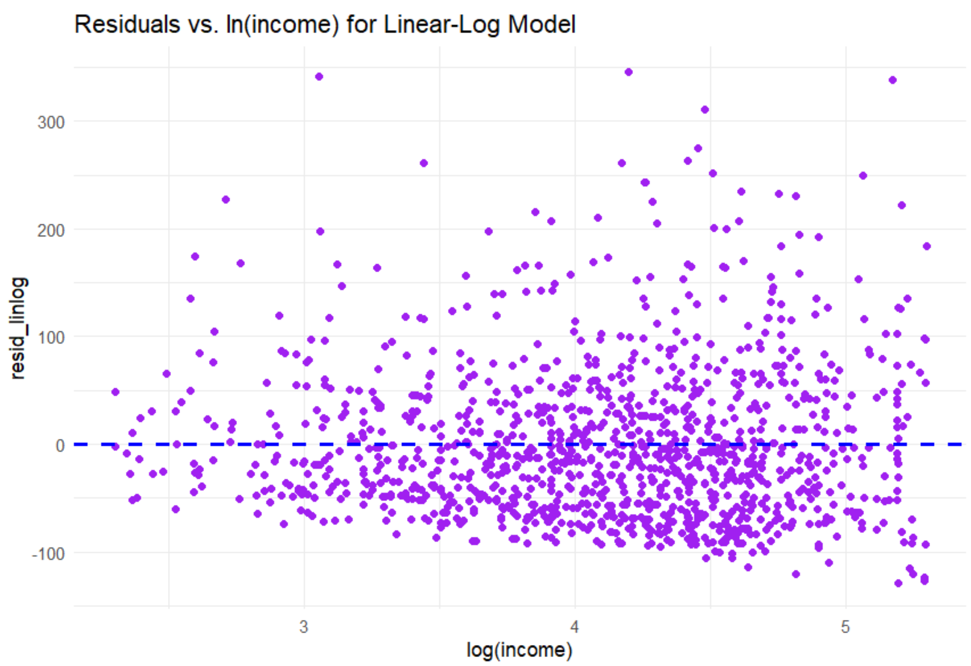
==== (i) Elasticity Estimates for Linear-Log Model ====

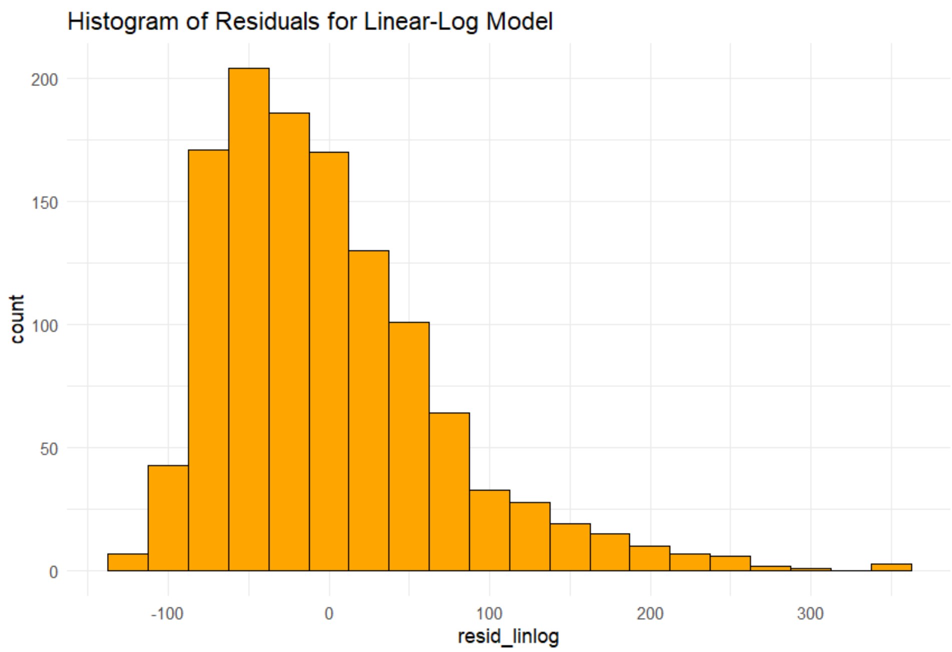
	income	elasticity	se	lower	upper
1	19	0.2495828	0.04706296	0.1573394	0.3418262
2	65	0.1909624	0.02755151	0.1369614	0.2449633
3	160	0.1629349	0.02005756	0.1236221	0.2022477

Interpretation:

由線性-log 模型下，各 income 水準下的 food 彈性是根據模型參數計算，可與 log-log 模型中常數彈性進行比較，檢視其是否在統計上顯著不同。

(j)





==== (j) Jarque–Bera Test for Linear-Log Residuals ====

data: resid_linlog

JB = 628.07, p-value < 2.2e-16

alternative hypothesis: greater

Interpretation:

檢定結果若顯著，則表示殘差偏離常態分布；否則可認為模型誤差近似常態。

(k)

從各模型的散佈圖、R^2 值、殘差診斷以及彈性估計來看：

- 線性模型 (food ~ income) 在原始尺度下解釋關係，但可能受到非線性與右偏的影響。

- log-log 模型 (ln(food) ~ ln(income)) 提供了恆定的彈性解釋，且散佈圖上的線性關係更明顯，同時計算出的 R^2（在轉換後尺度下）通常較高，殘差分布也較理想。

- 線性-log 模型 (food ~ ln(income)) 的關係解釋介於前兩者之間，其 R^2 與殘差行為可作為參考。

因此，根據資料擬合效果、彈性解釋的穩定性以及殘差診斷，我傾向選擇 log-log 模型作為對 food 與 income 關係的最佳描述。

原因包括：

1. log-log 模型直接提供常數彈性，易於解釋。
2. 散佈圖上 ln(food) 與 ln(income) 呈現較明顯的線性關係。
3. 模型 R^2 較高，且殘差分析顯示誤差結構較理想（儘管仍可能存在輕微偏離常態，但改善效果較好）。