

# HW0317

Yung-Jung Cheng

2025-03-23

## Q4

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:

$$\hat{RATING} = 64.289 + 0.990EXPER, \quad N = 50, \quad R^2 = 0.3793$$

Model 2:

$$\hat{RATING} = 39.464 + 15.312 \ln(EXPER), \quad N = 46, \quad R^2 = 0.6414$$

## Q4(a)

Sketch the fitted values from Model 1 for EXPER = 0 to 30 years.

## Ans

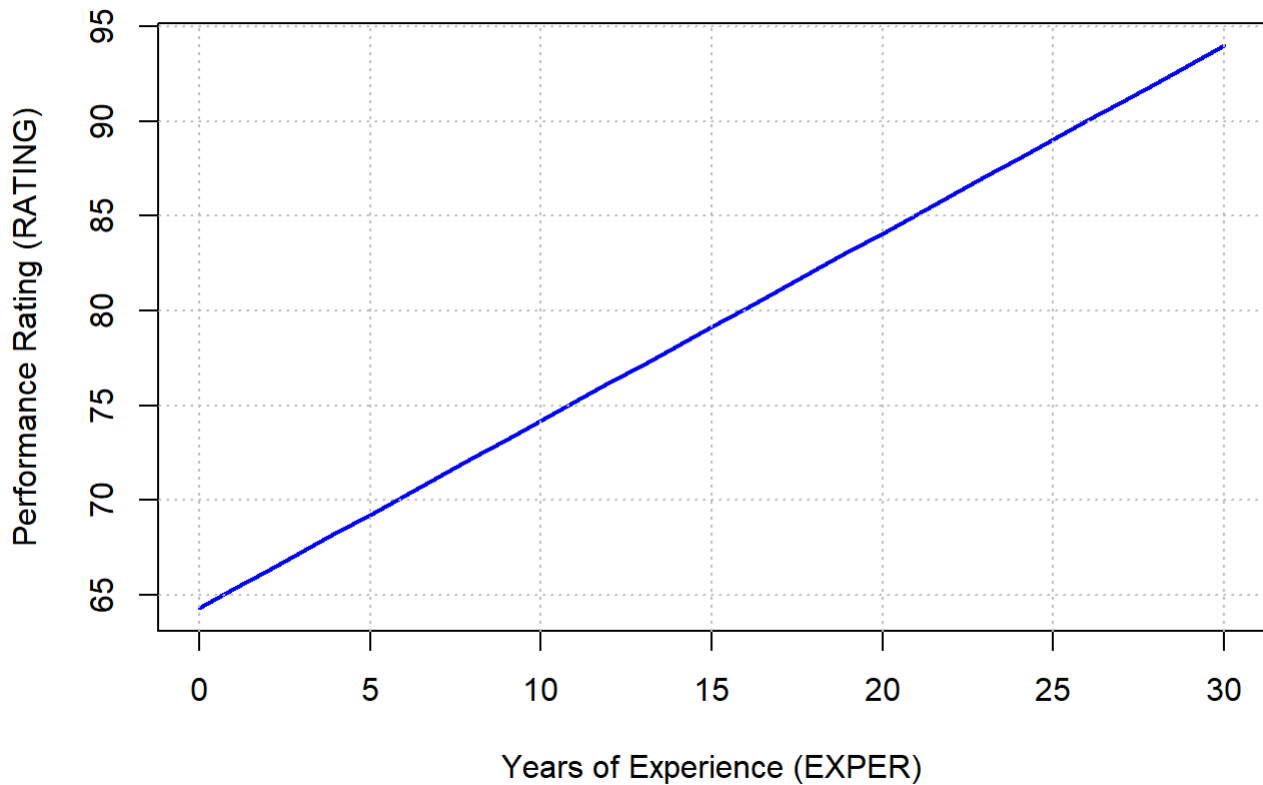
```
# 定義經驗年數範圍
exper <- 0:30

# 根據模型 1 計算預測的表現評分
rating <- 64.289 + 0.990 * exper

# 繪製圖形
plot(exper, rating, type = 'l', col = 'blue',
     main = "Model 1 Fitted Values",
     xlab = "Years of Experience (EXPER)",
     ylab = "Performance Rating (RATING)",
     lwd = 2)

# 添加網格
grid(nx = NULL, ny = NULL, col = "gray", lty = "dotted")
```

### Model 1 Fitted Values



## Q4(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.

## Ans

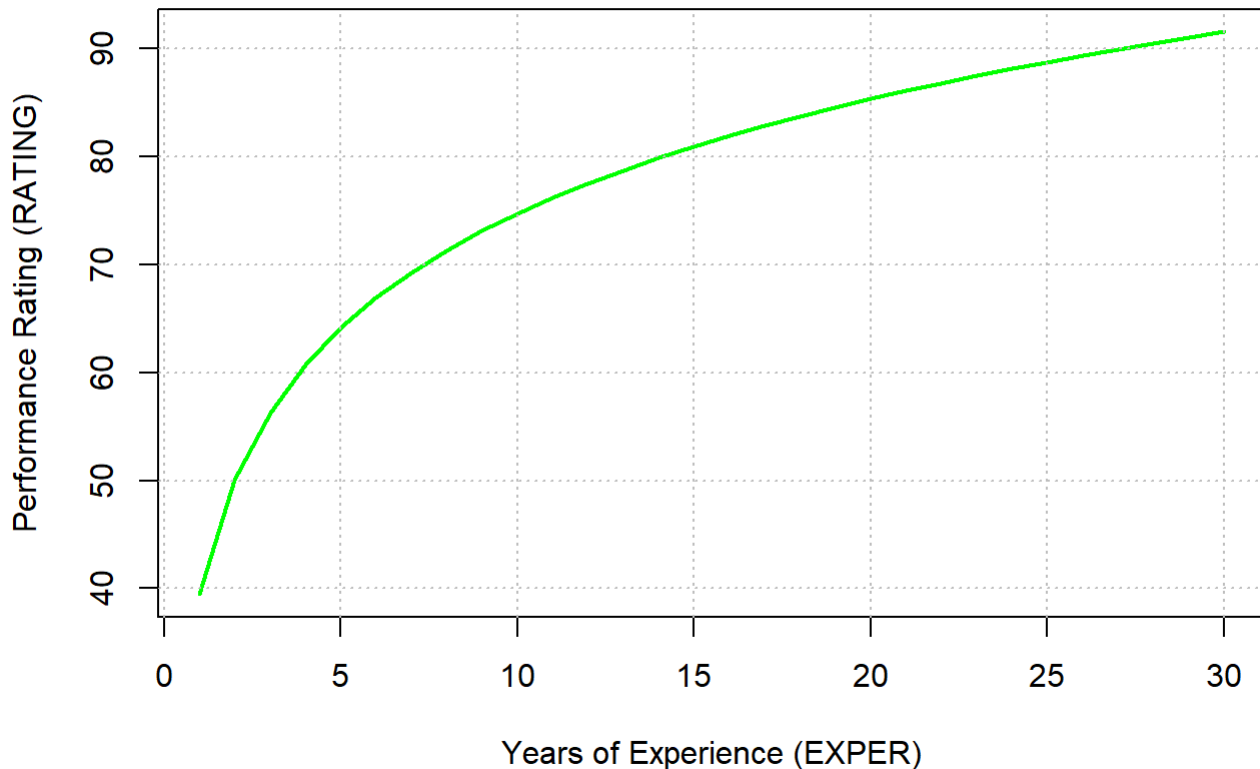
```
# 定義經驗年數範圍 (從 1 到 30)
exper <- 1:30

# 根據模型 2 計算預測的表現評分，使用自然對數
rating <- 39.464 + 15.312 * log(exper)

# 繪製圖形
plot(exper, rating, type = 'l', col = 'green',
     main = "Model 2 Fitted Values",
     xlab = "Years of Experience (EXPER)",
     ylab = "Performance Rating (RATING)",
     lwd = 2)

# 添加網格
grid(nx = NULL, ny = NULL, col = "gray", lty = "dotted")
```

## Model 2 Fitted Values



Since  $\ln(0)$  is undefined, we can not use model2 to predict those four artists with no experience.

### Q4(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.

### Ans

Marginal effect of model 1 is fitted and is 0.990. Also the model is linear, such that marginal effect on RATING of another year of experience for both 10 years of experience and 20 years of experience will be 0.990.

### Q4(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.

### Ans

$\beta = 15.312$

1. for a 10 years experience,  $15.312 \times \frac{1}{10} = 1.5312$
2. for a 20 years experience,  $15.312 \times \frac{1}{20} = 0.7656$

## Q4(e)

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

## Ans

Model 2 fits the data better, since  $0.6414 > 0.4858$

## Q4(f)

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

## Ans

**Model 2 is more reasonable** based on economic reasoning because it accounts for diminishing returns on experience. This model uses the logarithm of experience, reflecting a decrease in the impact of each additional year of experience as an artist becomes more skilled. This approach aligns better with real-world scenarios where improvements in performance tend to be greater at the beginning of a career and diminish over time.

## Q28

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: \$\$

$$\begin{aligned} YIELD_t &= \beta_0 + \beta_1 TIME + e_t \\ YIELD_t &= \alpha_0 + \alpha_1 \ln(TIME) + e_t \\ YIELD_t &= \gamma_0 + \gamma_1 TIME^2 + e_t \\ \ln(YIELD_t) &= \phi_0 + \phi_1 TIME + e_t \end{aligned}$$

\$\$

## Q28(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.

## Ans

```
# 數據準備
time_data <- wa_wheat$time
yield_data <- wa_wheat$northampton

# 模型擬合
model_linear <- lm(yield_data ~ time_data, data = wa_wheat)
model_log_time <- lm(yield_data ~ log(time_data), data = wa_wheat)
model_quadratic <- lm(yield_data ~ I(time_data^2), data = wa_wheat)
model_log_yield <- lm(log(yield_data) ~ time_data, data = wa_wheat)

# 檢視模型摘要 · 獲取R^2和殘差檢驗
summary(model_linear)
```

```
##
## Call:
## lm(formula = yield_data ~ time_data, data = wa_wheat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62394 -0.17302  0.03342  0.12996  0.72050
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.603245   0.081858   7.369 2.55e-09 ***
## time_data    0.023078   0.002908   7.935 3.69e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2791 on 46 degrees of freedom
## Multiple R-squared:  0.5778, Adjusted R-squared:  0.5687
## F-statistic: 62.96 on 1 and 46 DF,  p-value: 3.689e-10
```

```
summary(model_log_time)
```

```
##
## Call:
## lm(formula = yield_data ~ log(time_data), data = wa_wheat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78488 -0.20711 -0.06382  0.15447  0.91573
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.3510     0.1759   1.995  0.052 .
## log(time_data)  0.2790     0.0575   4.852 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3494 on 46 degrees of freedom
## Multiple R-squared:  0.3386, Adjusted R-squared:  0.3242
## F-statistic: 23.55 on 1 and 46 DF,  p-value: 1.44e-05
```

```
summary(model_quadratic)
```

```
##
## Call:
## lm(formula = yield_data ~ I(time_data^2), data = wa_wheat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56899 -0.14970  0.03119  0.12176  0.62049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.737e-01  5.222e-02   14.82  < 2e-16 ***
## I(time_data^2) 4.986e-04  4.939e-05   10.10 3.01e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2396 on 46 degrees of freedom
## Multiple R-squared:  0.689, Adjusted R-squared:  0.6822
## F-statistic: 101.9 on 1 and 46 DF, p-value: 3.008e-13
```

```
summary(model_log_yield)
```

```
##
## Call:
## lm(formula = log(yield_data) ~ time_data, data = wa_wheat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09292 -0.10049  0.07125  0.14140  0.40263
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.363938   0.076192  -4.777 1.85e-05 ***
## time_data    0.018632   0.002707   6.883 1.37e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2598 on 46 degrees of freedom
## Multiple R-squared:  0.5074, Adjusted R-squared:  0.4966
## F-statistic: 47.37 on 1 and 46 DF, p-value: 1.366e-08
```

```
# 正態性検測
shapiro.test(residuals(model_linear))
```

```
##
## Shapiro-Wilk normality test
##
## data:  residuals(model_linear)
## W = 0.98236, p-value = 0.6792
```

```
shapiro.test(residuals(model_log_time))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(model_log_time)  
## W = 0.96657, p-value = 0.1856
```

```
shapiro.test(residuals(model_quadratic))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(model_quadratic)  
## W = 0.98589, p-value = 0.8266
```

```
shapiro.test(residuals(model_log_yield))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(model_log_yield)  
## W = 0.86894, p-value = 7.205e-05
```

Based on the analysis, the **Quadratic Model (Model 3)** is the preferred choice. It has the highest  $R^2$  value of 0.689, indicating it explains the variability in the data better than the other models. The residuals are closest to a normal distribution, as confirmed by the Shapiro-Wilk test ( $p = 0.8266$ ), which supports the model's assumptions. Therefore, the quadratic model provides the best fit and most reliable statistical inferences for the data on wheat yield in Northampton.

## Q28(b)

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

## Ans

In the chosen Quadratic Model (Model 3), the coefficient of the time-squared variable ( $\beta_1 = 0.0004986$ ) represents the accelerating effect of time on wheat yield. Specifically, this coefficient indicates that as time progresses, the rate of increase in wheat yield accelerates. This suggests that wheat yield does not simply increase linearly over time but grows faster as time squares, reflecting potential improvements in technology, farming practices, or environmental factors.

## Q28(c)

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

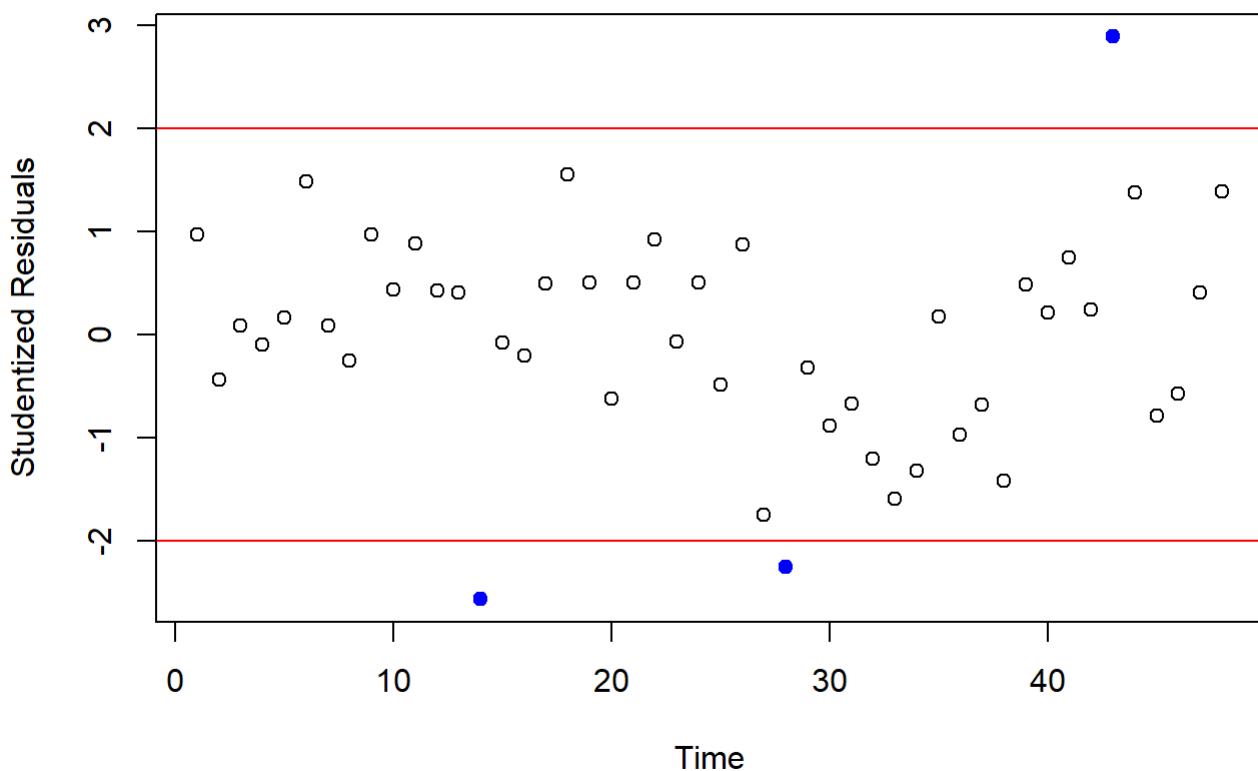
Ans

```
# 設定模型
model_quadratic <- lm(northampton ~ I(time^2), data = wa_wheat)

# 學生化殘差
student_resids <- rstudent(model_quadratic)
leverage_threshold <- 2 * mean(hatvalues(model_quadratic))
dffits_values <- dffits(model_quadratic)
dffits_threshold <- 2 * sqrt(2/nrow(wa_wheat))

# 繪製學生化殘差
plot(wa_wheat$time, student_resids, ylab="Studentized Residuals", xlab="Time", main="Plot of Studentized Residuals")
abline(h = c(-2, 2), col = "red")
# 標記異常值
outliers_resids <- which(abs(student_resids) > 2)
points(wa_wheat$time[outliers_resids], student_resids[outliers_resids], col = "blue", pch = 19)
```

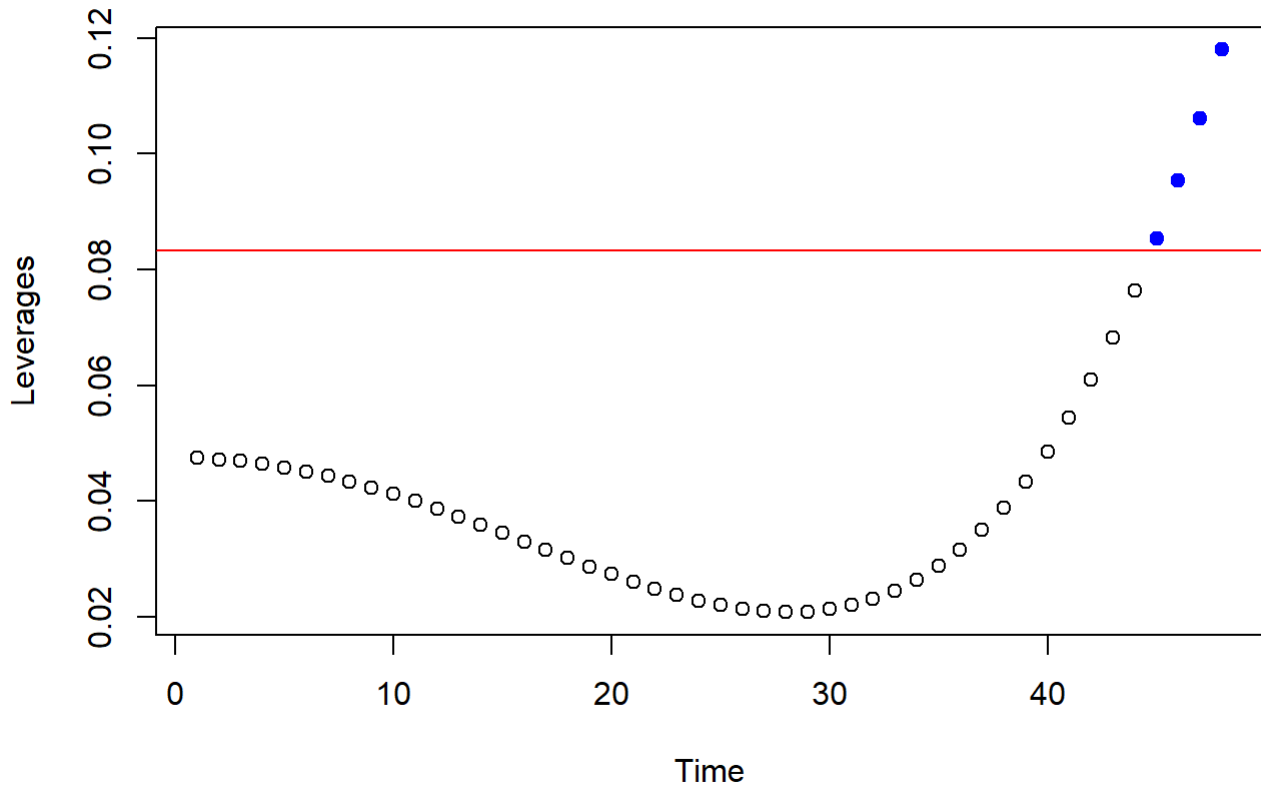
**Plot of Studentized Residuals**



```
# 繪製杠杆值
plot(wa_wheat$time, hatvalues(model_quadratic), ylab="Leverages", xlab="Time", main="Plot of Leverages")
abline(h = leverage_threshold, col = "red")
# 標記異常值
outliers_leverage <- which(hatvalues(model_quadratic) > leverage_threshold)
points(wa_wheat$time[outliers_leverage], hatvalues(model_quadratic)[outliers_leverage], col = "blue", pch = 19)
```

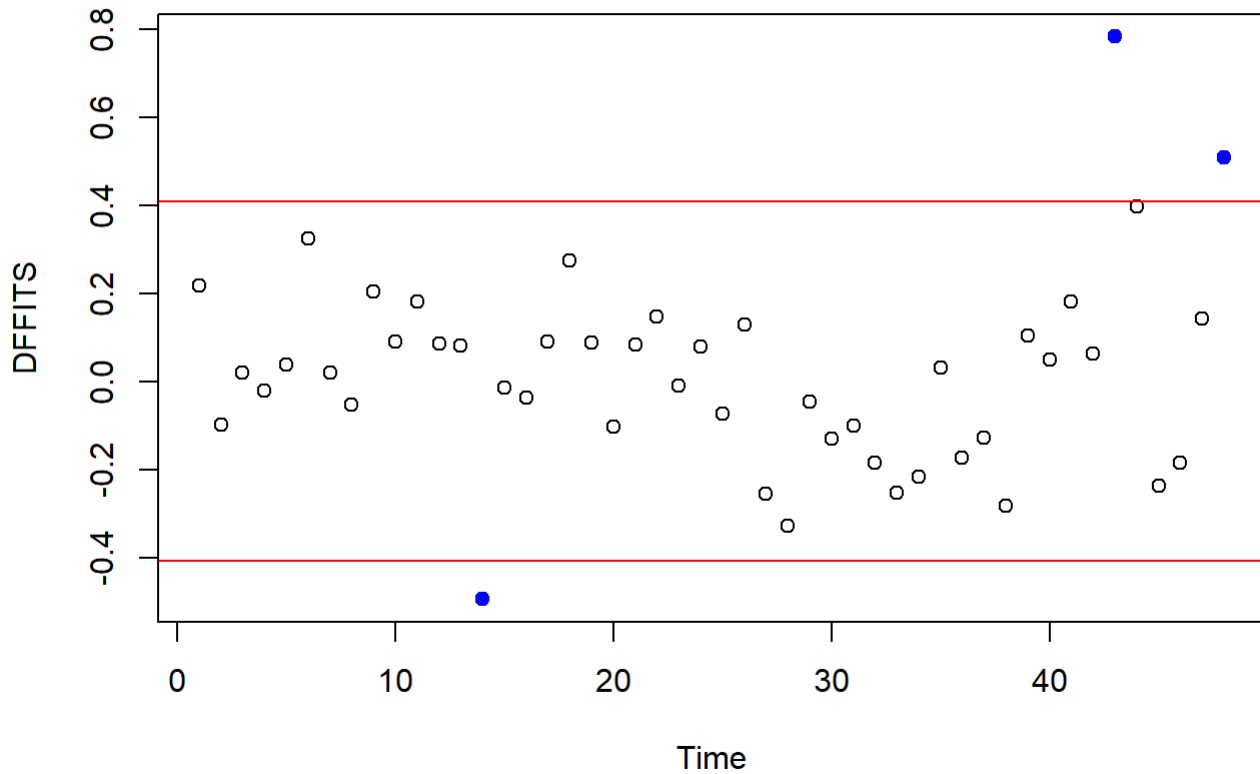


Plot of Leverages



```
# 繪製DFFITS
plot(wa_wheat$time, dffits_values, ylab="DFFITS", xlab="Time", main="DFFITS Values")
abline(h = c(-dffits_threshold, dffits_threshold), col = "red")
# 標記異常值
outliers_dffits <- which(abs(dffits_values) > dffits_threshold)
points(wa_wheat$time[outliers_dffits], dffits_values[outliers_dffits], col = "blue", pch = 19)
```

## DFFITS Values



```
# 打印出所有異常值的時間和北安普敦產量
wa_wheat[unique(c(outliers_resids, outliers_leverage, outliers_dffits)), ]
```

```
##      northampton chapman mullewa greenough time
## 14      0.3024   0.4167   0.3965    0.4369   14
## 28      0.6539   0.5827   0.4252    0.9759   28
## 43      2.3161   2.0244   1.6880    1.8081   43
## 45      1.6040   1.4769   1.3871    1.5674   45
## 46      1.6980   1.4430   1.4558    1.6893   46
## 47      1.9691   1.7107   1.6571    1.7191   47
## 48      2.2318   1.8435   1.7992    2.2353   48
```

## Q28(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?

# Ans

```
# 過濾1996年及以前的數據 ( 從1950年開始 · 對應到1996年為 time = 47 )
data_train <- wa_wheat[wa_wheat$time <= 47, ]

# 擬合二次模型
model_train <- lm(northampton ~ I(time^2), data = data_train)

# 進行1997年的預測 ( 對應到 time = 48 )
data_1997 <- wa_wheat[wa_wheat$time == 48, ]
predict_1997 <- predict(model_train, newdata = data_1997, interval = "prediction", level = 0.95)

# 輸出預測結果和95%預測區間
print(predict_1997)
```

```
##           fit           lwr           upr
## 48 1.881111 1.372403 2.389819
```

```
# 檢查1997年實際產量是否在預測區間內
actual_1997 <- data_1997$northampton
print(actual_1997)
```

```
## [1] 2.2318
```

```
within_interval <- actual_1997 >= predict_1997[1, "lwr"] & actual_1997 <= predict_1997[1, "upr"]
print(within_interval)
```

```
## [1] TRUE
```

```

# 建立預測資料框
df_pred <- data.frame(
  time = 48,
  fit = predict_1997[1, "fit"],
  lwr = predict_1997[1, "lwr"],
  upr = predict_1997[1, "upr"],
  actual = actual_1997
)

# 過去觀測資料 (1950-1996)
plot_data <- wa_wheat[wa_wheat$time <= 47, ]

# 繪圖
ggplot() +
  # 觀測資料點
  geom_point(data = plot_data, aes(x = time, y = northampton), color = "black") +

  # 擬合的二次曲線
  geom_smooth(data = plot_data, aes(x = time, y = northampton), method = "lm",
    formula = y ~ I(x^2), se = FALSE, color = "blue") +

  # 1997 預測點
  geom_point(data = df_pred, aes(x = time, y = fit), color = "blue", shape = 17, size = 3) +

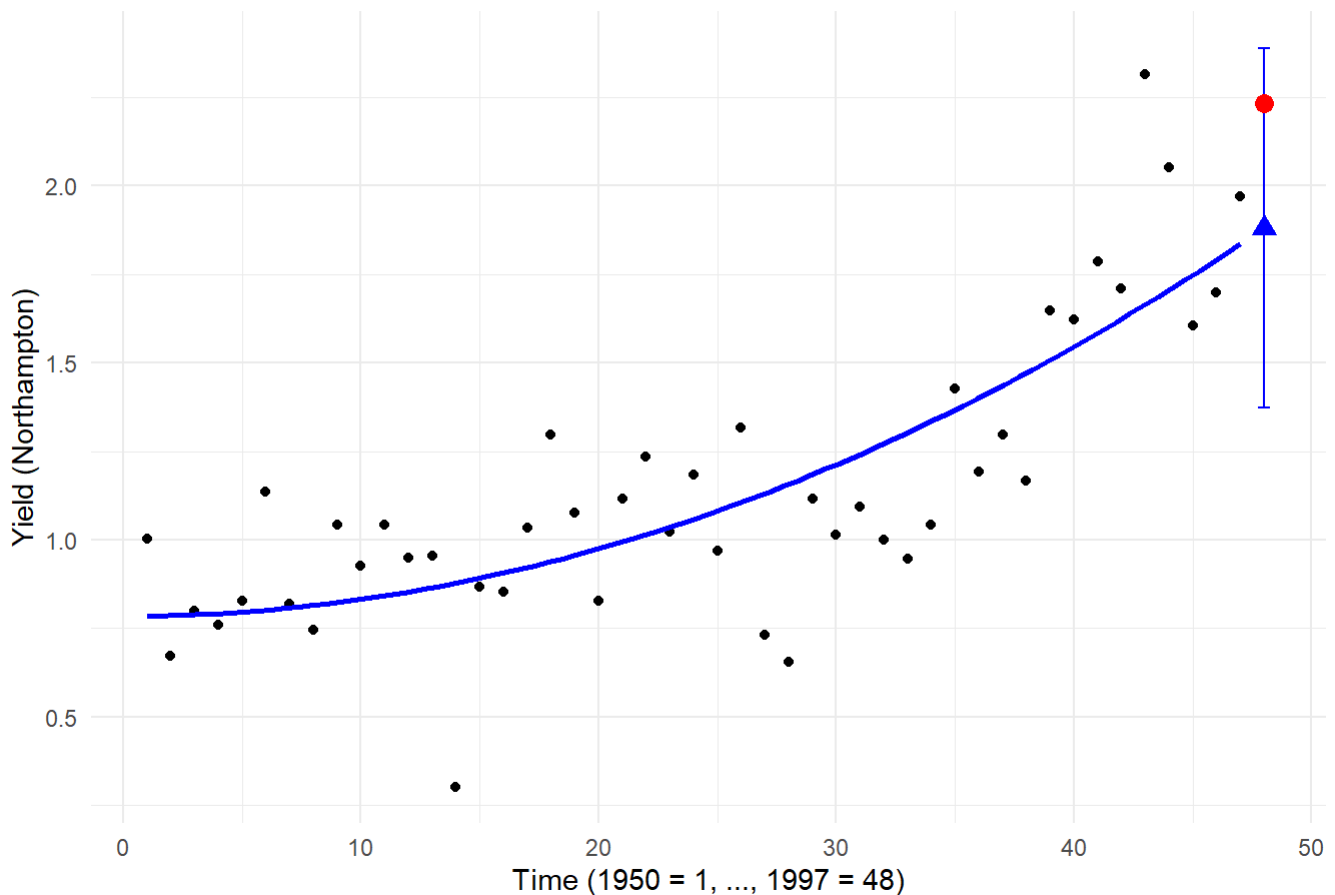
  # 95% 預測區間
  geom_errorbar(data = df_pred, aes(x = time, ymin = lwr, ymax = upr), width = 0.5, color =
"blue") +

  # 實際觀測值
  geom_point(data = df_pred, aes(x = time, y = actual), color = "red", shape = 19, size = 3)
+

  labs(title = "Prediction for 1997 with 95% Prediction Interval",
    x = "Time (1950 = 1, ..., 1997 = 48)",
    y = "Yield (Northampton)") +
  theme_minimal()

```

### Prediction for 1997 with 95% Prediction Interval



Yes, the interval contain true value

## Q29

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.

### Q29(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.

# Ans

```
library(ggplot2)
library(tseries) # Jarque-Bera test 用
```

```
# 抽取變數
```

```
food <- cex5_small$food
```

```
income <- cex5_small$income
```

```
# ===== 敘述統計 =====
```

```
food_stats <- c(
  mean = mean(food),
  median = median(food),
  min = min(food),
  max = max(food),
  sd = sd(food)
)
```

```
income_stats <- c(
  mean = mean(income),
  median = median(income),
  min = min(income),
  max = max(income),
  sd = sd(income)
)
```

```
print(round(food_stats, 4))
```

```
##      mean   median    min    max    sd
## 114.4431  99.8000   9.6300 476.6700 72.6575
```

```
print(round(income_stats, 4))
```

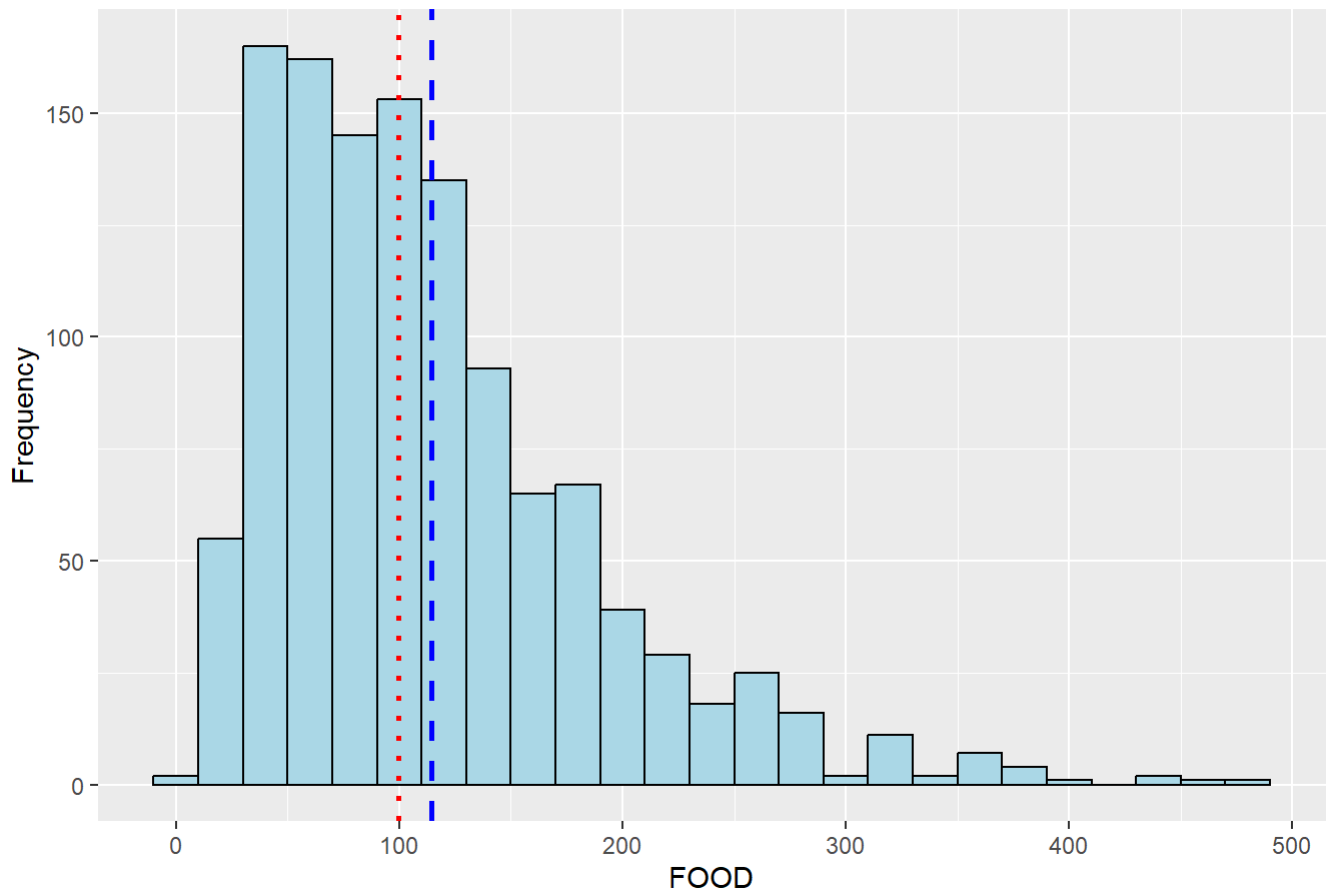
```
##      mean   median    min    max    sd
##  72.1426  65.2900  10.0000 200.0000 41.6523
```

```
# ===== 繪製直方圖 =====
```

```
# food histogram
```

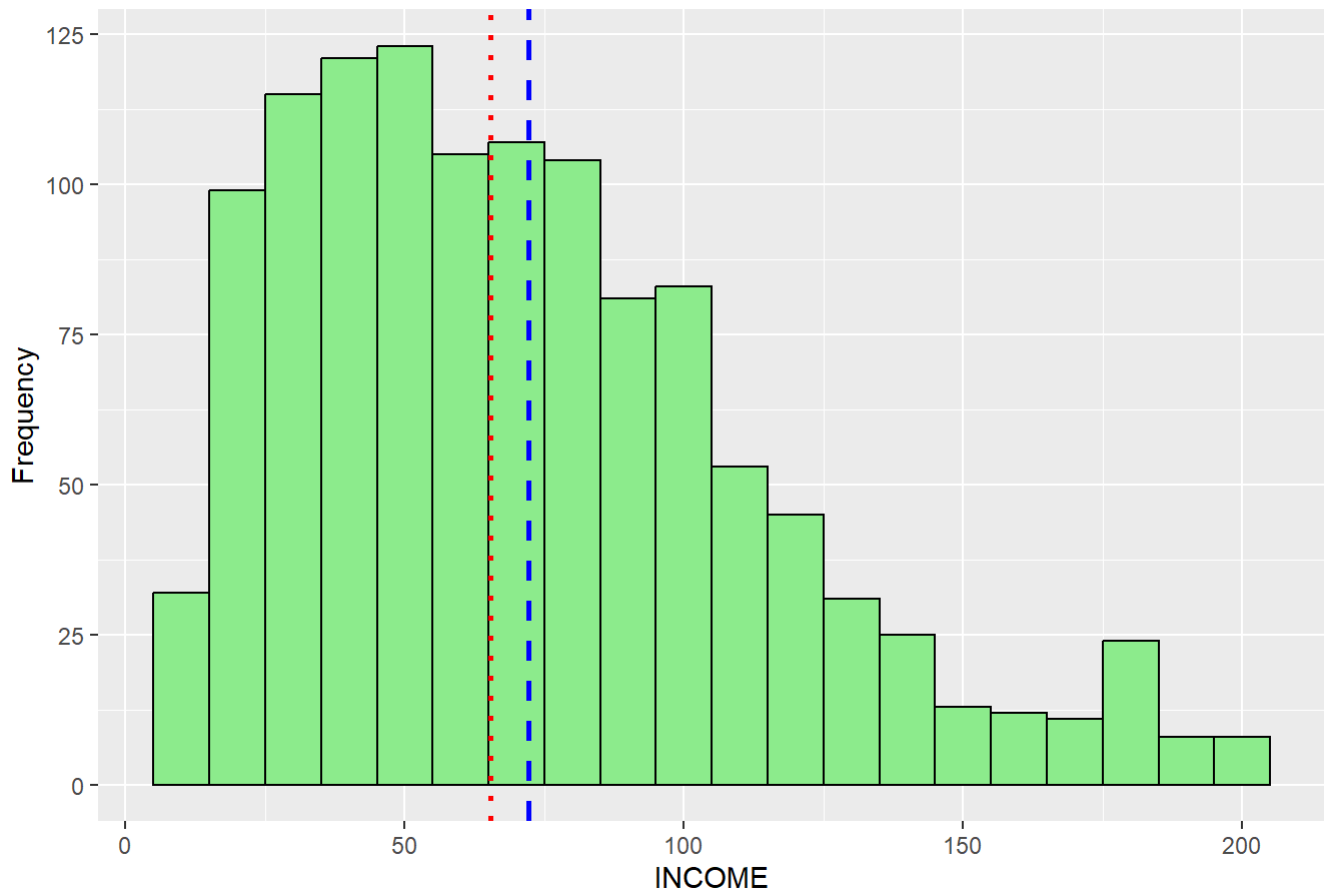
```
ggplot(cex5_small, aes(x = food)) +
  geom_histogram(binwidth = 20, fill = "lightblue", color = "black") +
  geom_vline(aes(xintercept = mean(food)), color = "blue", linetype = "dashed", linewidth =
1) +
  geom_vline(aes(xintercept = median(food)), color = "red", linetype = "dotted", linewidth =
1) +
  labs(title = "Histogram of FOOD", x = "FOOD", y = "Frequency")
```

Histogram of FOOD



```
# income histogram
ggplot(cex5_small, aes(x = income)) +
  geom_histogram(binwidth = 10, fill = "lightgreen", color = "black") +
  geom_vline(aes(xintercept = mean(income)), color = "blue", linetype = "dashed", linewidth =
1) +
  geom_vline(aes(xintercept = median(income)), color = "red", linetype = "dotted", linewidth
= 1) +
  labs(title = "Histogram of INCOME", x = "INCOME", y = "Frequency")
```

# Histogram of INCOME



```
# ===== Jarque-Bera 常態性検定 =====
jb_food <- jarque.bera.test(food)
jb_income <- jarque.bera.test(income)

print(jb_food)
```

```
##
## Jarque Bera Test
##
## data: food
## X-squared = 648.65, df = 2, p-value < 2.2e-16
```

```
print(jb_income)
```

```
##
## Jarque Bera Test
##
## data: income
## X-squared = 148.21, df = 2, p-value < 2.2e-16
```

## Summary Statistics

- **FOOD:**
  - Mean = 114.44
  - Median = 99.80
  - Minimum = 9.63
  - Maximum = 476.67



- Standard Deviation = 72.66
- **INCOME:**
  - Mean = 72.14
  - Median = 65.29
  - Minimum = 10.00
  - Maximum = 200.00
  - Standard Deviation = 41.65

---

## Histograms

- Both histograms are **right-skewed**, not symmetrical or bell-shaped.
- In both cases, the **mean is greater than the median**.

---

## Jarque–Bera Normality Test

- **FOOD:**
  - $JB = 648.65, p < 2.2e-16 \rightarrow$  **Not normally distributed**
- **INCOME:**
  - $JB = 148.21, p < 2.2e-16 \rightarrow$  **Not normally distributed**

## Q29(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  $\beta_2$ . Have we estimated the effect of changing income on average FOOD relatively precisely, or not?

## Ans

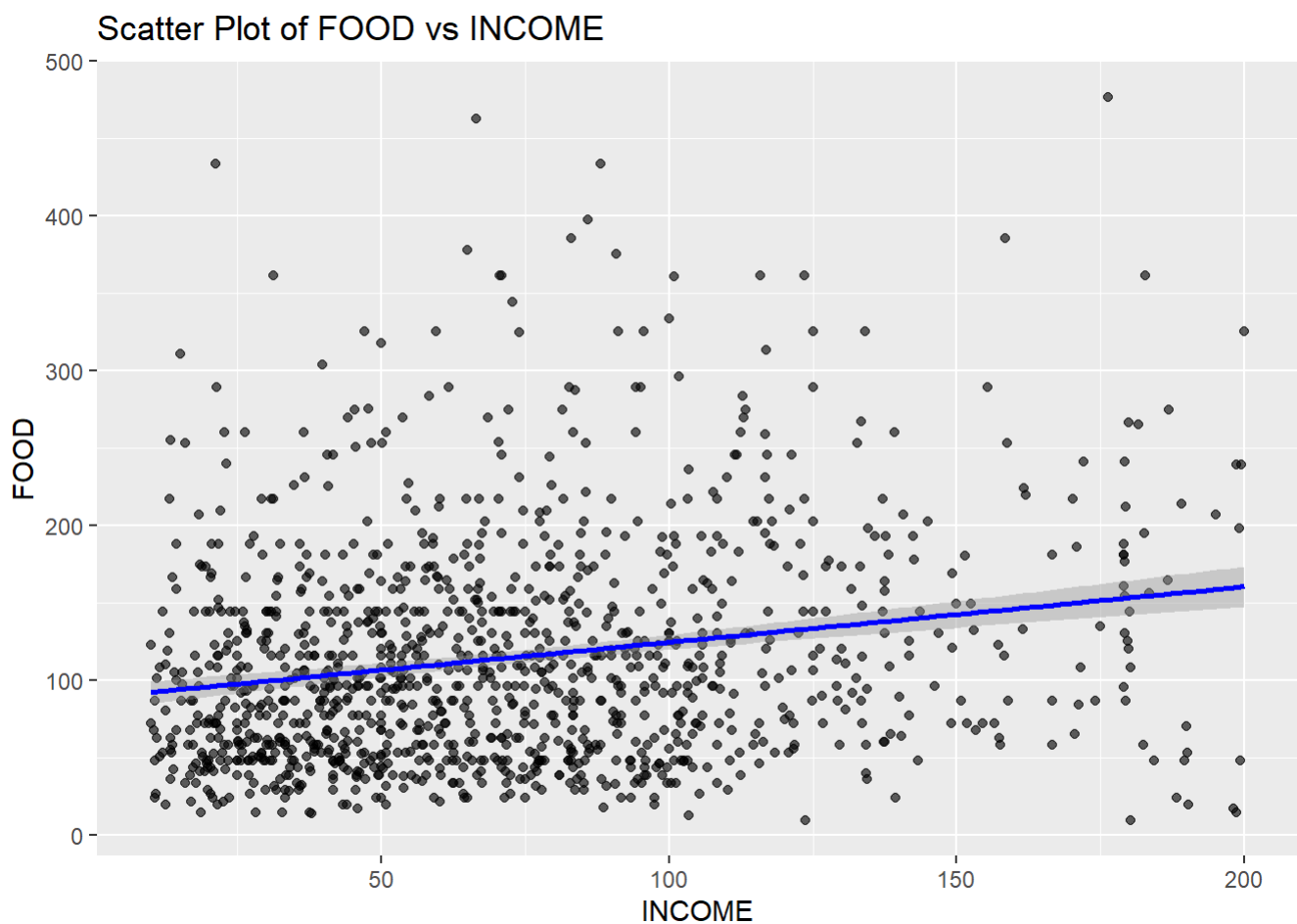
```
# 建立線性模型
model_lm <- lm(food ~ income, data = cex5_small)

# 顯示模型摘要 (包含估計值、t 值與 p 值)
summary(model_lm)
```

```
##
## 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
```

```
# 畫出散佈圖與回歸線
ggplot(cex5_small, aes(x = income, y = food)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Scatter Plot of FOOD vs INCOME",
       x = "INCOME", y = "FOOD")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
# 建立 62 的 95% 信賴區間
confint(model_lm, level = 0.95)
```

```
##                2.5 %    97.5 %
## (Intercept) 80.5064570 96.626543
## income      0.2619215  0.455452
```

We estimated the model:

$$FOOD = \beta_1 + \beta_2 \cdot INCOME + e$$

Estimated regression equation:

$$\widehat{FOOD} = 88.5665 + 0.3587 \cdot INCOME$$

95% Confidence Interval for  $\beta_2$ :

- [0.2619, 0.4555]

Interpretation:

- On average, for each additional unit of income, food expenditure increases by **approximately 0.36 units**.
- Since the confidence interval does **not include 0** and the p-value is very small ( $p < 0.0001$ ), income has a **statistically significant** positive effect on food spending.
- The narrow confidence interval indicates a **precise estimate** of the slope.

Scatter plot observation:

- The confidence band around the regression line shows the uncertainty in estimating the **mean** response.
- It is expected that many actual data points fall outside this band, since it is **not a prediction interval**.
- Therefore, the presence of many points outside the band does **not imply a poor fit**.

## Q29 (c)

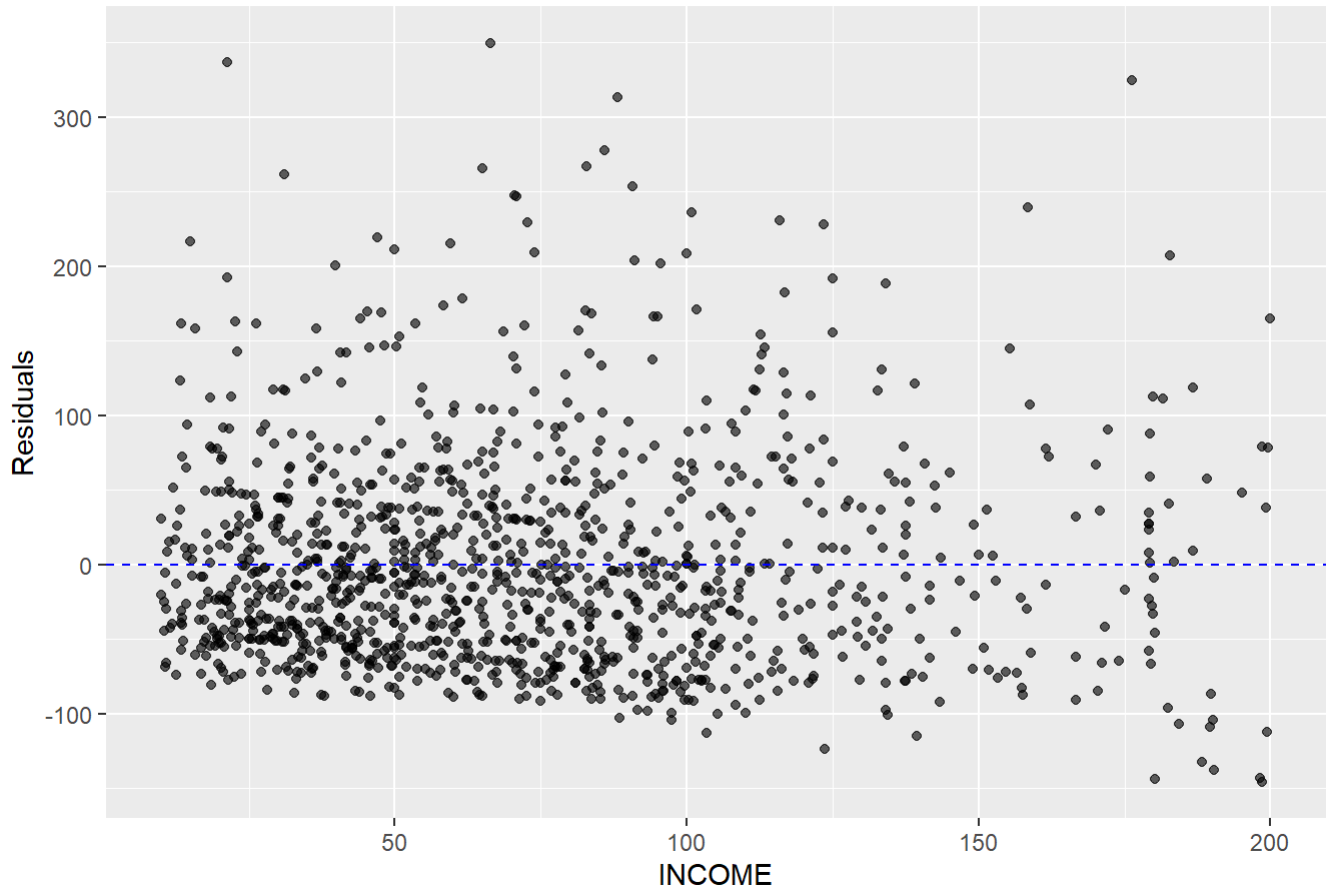
Obtain the least squares residuals from the regression in (b) and plot them against INCOME. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. Is it more important for the variables FOOD and INCOME to be normally distributed, or that the random error  $e$  be normally distributed? Explain your reasoning.

## Ans

```
# 提取殘差
residuals_lm <- residuals(model_lm)

# 殘差 vs INCOME
ggplot(data = cex5_small, aes(x = income, y = residuals_lm)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "blue", linetype = "dashed") +
  labs(title = "Residuals vs INCOME", x = "INCOME", y = "Residuals")
```

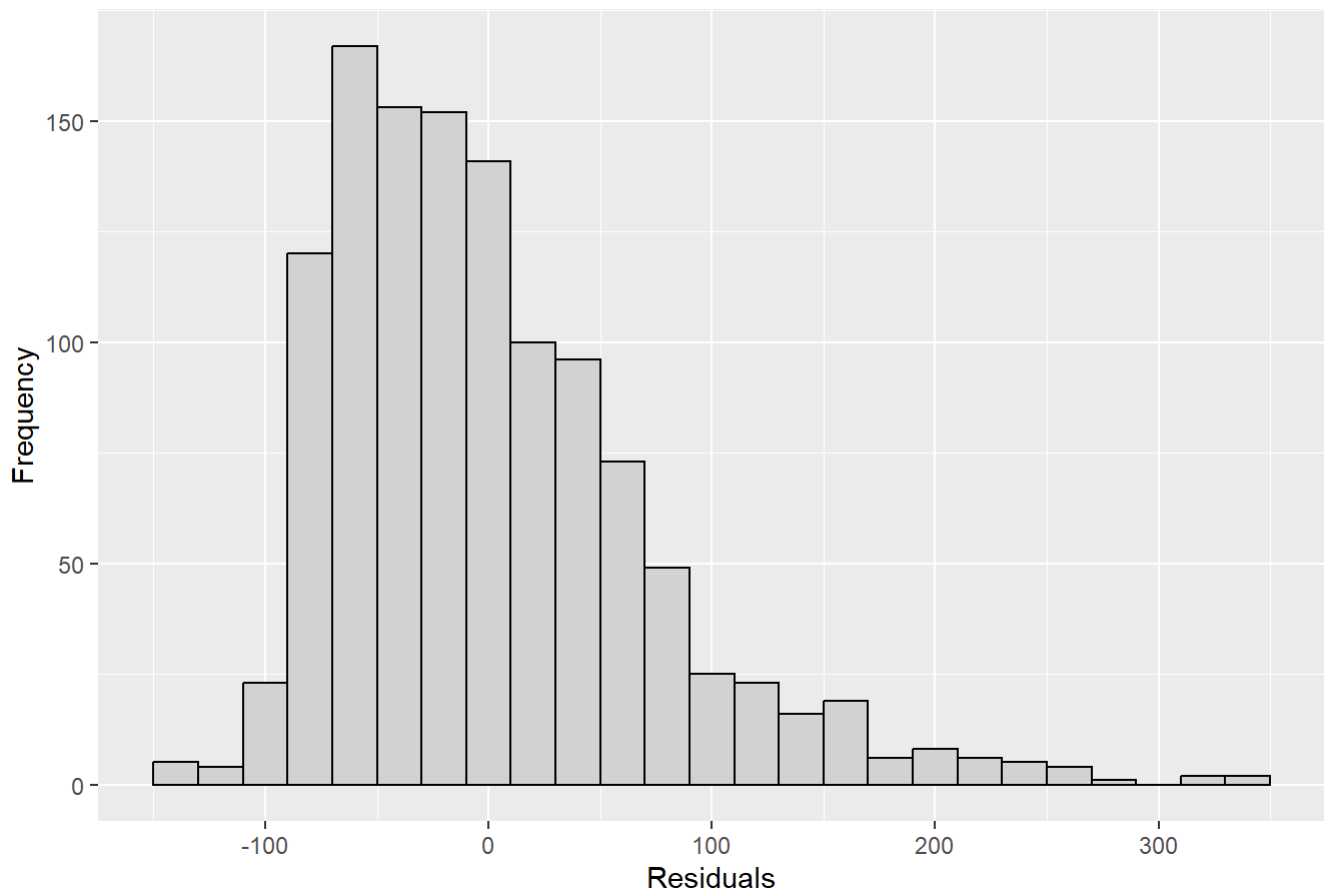
Residuals vs INCOME



```
# 殘差直方圖
```

```
ggplot(data.frame(residuals = residuals_lm), aes(x = residuals)) +  
  geom_histogram(binwidth = 20, fill = "lightgray", color = "black") +  
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency")
```

### Histogram of Residuals



```
# Jarque-Bera test for residuals
jarque.bera.test(residuals_lm)
```

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

```
# 執行 Breusch-Pagan test
bptest(model_lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_lm
## BP = 14.086, df = 1, p-value = 0.0001746
```

#### Breusch-Pagan Test for Heteroskedasticity:

- BP statistic = 14.086
- Degrees of freedom = 1
- p-value = 0.00017

#### Conclusion:

- We reject the null hypothesis of constant variance.
- There is statistical evidence of **heteroskedasticity** in the residuals.

- This supports the earlier visual observation that residual spread increases with income.
- As a result, standard errors from the OLS regression may be **biased**, and inference should be interpreted with caution or corrected using robust methods.

## Q29(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?

## Ans

```
# 目標點
income_points <- c(19, 65, 160)

# 預測 food (回歸線上的點，不是隨機)
food_hat <- predict(model_lm, newdata = data.frame(income = income_points))

# 斜率估計值與標準誤
b2 <- coef(model_lm)["income"]
se_b2 <- summary(model_lm)$coefficients["income", "Std. Error"]

# 建立彈性點估計
elasticity_point <- b2 * income_points / food_hat

# delta method approximation for SE of elasticity
# Var(E) ≈ (INCOME / FOOD)^2 * Var(b2)
elasticity_se <- (income_points / food_hat)^2 * se_b2^2
elasticity_ci_lower <- elasticity_point - 1.96 * sqrt(elasticity_se)
elasticity_ci_upper <- elasticity_point + 1.96 * sqrt(elasticity_se)

# 結果整理
elasticity_table <- data.frame(
  INCOME = income_points,
  FOOD_hat = round(food_hat, 4),
  Elasticity = round(elasticity_point, 4),
  Lower_95CI = round(elasticity_ci_lower, 4),
  Upper_95CI = round(elasticity_ci_upper, 4)
)

print(elasticity_table)
```

##	INCOME	FOOD_hat	Elasticity	Lower_95CI	Upper_95CI
## 1	19	95.3815	0.0715	0.0522	0.0907
## 2	65	111.8811	0.2084	0.1522	0.2645
## 3	160	145.9564	0.3932	0.2872	0.4992

We computed the point and 95% confidence interval estimates of the elasticity of food expenditure at different income levels, based on the linear model:

$$\text{Elasticity} = \hat{\beta}_2 \cdot \frac{INCOME}{\widehat{FOOD}}$$

INCOME	Predicted FOOD	Elasticity	95% CI
19	95.3815	0.0715	[0.0522, 0.0907]
65	111.8811	0.2084	[0.1522, 0.2645]
160	145.9564	0.3932	[0.2872, 0.4992]

#### Observations:

- Elasticity increases with income, indicating greater responsiveness of food expenditure at higher income levels.
- The confidence intervals do not significantly overlap, suggesting the differences are statistically meaningful.
- All elasticity values are below 1, consistent with food being a **necessity good** in economic theory.

## Q29(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<sup>2</sup> for the log-log model and compare it to the R<sup>2</sup> from the linear model. Which of the models seems to fit the data better?

## Ans

```
# 取 Log 變數
cex5_small$log_food <- log(cex5_small$food)
cex5_small$log_income <- log(cex5_small$income)

# 建立 Log-Log 模型
model_loglog <- lm(log_food ~ log_income, data = cex5_small)

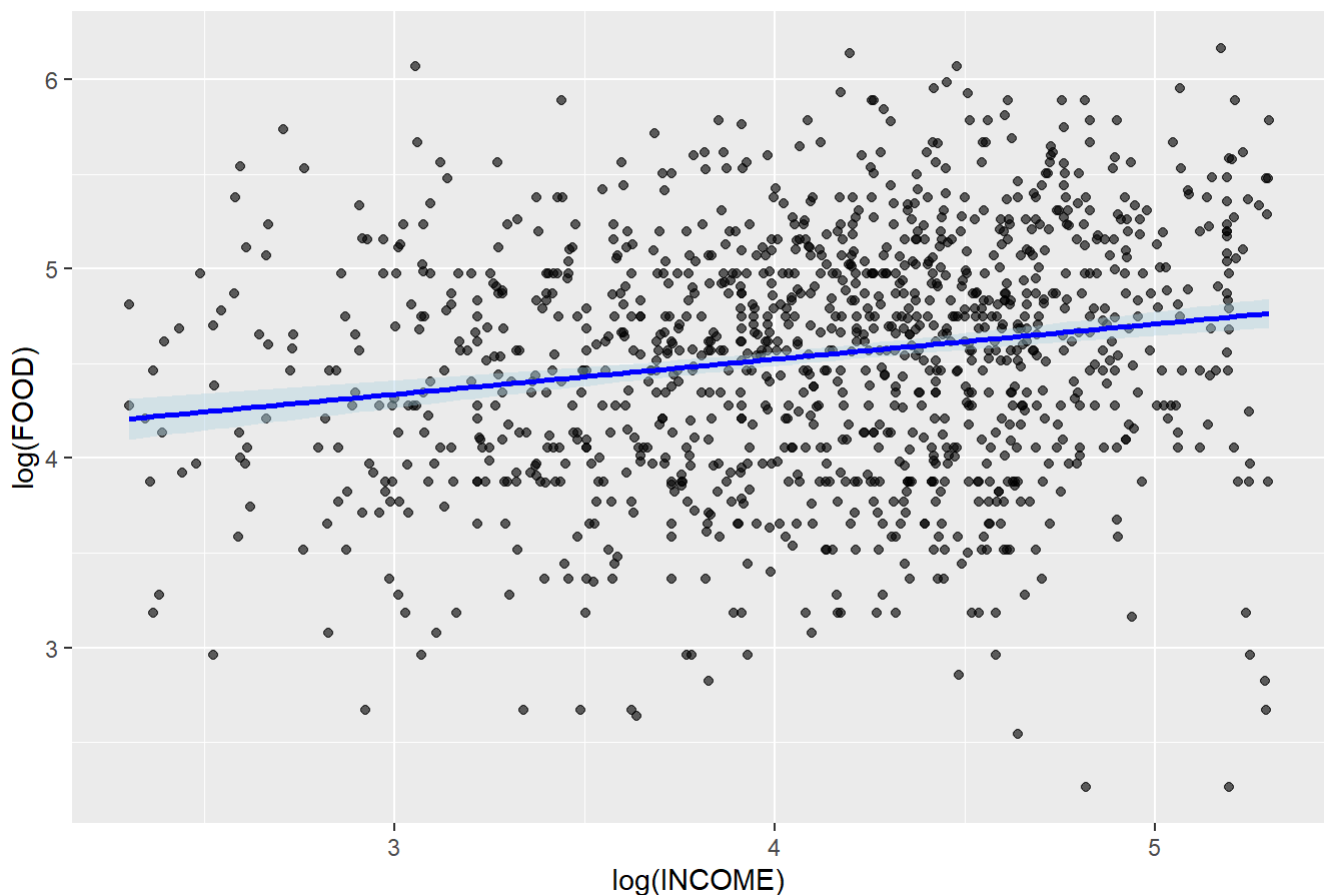
# 顯示模型摘要
summary(model_loglog)
```

```
##
## Call:
## lm(formula = log_food ~ log_income, data = cex5_small)
##
## 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
```

```
# 畫圖: Ln(FOOD) vs Ln(INCOME)
ggplot(cex5_small, aes(x = log_income, y = log_food)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue", fill = "lightblue") +
  labs(title = "Scatter Plot of log(FOOD) vs log(INCOME)",
       x = "log(INCOME)", y = "log(FOOD)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Scatter Plot of log(FOOD) vs log(INCOME)





We estimated the model:

$$\ln(FOOD) = \gamma_1 + \gamma_2 \cdot \ln(INCOME) + e$$

- Estimated elasticity ( $\gamma_2$ ): **0.1863**
- $R^2 = \mathbf{0.0332}$ , slightly lower than the linear model  $R^2 = \mathbf{0.0423}$

#### Visual comparison:

- The log-log scatter plot shows a clearer linear trend.
- Points are more symmetrically distributed around the line.

#### Conclusion:

- The log-log model gives a more well-defined visual relationship.
- However, it explains slightly **less variation** in the data than the linear model.

## Q29(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.

## Ans

```
# 提取彈性與標準誤
gamma2 <- coef(model_loglog)["log_income"]
se_gamma2 <- summary(model_loglog)$coefficients["log_income", "Std. Error"]

# 建立 95% 信賴區間
gamma2_ci <- c(
  lower = gamma2 - 1.96 * se_gamma2,
  upper = gamma2 + 1.96 * se_gamma2
)

# 顯示結果
round(c(Elasticity = gamma2, gamma2_ci), 4)
```

```
## Elasticity.log_income      lower.log_income      upper.log_income
##                0.1863                0.1294                0.2432
```

- Estimated elasticity ( $\gamma_2$ ) = **0.1863**
- 95% confidence interval: **[0.1294, 0.2432]**

#### Comparison with part (d):

INCOME	Elasticity (d)	95% CI (d)	Inside log-log CI?
19	0.0715	[0.0522, 0.0907]	✗ No
65	0.2084	[0.1522, 0.2645]	✓ Yes
160	0.3932	[0.2872, 0.4992]	✗ No

#### Conclusion:

- The elasticity from the log-log model is close to that at income = 65 in part (d), but differs from those at income = 19 and 160.
- This shows that the log-log model assumes a constant elasticity, while the linear model reveals that elasticity **increases with income**, which may better reflect economic behavior.

## Q29(g)

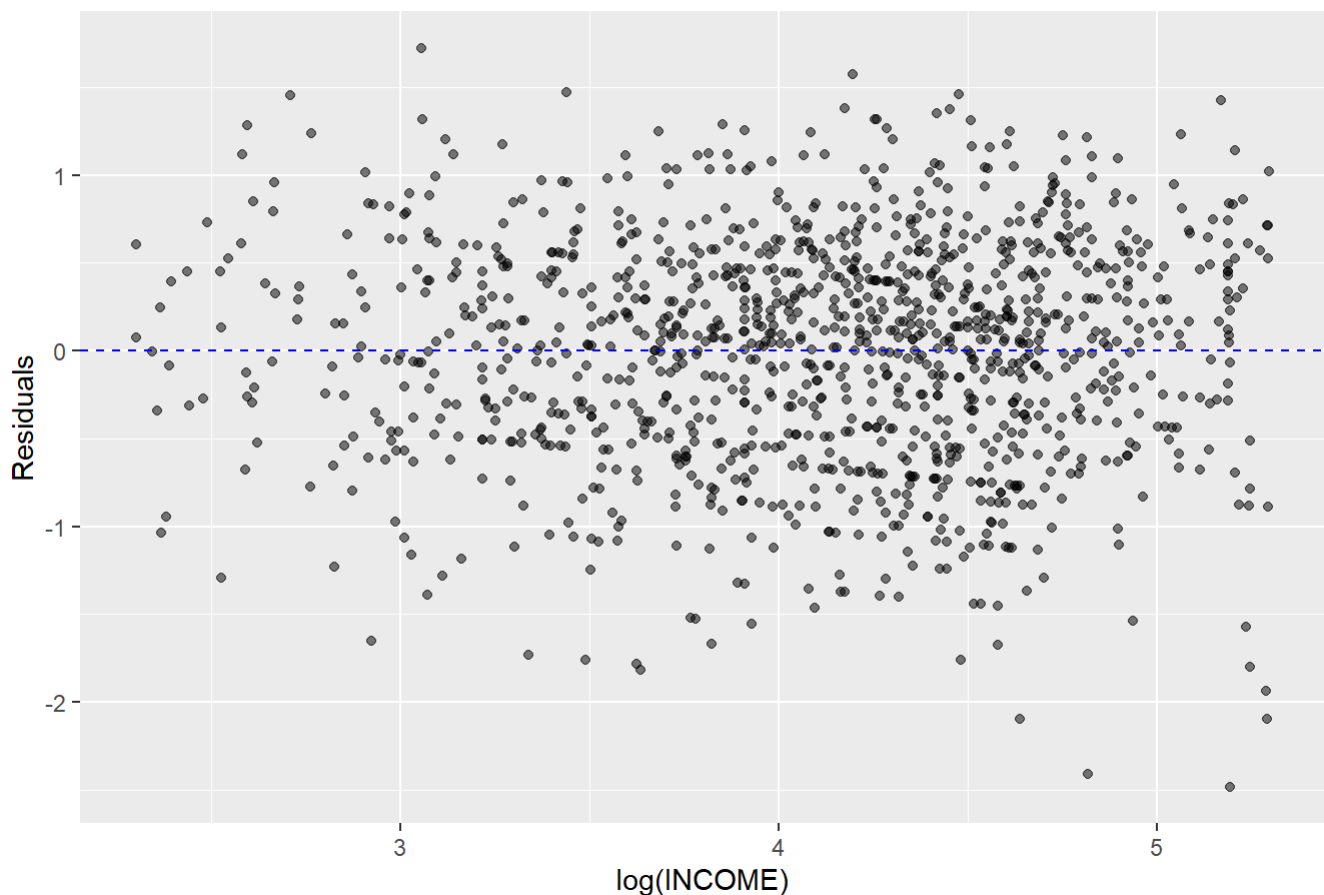
Obtain the least squares residuals from the log-log model and plot them against  $\ln(\text{INCOME})$ . Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

## Ans

```
# 取出 Log-Log 模型的殘差
residuals_loglog <- residuals(model_loglog)

# 殘差 vs ln(INCOME)
ggplot(cex5_small, aes(x = log_income, y = residuals_loglog)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "blue", linetype = "dashed") +
  labs(title = "Residuals vs log(INCOME)",
       x = "log(INCOME)", y = "Residuals")
```

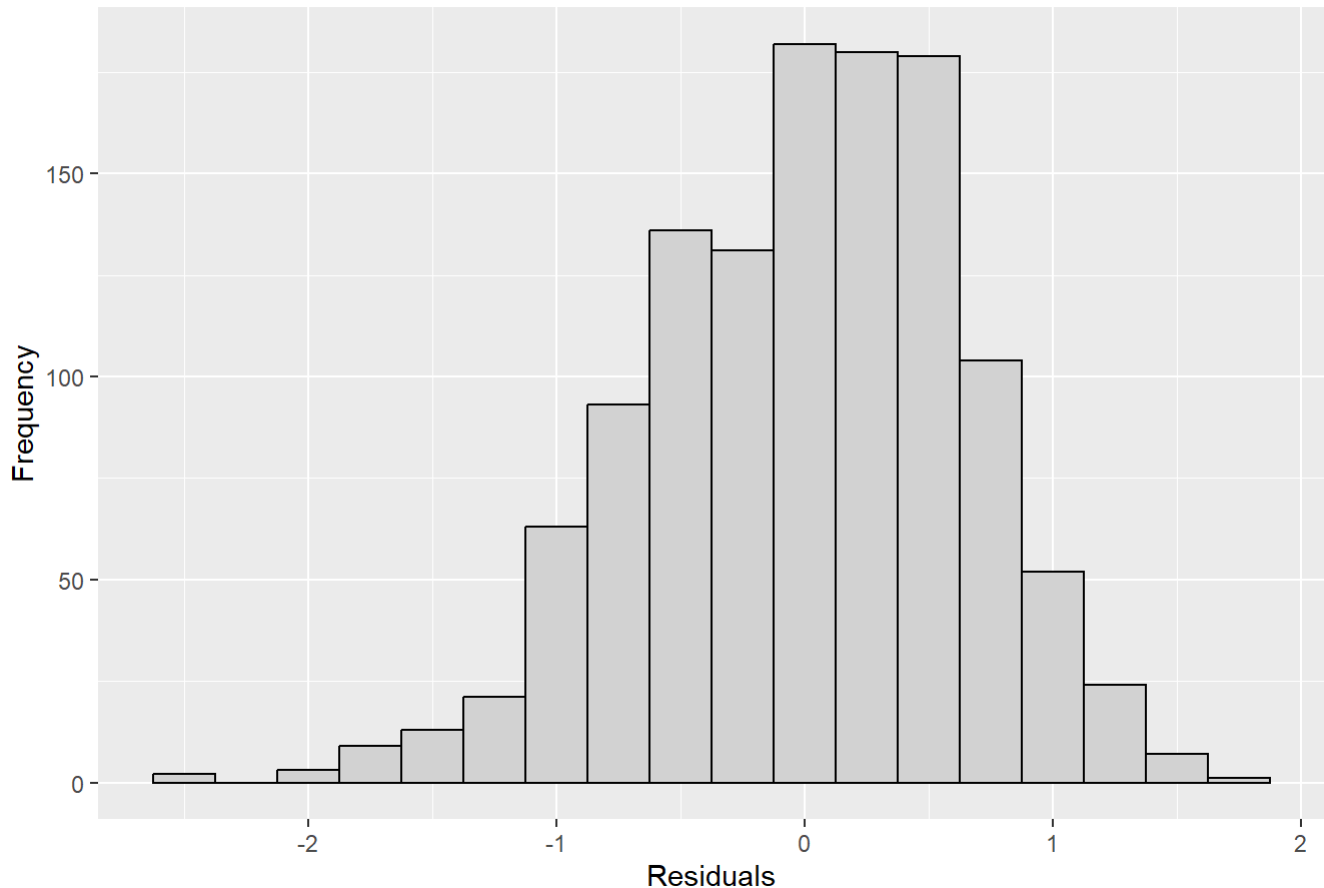
Residuals vs log(INCOME)



```
# 殘差直方圖
```

```
ggplot(data.frame(residuals = residuals_loglog), aes(x = residuals)) +  
  geom_histogram(binwidth = 0.25, fill = "lightgray", color = "black") +  
  labs(title = "Histogram of Residuals (log-log model)",  
        x = "Residuals", y = "Frequency")
```

Histogram of Residuals (log-log model)



```
# Jarque-Bera 檢定
```

```
jarque.bera.test(residuals_loglog)
```

```
##
```

```
## Jarque Bera Test
```

```
##
```

```
## data: residuals_loglog
```

```
## X-squared = 25.85, df = 2, p-value = 2.436e-06
```

### Residuals vs log(INCOME):

- The residuals are randomly scattered around zero, with no clear pattern.
- No evidence of heteroskedasticity or model misfit.

### Histogram of residuals:

- Roughly bell-shaped, slightly left-skewed.
- More symmetric than in the linear model.

### Jarque-Bera test:

- JB = 25.85, p-value = 2.44e-06 → Reject normality.
- However, this is a **major improvement** compared to the linear model (JB = 624).

## Conclusion:

- The residuals from the log-log model are **not perfectly normal**, but much closer to normality.
- The log-log model shows a better error structure and improves model assumptions.

## Q29(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?

## Ans

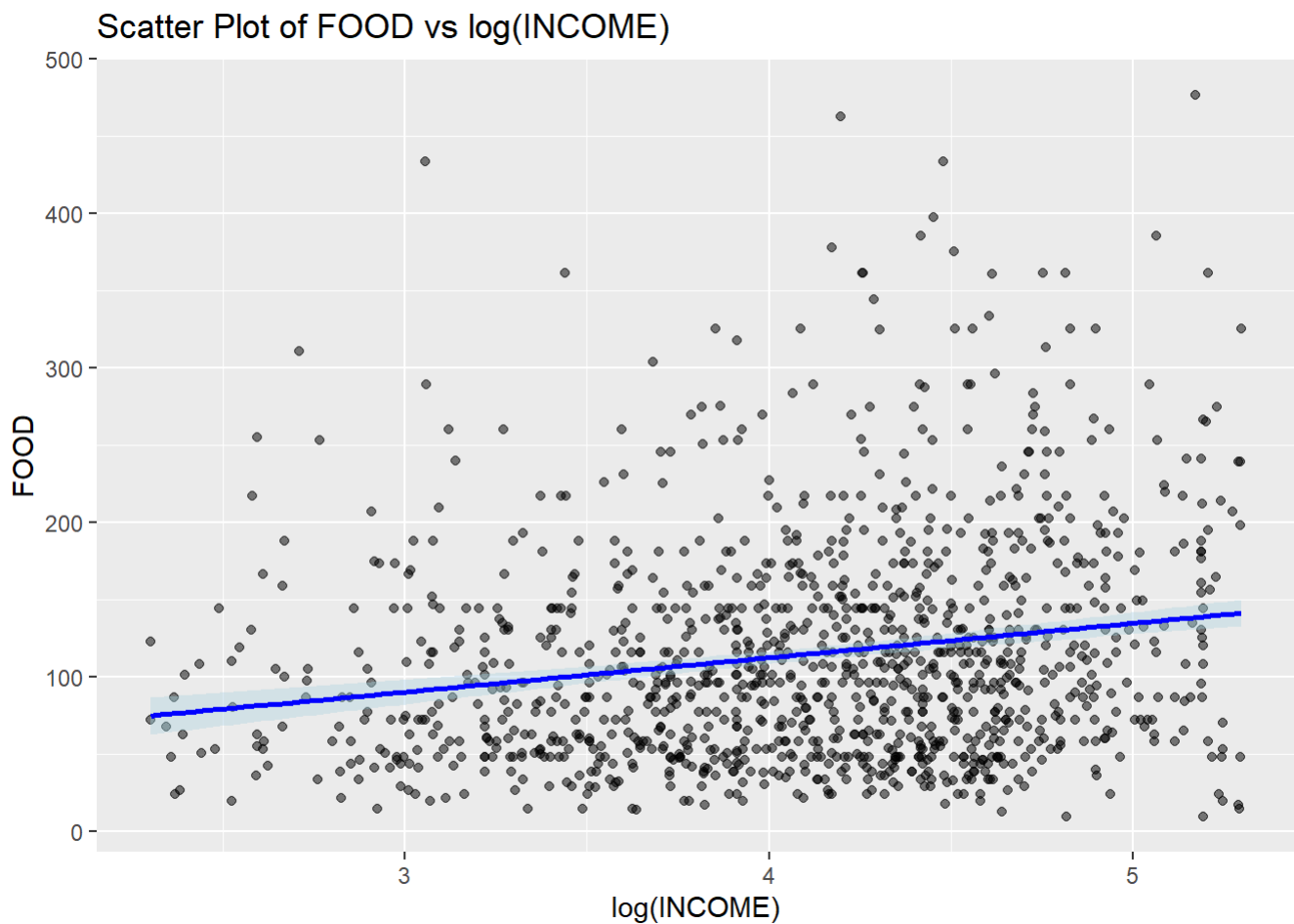
```
# 建立 linear-log 模型
model_linlog <- lm(food ~ log_income, data = cex5_small)

# 顯示模型摘要 (含 R²)
summary(model_linlog)
```

```
##
## Call:
## lm(formula = food ~ log_income, data = cex5_small)
##
## 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
```

```
# 畫出 FOOD vs Log(INCOME)
ggplot(cex5_small, aes(x = log_income, y = food)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "blue", fill = "lightblue") +
  labs(title = "Scatter Plot of FOOD vs log(INCOME)",
       x = "log(INCOME)", y = "FOOD")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We estimated the model:

$$FOOD = \alpha_1 + \alpha_2 \cdot \log(INCOME) + e$$

**Estimated coefficients:**

- Intercept = 23.568
- Slope = 22.187
- $R^2 = 0.038$

**Visual comparison:**

- The scatter plot shows a clearer trend than the linear model in (b), but still with substantial spread.
- It is more linear than the original model but less symmetric than the log-log model in (e).

**Conclusion:**

- The linear-log model slightly improves over the log-log model in terms of  $R^2$  (0.038 vs 0.0332).
- However, it still performs slightly worse than the original linear model ( $R^2 = 0.0423$ ).
- None of the models fit particularly well, but each has different strengths.

## Q29(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.

# Ans

```
# 資料點
income_vals <- c(19, 65, 160)

# 預測對應的 food (模型 fitted 值)
food_hat_linlog <- predict(model_linlog, newdata = data.frame(log_income = log(income_vals)))

# 係數與標準誤
alpha2 <- coef(model_linlog)["log_income"]
se_alpha2 <- summary(model_linlog)$coefficients["log_income", "Std. Error"]

# 彈性估計與信賴區間 (Delta method)
elasticity_point <- (alpha2 / food_hat_linlog) * income_vals
elasticity_se <- ((income_vals / food_hat_linlog)^2) * se_alpha2^2
elasticity_ci_lower <- elasticity_point - 1.96 * sqrt(elasticity_se)
elasticity_ci_upper <- elasticity_point + 1.96 * sqrt(elasticity_se)

# 結果整理
elasticity_table_i <- data.frame(
  INCOME = income_vals,
  FOOD_hat = round(food_hat_linlog, 4),
  Elasticity = round(elasticity_point, 4),
  Lower_95CI = round(elasticity_ci_lower, 4),
  Upper_95CI = round(elasticity_ci_upper, 4)
)

print(elasticity_table_i)
```

```
##   INCOME FOOD_hat Elasticity Lower_95CI Upper_95CI
## 1    19  88.8979    4.7421    3.3910    6.0932
## 2    65 116.1872   12.4126    8.8760   15.9491
## 3   160 136.1733   26.0696   18.6418   33.4974
```

We estimated elasticity at three income levels using:

$$\text{Elasticity} = \frac{\alpha_2}{\widehat{FOOD}} \cdot INCOME$$

INCOME	Predicted FOOD	Elasticity	95% CI
19	88.8979	4.7421	[3.3910, 6.0932]
65	116.1872	12.4126	[8.8760, 15.9491]
160	136.1733	26.0696	[18.6418, 33.4974]

## Comparison:

- Elasticity estimates are **much larger** than those from the linear and log-log models.
- Confidence intervals do **not overlap**, indicating statistical difference.
- This suggests the linear-log model implies **extremely high elasticity**, which is **unrealistic** for food expenditure.

## Conclusion:

- The elasticity estimates from the linear-log model are statistically and economically **dissimilar** from the others.
- This model may **overstate the responsiveness** of food expenditure to income.

## Q29(j)

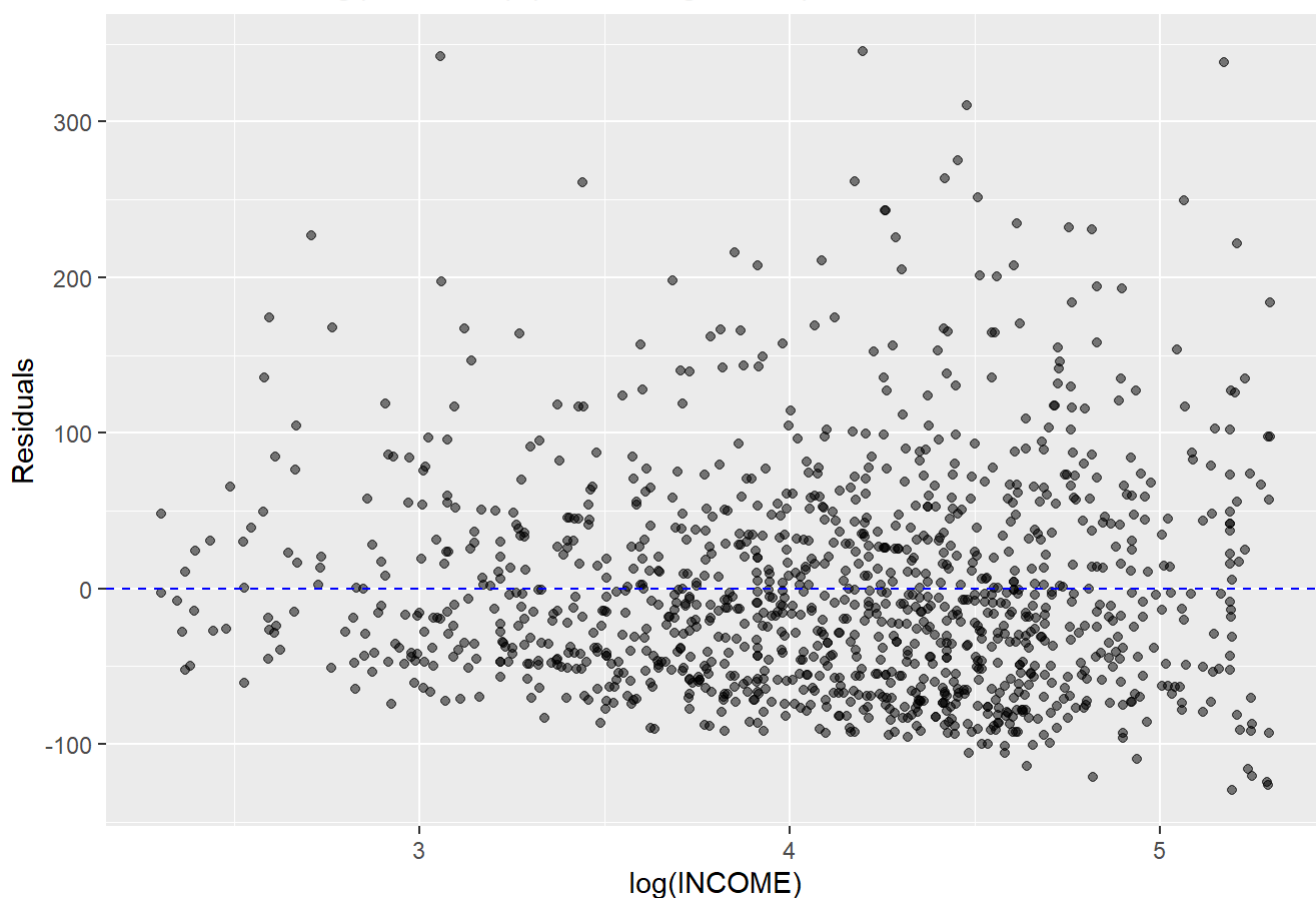
Obtain the least squares residuals from the linear-log model and plot them against  $\ln(\text{INCOME})$ . Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

## Ans

```
# 取得殘差
residuals_linlog <- residuals(model_linlog)

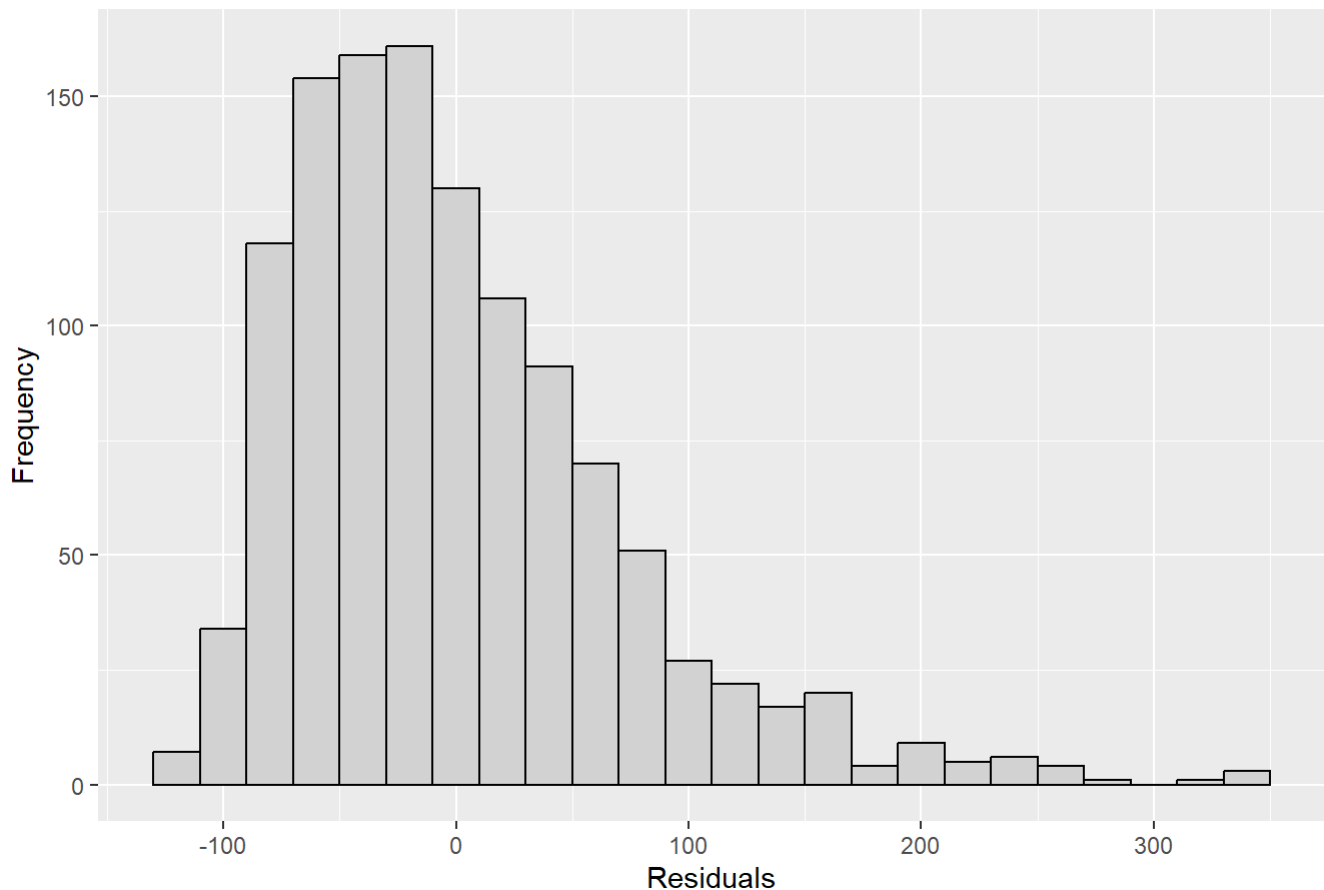
# 殘差 vs Ln(INCOME)
ggplot(cex5_small, aes(x = log_income, y = residuals_linlog)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "blue", linetype = "dashed") +
  labs(title = "Residuals vs log(INCOME) (Linear-Log Model)",
       x = "log(INCOME)", y = "Residuals")
```

Residuals vs log(INCOME) (Linear-Log Model)



```
# 殘差直方圖
ggplot(data.frame(residuals = residuals_linlog), aes(x = residuals)) +
  geom_histogram(binwidth = 20, fill = "lightgray", color = "black") +
  labs(title = "Histogram of Residuals (Linear-Log Model)",
       x = "Residuals", y = "Frequency")
```

Histogram of Residuals (Linear-Log Model)



```
# Jarque-Bera test
jarque.bera.test(residuals_linlog)
```

```
##
##  Jarque Bera Test
##
## data:  residuals_linlog
## X-squared = 628.07, df = 2, p-value < 2.2e-16
```

#### Residuals vs log(INCOME):

- The residuals show a right-skewed pattern with increasing spread at higher income levels.
- Suggests some heteroskedasticity.

#### Histogram of residuals:

- Strongly right-skewed, not bell-shaped.
- Indicates departure from normality.

#### Jarque-Bera test:

- JB = 628.07, p-value < 2.2e-16 → Strong rejection of normality.

#### Conclusion:

- The residuals from the linear-log model are **not normally distributed**.
- The model violates the normality assumption, and the error structure is not well-behaved.
- Compared to the log-log model, this one performs **worse** in terms of residual normality.



# Q29(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.




## Ans

After comparing all three models, I choose the **log-log model**:

$$\log(FOOD) = \gamma_1 + \gamma_2 \cdot \log(INCOME) + e$$

Although its  $R^2$  is slightly lower than the linear model, the log-log model has the **best residual behavior**. The residuals are much closer to normal, show no strong pattern, and the Jarque–Bera test statistic is far lower than for the other models.

### Comparison of Models

Model	Equation	Elasticity	$R^2$	JB Test (p-value)	Residual Shape
Linear	$FOOD = \beta_1 + \beta_2 \cdot INCOME$	Varies (small)	0.0423	624.19 (p < 2e-16) 	Right-skewed, wide
Log-Log	$\log(FOOD) = \gamma_1 + \gamma_2 \cdot \log(INCOME)$	Constant $\approx$ 0.186	0.0332	25.85 (p = 2e-6) 	Symmetric, centered
Linear-Log	$FOOD = \alpha_1 + \alpha_2 \cdot \log(INCOME)$	Varies (too large)	0.0380	628.07 (p < 2e-16) 	Strongly right-skewed

### Conclusion:

- The **log-log model** provides the best overall trade-off:
  - It has **stable, approximately normal residuals**.
  - Its elasticity estimate is constant and **economically plausible** (below 1).
- Despite a lower  $R^2$ , the **model assumptions are better satisfied**.
- Therefore, I select the log-log specification as the most appropriate model for analyzing food expenditure.