**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:

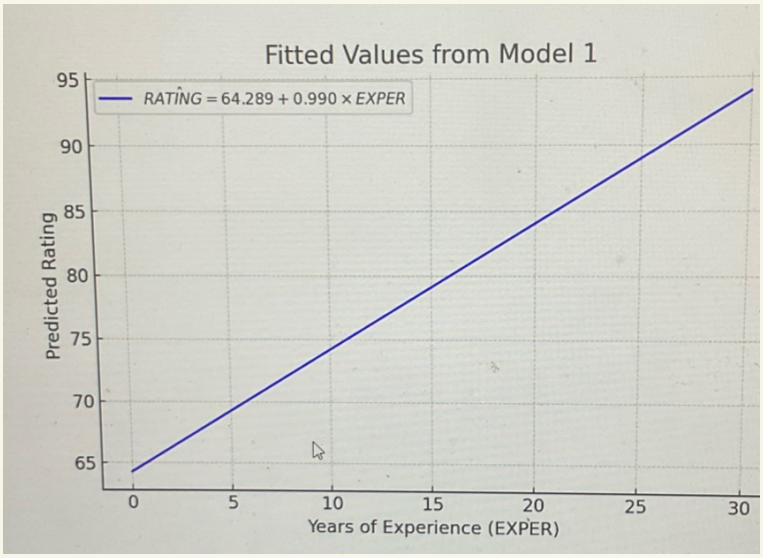
Model 1:

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

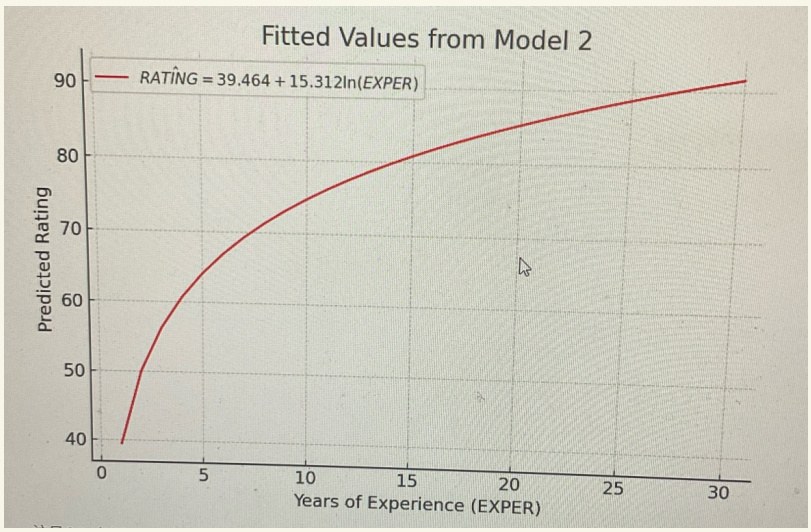
Model 2:

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

**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.




∵ ln(0) doesn't exist
∴ experience = 0 can't be used

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.

$$\text{marginal effect} = \frac{d(RATING)}{d(EXPER)} = 0.990$$

∴ No matter what value experience is, increase one year of EXPER, Rating increase 0.99 point

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.

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

(i) EXPER = 10, $\text{marginal effect} = \frac{15.312}{10} = 1.5312$

(ii) EXPER = 20, $\text{marginal effect} = \frac{15.312}{20} = 0.7656$

當藝術家有10年經驗，額外1年經驗提升1.5312分評分

20 ⋯⋯⋯⋯⋯⋯ 0.7656分評分

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

Model 1: $R^2 = 0.3793$

Model 1 (with some experience yield): $R^2 = 0.4858$

Model 2: $R^2 = 0.6414$

∵ Model 2 has larger $R^2$ ∴ it can fit data better

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

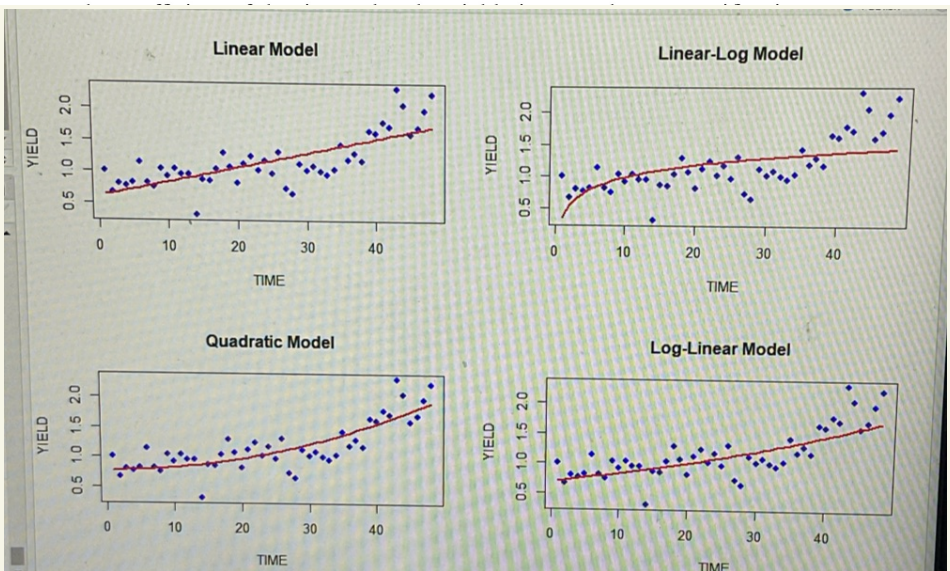
Model 2更符合經濟學原理，在現實情況中，職場經驗在前幾年帶來更大的回報，但到了一定階段，額外的經驗不會有太大增值，符合邊際報酬遞減現象

**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:

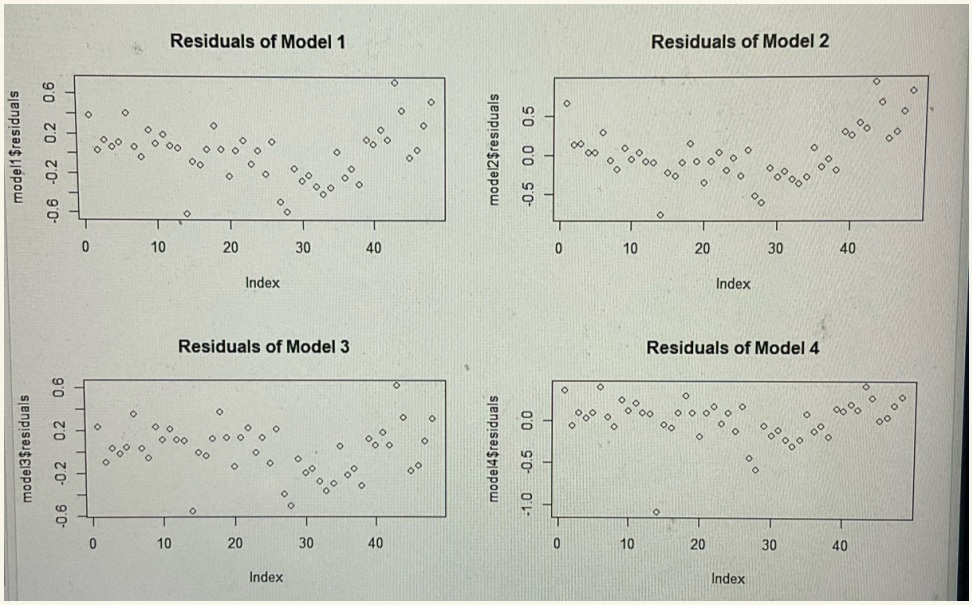
$$YIELD_t = \beta_0 + \beta_1 TIME + e_t$$
$$YIELD_t = \alpha_0 + \alpha_1 \ln(TIME) + e_t$$
$$YIELD_t = \gamma_0 + \gamma_1 TIME^2 + e_t$$
$$\ln(YIELD_t) = \phi_0 + \phi_1 TIME + e_t$$

**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.

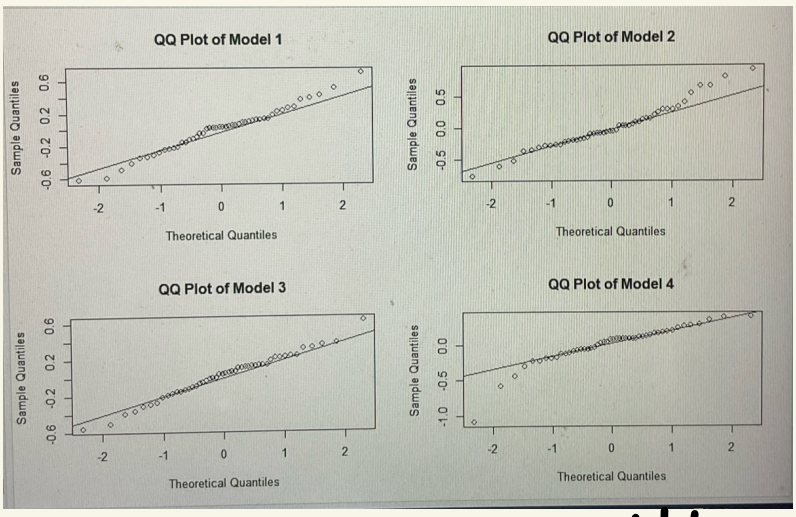
(i)



(ii)




(iii)




```
> print(result_table)
    Model                                  Test         W  p_value
1 Model 1 Shapiro-Wilk Normality Test 0.9823582    0.6792
2 Model 2 Shapiro-Wilk Normality Test 0.9665694    0.1856
3 Model 3 Shapiro-Wilk Normality Test 0.9858917    0.8266
4 Model 4 Shapiro-Wilk Normality Test 0.8689369 7.205e-05
```

根據 shapiro-Wilk 正態性檢驗，如果P值 > 0.05，說明沒有顯注偏離正態分佈
∴ model 1, 2, 3 的殘差可以視為正態分佈...

(iv)

```
> names(adj_r2_values) <- c("Model 1", "Model 2", "Model 3", "Model 4")
> print(adj_r2_values)
  Model 1   Model 2   Model 3   Model 4
0.5686594 0.3241945 0.6822494 0.4966469
```

Model 3 adj-$R^2$ 最大，且殘差為正態
因此是最優模型

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

```
最優模型為: Model 3
> #(b)
> summary(model3)

Call:
lm(formula = YIELD ~ TIME2, data = northampton_data)

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 ***
TIME2       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
```

$\beta_1 = 0.000486$，表示 TIME$^2$ 與 YIELD 之間是正相關

t value = 10.10，p-value = 3.01e-13 表示 TIME$^2$ 對於 YIELD 有很強的統計顯著性

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

| | residuals | leverage | dfbetas (Intercept) | dfbetas TIME2 | dffits |
|---|---|---|---|---|---|
| | -2.5606825 | 0.03593775 | -0.489450177 | 0.320519995 | -0.4944002 |
| | -2.2468473 | 0.02083617 | -0.219908713 | 0.003822742 | -0.3277591 |
| | 2.8894474 | 0.06829921 | -0.202525494 | 0.652179762 | 0.7823199 |
| | 1.3788286 | 0.07643579 | -0.116349648 | 0.338316939 | 0.3966661 |
| | -0.7794852 | 0.08542511 | 0.077302871 | -0.207150911 | -0.2382270 |
| | -0.5694893 | 0.09531255 | 0.065201030 | -0.163400687 | -0.1848466 |
| 7 | 0.4111600 | 0.10614453 | -0.053605470 | 0.127022369 | 0.1416857 |
| 7 | 1.3884647 | 0.11796846 | -0.203926321 | 0.460766575 | 0.5077802 |

studentized threshold = ±2

unusual observation:

14. 28. 43

| | outlier_residuals | outlier_leverage | outlier_dfbetas | outlier_dffits | outlier_dfbetas_TIME2 |
|---|---|---|---|---|---|
| 14 | TRUE | FALSE | TRUE | TRUE | TRUE |
| 28 | TRUE | FALSE | FALSE | FALSE | FALSE |
| 43 | TRUE | FALSE | TRUE | TRUE | TRUE |
| 44 | FALSE | FALSE | TRUE | FALSE | TRUE |
| 45 | FALSE | TRUE | FALSE | FALSE | FALSE |
| 46 | FALSE | TRUE | FALSE | FALSE | FALSE |
| 47 | FALSE | TRUE | FALSE | FALSE | FALSE |
| 48 | FALSE | TRUE | TRUE | TRUE | TRUE |

Leverage threshold = $\frac{2p}{n} = 0.083$

unusual observation:

45. 46. 47. 48

dfbetas threshold = $\pm\frac{2}{\sqrt{n}} = 0.288$

unusual observation:

14. 43. 44. 48

dffits threshold = $\pm\frac{2}{\sqrt{\frac{p}{n}}} = 0.408$

unusual observation:

14. 43. 48

**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?

```
> print(pred_1997)
       fit      lwr      upr
1 1.875113 1.367898 2.382328
> # 1997 年的實際值
> actual_1997 <- wa_wheat$northampton[wa_wheat$time == 48]
> # 比較
> print(actual_1997)
[1] 2.2318
```

95% PI
=[1.368, 2.382]
2.2318有落在這個區間

4.28 R

```r
data("wa_wheat")
northampton_data <- data.frame(
  YIELD = wa_wheat$northampton,
  TIME = wa_wheat$time
)
#a(i)
model1 <- lm(YIELD ~ TIME, data = northampton_data)
northampton_data$ln_TIME <- log(northampton_data$TIME)
model2 <- lm(YIELD ~ ln_TIME, data = northampton_data)
northampton_data$TIME2 <- northampton_data$TIME^2
model3 <- lm(YIELD ~ TIME2, data = northampton_data)
northampton_data$ln_YIELD <- log(northampton_data$YIELD)
model4 <- lm(ln_YIELD ~ TIME, data = northampton_data)
par(mfrow = c(2, 2))  # 設定 2x2 圖形排列
plot(northampton_data$TIME, northampton_data$YIELD, main = "Linear Model",
     xlab = "TIME", ylab = "YIELD", col = "blue", pch = 16)
lines(northampton_data$TIME, fitted(model1), col = "red", lwd = 2)
plot(northampton_data$TIME, northampton_data$YIELD, main = "Linear-Log Model",
     xlab = "TIME", ylab = "YIELD", col = "blue", pch = 16)
lines(northampton_data$TIME, fitted(model2), col = "red", lwd = 2)
plot(northampton_data$TIME, northampton_data$YIELD, main = "Quadratic Model",
     xlab = "TIME", ylab = "YIELD", col = "blue", pch = 16)
lines(northampton_data$TIME, fitted(model3), col = "red", lwd = 2)
plot(northampton_data$TIME, northampton_data$YIELD, main = "Log-Linear Model",
     xlab = "TIME", ylab = "YIELD", col = "blue", pch = 16)
lines(northampton_data$TIME, exp(fitted(model4)), col = "red", lwd = 2)  # 轉回 YIELD
#a(ii)
par(mfrow = c(2,2))
plot(model1$residuals, main="Residuals of Model 1")
plot(model2$residuals, main="Residuals of Model 2")
plot(model3$residuals, main="Residuals of Model 3")
plot(model4$residuals, main="Residuals of Model 4")
```

```r
#a(iii)
par(mfrow = c(2,2))
qqnorm(model1$residuals, main="QQ Plot of Model 1"); qqline(model1$residuals)
qqnorm(model2$residuals, main="QQ Plot of Model 2"); qqline(model2$residuals)
qqnorm(model3$residuals, main="QQ Plot of Model 3"); qqline(model3$residuals)
qqnorm(model4$residuals, main="QQ Plot of Model 4"); qqline(model4$residuals)
model1_result <- shapiro.test(model1$residuals)
model2_result <- shapiro.test(model2$residuals)
model3_result <- shapiro.test(model3$residuals)
model4_result <- shapiro.test(model4$residuals)
result_table <- data.frame(    #創建一個包含每個模型檢驗結果的 data.frame
  Model = c("Model 1", "Model 2", "Model 3", "Model 4"),
  Test = rep("Shapiro-Wilk Normality Test", 4),  # 添加檢驗名稱
  W = c(model1_result$statistic, model2_result$statistic, model3_result$statistic, model4_result$statistic),
  p_value = c(format(model1_result$p.value, digits = 4),
              format(model2_result$p.value, digits = 4),
              format(model3_result$p.value, digits = 4),
              format(model4_result$p.value, digits = 4))
)
print(result_table)
#a(iv)
adj_r2_values <- c(
  summary(model1)$adj.r.squared,
  summary(model2)$adj.r.squared,
  summary(model3)$adj.r.squared,
  summary(model4)$adj.r.squared
)
names(adj_r2_values) <- c("Model 1", "Model 2", "Model 3", "Model 4")
print(adj_r2_values)
best_model <- names(which.max(adj_r2_values))
cat("最優模型為:", best_model, "\n")
```

```r
#(b)
summary(model3)

#(c)
wa_wheat$residuals <- rstudent(model3)   # Studentized residuals
wa_wheat$leverage <- hatvalues(model3)   # Leverage 值
wa_wheat$dfbetas  <- dfbetas(model3)     # DFBETAS 值
wa_wheat$dffits   <- dffits(model3)      # DFFITS 值


n <- nrow(wa_wheat)   # 觀測值個數
p <- length(coef(model3))  # 回歸參數個數

# 常見的異常值判斷標準
threshold_residual <- 2  # Studentized residuals 絕對值大於 2
threshold_leverage <- 2 * p / n  # Leverage 高於 2p/n
threshold_dfbetas  <- 2 / sqrt(n)  # DFBETAS 絕對值大於 2/sqrt(n)
threshold_dffits   <- 2 * sqrt(p / n)  # DFFITS 絕對值大於 2 * sqrt(p/n)

# 標記異常值
wa_wheat$outlier_residuals <- abs(wa_wheat$residuals) > threshold_residual
wa_wheat$outlier_leverage  <- wa_wheat$leverage > threshold_leverage
wa_wheat$outlier_dfbetas_TIME2 <- abs(wa_wheat$dfbetas[, "TIME2"]) > threshold_dfbetas
wa_wheat$outlier_dfbetas <- wa_wheat$outlier_dfbetas_TIME2
wa_wheat$outlier_dffits    <- abs(wa_wheat$dffits) > threshold_dffits

# 過濾出異常觀測值
outliers <- wa_wheat[wa_wheat$outlier_residuals |
                     wa_wheat$outlier_leverage |
                     wa_wheat$outlier_dffits |
                     wa_wheat$outlier_dfbetas_TIME2 , ]
print(outliers)
```

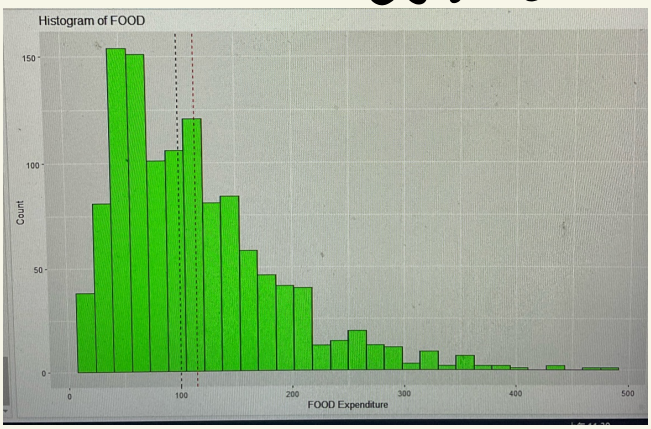
```r
#(d)
data_1996 <- subset(wa_wheat, time < 47)
model3_1996 <- lm(northampton ~ I(time^2), data = wa_wheat)
# 找出 1997 年的 time 值
time_1997 <- wa_wheat$time[wa_wheat$time == 47]
# 對 1997 年進行預測
pred_1997 <- predict(model3_1996,
                     newdata = data.frame(time = time_1997),
                     interval = "prediction", level = 0.95)
print(pred_1997)
# 1997 年的實際值
actual_1997 <- wa_wheat$northampton[wa_wheat$time == 48]
print(actual_1997)
```

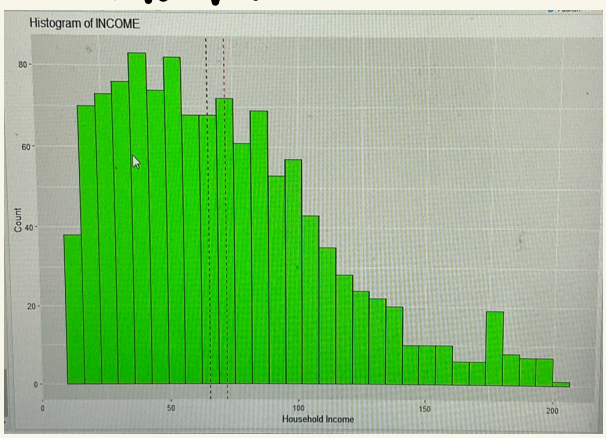
**4.29** Consider a model for household expenditure as a function of household income using the 2013 data from the Consumer Expenditure Survey, *cex5_small*. The data file *cex5* contains more observations. Our attention is restricted to three-person households, consisting of a husband, a wife, plus 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.

```
> print(summary_stats)
  Variable      Mean Median   Min    Max  Std_Dev
1     FOOD 114.44311  99.80  9.63 476.67 72.65750
2   INCOME  72.14264  65.29 10.00 200.00 41.65228
```

red line: mean , black line: median



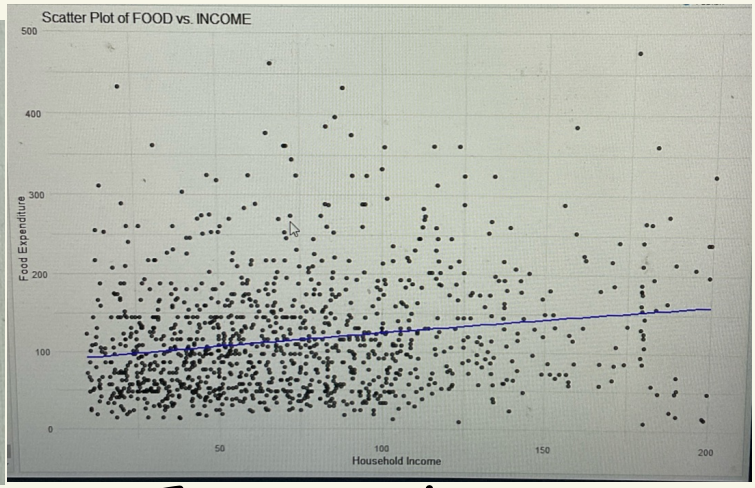

兩個分佈 mean > median ∴都是右偏分布
而不是鐘型或對稱分布

```
        Jarque Bera Test

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

> print(jb_income)

        Jarque Bera Test

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

兩個檢驗 p-value
都非常小 ∴ reject $H_0$
都不符合常態分布

**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?

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

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
```



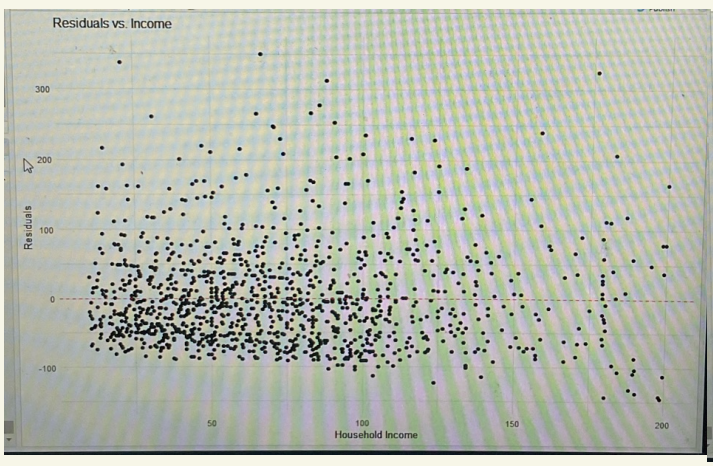

```
                2.5 %    97.5 %
(Intercept) 80.5064570 96.626543
income       0.2619215  0.455452
>
```

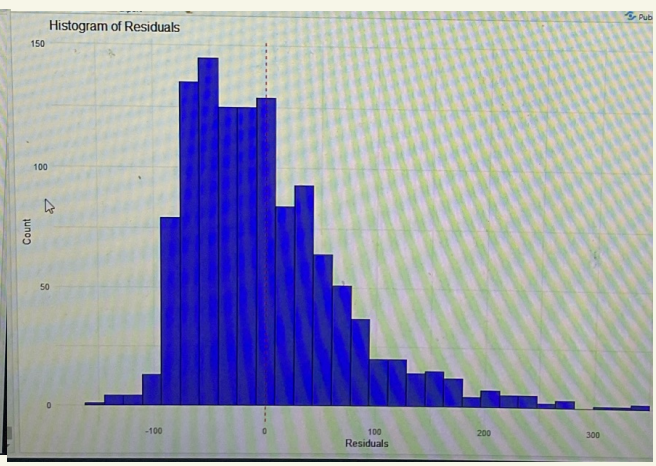
95% interval estimate

[0.2619, 0.4555]

the effect of changing income on food expenditure may not be estimated very precisely, as the low $R^2$ value indicates that the model explain only a small proportion of the variation in food expenditure.

**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.



```
        Jarque Bera Test

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

p-value<0.05 ∴ reject $H_0$
不符合常態分布

殘差大致分佈在0附近，沒有嚴重偏移，但在高收入(Income>150)時，殘差的變異數變大，可能有異質變異數問題。

It is more important for the random error to be normally distributed, because the validity of hypothesis tests in regression analysis relies on the normality of the residuals

**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?

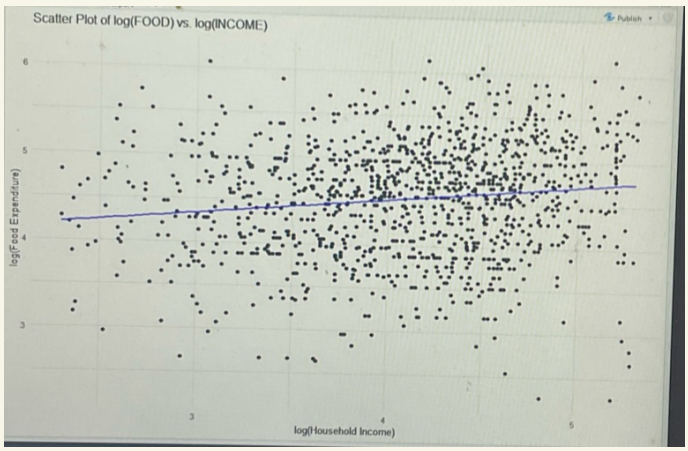
$$E = \beta_2 \times \frac{INCOME}{Food} \qquad \beta_2 = 0.3587$$

```
> print(elasticity_df)
  Income Predicted FOOD Elasticity Lower_95CI Upper_95CI
1     19       95.38155 0.07145302 0.05217047 0.09073558
2     65      111.88114 0.20839527 0.15215702 0.26463353
3    160      145.95638 0.39321338 0.28709948 0.49932727
>
```

the estimated elasticities are dissimilar.
INCOME increase lead to a higher elasticity.

the interval estimated don't overlap(不重疊).
Based on Economics principles, food is necessity good (必需品), so elasticity must less than 1 . and with income increase, the elasticity will decrease.

**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?




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

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
```

```
> print(paste("Generalized R² for Log-Log Model: ", generalized_r2_log_log))
[1] "Generalized R2 for Log-Log Model:  0.0396516120669934"
>
```

generalized log-log model $R^2 = 0.0397 < 0.0423 =$ linear model $R^2$

∴ linear model fit the data better

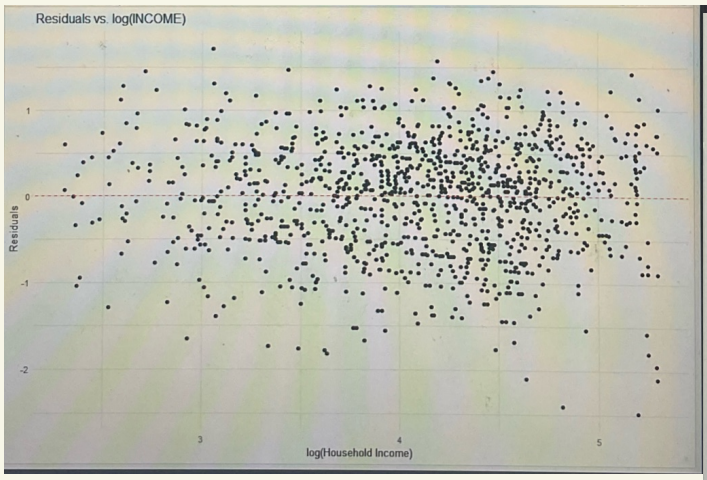
**f.** Construct a point and 95% interval estimate of the elasticity for the log-log model. Is the elasticity of food expenditure from the log-log model similar to that in part (d), or dissimilar? Provide statistical evidence for your claim.

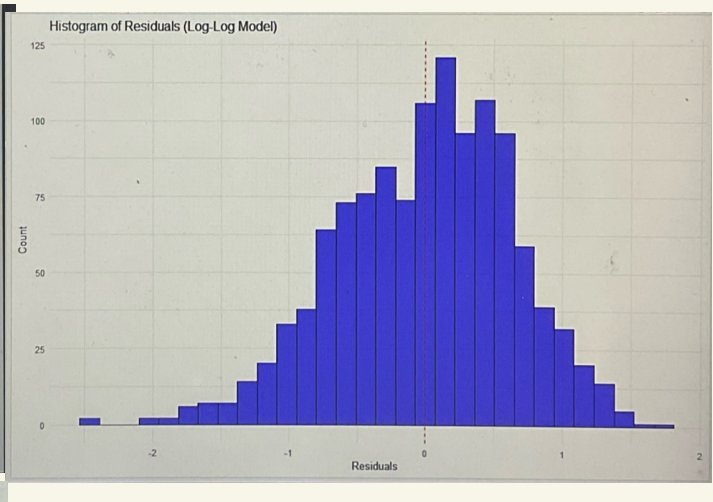
```
> confint(model_loglog, level = 0.95)
                2.5 %    97.5 %
(Intercept) 3.5428135 4.0150507
log(income) 0.1293432 0.2432675
>
```

95% interval estimate
$= [0.1293, 0.2433]$

$E = \gamma_2 = 0.1863$, it is a constant so dissimilar to the elasticity of linear model.

**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?



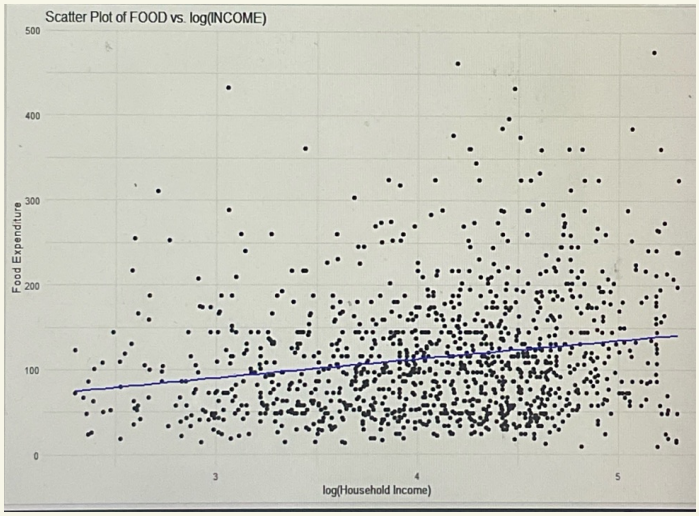

```
        Jarque Bera Test

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

residuals seems no any strong pattern

In Jarque Bera Test, p-value < 0.05.

∴ reject $H_0$, 不符合常態分布

**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?




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

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
```

$R^2$ of linear model $= 0.0423$

Generalized $R^2$ of log-log model $= 0.0397$

$R^2$ of linear-log model $= 0.038$

By $R^2$, linear model 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.

```
> confint(model_linlog, level = 0.95)
                2.5 %   97.5 %
(Intercept) -2.661958 49.79892
log(income) 15.859458 28.51531
>
```

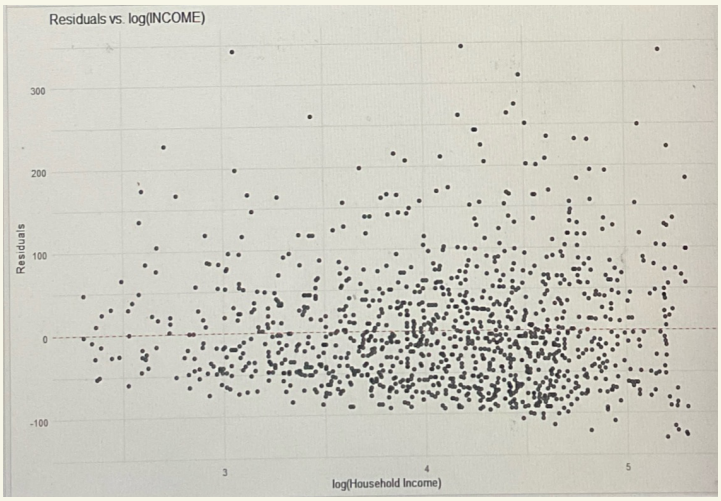
95% interval estimate $= [15.859, 28.515]$

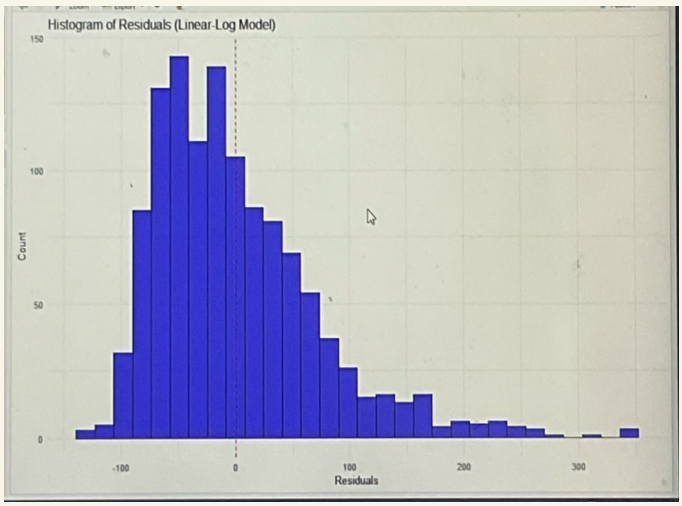
```
> print(income_values)
  income predicted_food elasticity elasticity_lower elasticity_upper
1     19       88.89627  0.2495830        0.1784889        0.3206771
2     65      116.18513  0.1909625        0.1365665        0.2453584
3    160      136.17088  0.1629350        0.1165227        0.2093473
> #(i)
```

$$E = \frac{\alpha_2}{INCOME}$$

Income increase lead to a decrease on elasticity, so it is dissimilar to other models

**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?




```
        Jarque Bera Test

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

residuals seems no any strong pattern

In Jarque Bera Test, p-value < 0.05.

∴ reject $H_0$, 不符合常態分布

**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.

linear model 的所得彈性隨收入增加而上升，不符合經濟學原理首先排除，log-log model 相比之下 $R^2$ 更高且殘差更接近隨機分佈，因此選 log-log model

4.29 R

```r
data(cex5_small)
df <- cex5_small
#(a)
summary_stats <- data.frame(
  Variable = c("FOOD", "INCOME"),
  Mean = c(mean(df$food), mean(df$income)),
  Median = c(median(df$food), median(df$income)),
  Min = c(min(df$food), min(df$income)),
  Max = c(max(df$food), max(df$income)),
  Std_Dev = c(sd(df$food), sd(df$income))
)
print(summary_stats)
ggplot(df, aes(x = food)) +
  geom_histogram(color="black", fill="green", bins=30, alpha=0.7) +
  geom_vline(aes(xintercept = mean(food)), color="red", linetype="dashed") +
  geom_vline(aes(xintercept = median(food)), color="black", linetype="dashed") +
  labs(title="Histogram of FOOD", x="FOOD Expenditure", y="Count")
ggplot(df, aes(x = income)) +
  geom_histogram(color="black", fill="green", bins=30, alpha=0.7) +
  geom_vline(aes(xintercept = mean(income)), color="red", linetype="dashed") +
  geom_vline(aes(xintercept = median(income)), color="black", linetype="dashed") +
  labs(title="Histogram of INCOME", x="Household Income", y="Count")
jb_food <- jarque.bera.test(df$food)
jb_income <- jarque.bera.test(df$income)
print(jb_food)
print(jb_income)

#(b)
model_linear <- lm(food ~ income, data = df)
summary(model_linear)
ggplot(df, aes(x = income, y = food)) +
  geom_point(alpha = 0.6) +  # Scatter points
  geom_smooth(method = "lm", color = "blue", se = FALSE) +  # Regression line
  labs(title = "Scatter Plot of FOOD vs. INCOME",
       x = "Household Income",
       y = "Food Expenditure") +
  theme_minimal()
confint(model_linear, level = 0.95)

#(c)
df$residuals <- residuals(model_linear)
ggplot(df, aes(x = income, y = residuals)) +
  geom_point(alpha = 2) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs. Income",
       x = "Household Income",
       y = "Residuals") +
  theme_minimal()
ggplot(df, aes(x = residuals)) +
  geom_histogram(color="black", fill="blue", bins=30, alpha=0.7) +
  geom_vline(aes(xintercept = mean(residuals)), color="red", linetype="dashed") +
  labs(title="Histogram of Residuals", x="Residuals", y="Count") +
  theme_minimal()
jb_test_residuals <- jarque.bera.test(df$residuals)
print(jb_test_residuals)
```

```r
#(d)
income_values <- data.frame(income = c(19, 65, 160))
predicted_food <- predict(model_linear, newdata = income_values)
elasticity <- 0.3587 * (income_values$income / predicted_food)
elasticity_df <- data.frame(income_values$income, predicted_food, elasticity)
colnames(elasticity_df) <- c("Income", "Predicted FOOD", "Elasticity")
elasticity_lower <- 0.2619 * (income_values$income / predicted_food)
elasticity_upper <- 0.4555 * (income_values$income / predicted_food)
elasticity_df$Lower_95CI <- elasticity_lower
elasticity_df$Upper_95CI <- elasticity_upper
print(elasticity_df)
```

```r
#(e)
model_loglog <- lm(log(food) ~ log(income), data = df)
ggplot(df, aes(x = log(income), y = log(food))) +
  geom_point(alpha = 2) +  # Scatter points
  geom_smooth(method = "lm", color = "blue", se = FALSE) +  # Regression line
  labs(title = "Scatter Plot of log(FOOD) vs. log(INCOME)",
       x = "log(Household Income)",
       y = "log(Food Expenditure)") +
  theme_minimal()
summary(model_loglog)
# 確保 df 中的 food 和 income 都大於 0
data_log <- subset(df, food > 0 & income > 0)
# 創建 log 變數
data_log$ln_food <- log(data_log$food)
data_log$ln_income <- log(data_log$income)
model_loglog <- lm(ln_food ~ ln_income, data = data_log)

# 預測 Log-Log 模型的預測值
data_log$food_pred <- exp(predict(model_loglog))  # 需要將預測值反log，回到原始單位

# 計算 Generalized R² 的函數
generalized_r2_corr <- function(actual, predicted) {
  r_g <- cor(actual, predicted)^2  # 計算相關係數平方
  return(r_g)
}
# 計算 Log-Log Model 的 Generalized R²
generalized_r2_log_log <- generalized_r2_corr(data_log$food, data_log$food_pred)
print(paste("Generalized R² for Log-Log Model: ", generalized_r2_log_log))

#(f)
elasticity_loglog <- coef(model_loglog)["log(income)"]
print(elasticity_loglog)
confint(model_loglog, level = 0.95)
```

```r
#(g)
df$residuals_loglog <- residuals(model_loglog)
ggplot(df, aes(x = log(income), y = residuals_loglog)) +
  geom_point(alpha = 2) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs. log(INCOME)",
       x = "log(Household Income)",
       y = "Residuals") +
  theme_minimal()
ggplot(df, aes(x = residuals_loglog)) +
  geom_histogram(color="black", fill="blue", bins=30, alpha=0.7) +
  geom_vline(aes(xintercept = mean(residuals_loglog)), color="red", linetype="dashed") +
  labs(title="Histogram of Residuals (Log-Log Model)", x="Residuals", y="Count") +
  theme_minimal()
jb_test_residuals_loglog <- jarque.bera.test(df$residuals_loglog)
print(jb_test_residuals_loglog)

#(h)
model_linlog <- lm(food ~ log(income), data = df)
summary(model_linlog)
ggplot(df, aes(x = log(income), y = food)) +
  geom_point(alpha = 2) +  # Scatter points
  geom_smooth(method = "lm", color = "blue", se = FALSE) +  # Regression line
  labs(title = "Scatter Plot of FOOD vs. log(INCOME)",
       x = "log(Household Income)",
       y = "Food Expenditure") +
  theme_minimal()
```

```r
#(i)
confint(model_linlog, level = 0.95)
income_values <- data.frame(income = c(19, 65, 160))
income_values$predicted_food <- 23.568 + 22.187 * log(income_values$income)
income_values$elasticity <- 22.187 / income_values$predicted_food
income_values$elasticity_lower <- 15.867 / income_values$predicted_food
income_values$elasticity_upper <- 28.507 / income_values$predicted_food
print(income_values)

#(j)
df$residuals_linlog <- residuals(model_linlog)
ggplot(df, aes(x = log(income), y = residuals_linlog)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs. log(INCOME)",
       x = "log(Household Income)",
       y = "Residuals") +
  theme_minimal()
ggplot(df, aes(x = residuals_linlog)) +
  geom_histogram(color="black", fill="blue", bins=30, alpha=0.7) +
  geom_vline(aes(xintercept = mean(residuals_linlog)), color="red", linetype="dashed") +
  labs(title="Histogram of Residuals (Linear-Log Model)", x="Residuals", y="Count") +
  theme_minimal()
jb_test_residuals_linlog <- jarque.bera.test(df$residuals_linlog)
print(jb_test_residuals_linlog)
```