**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:

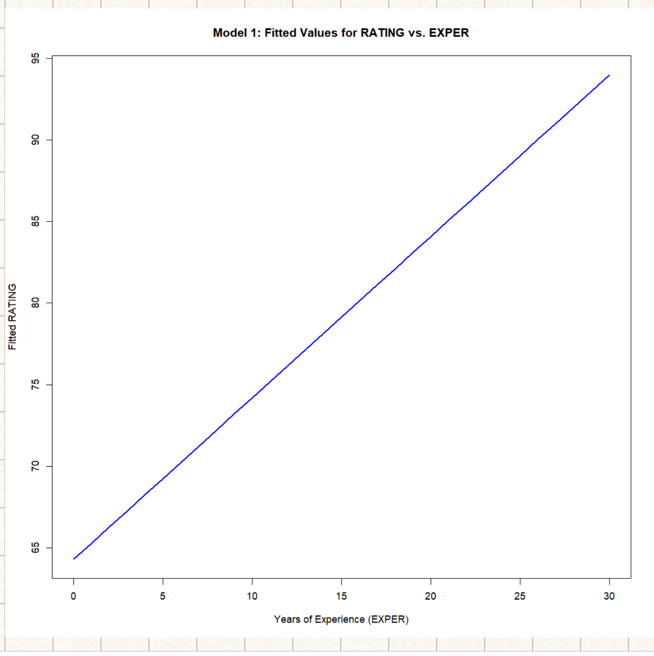
Model 1:

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

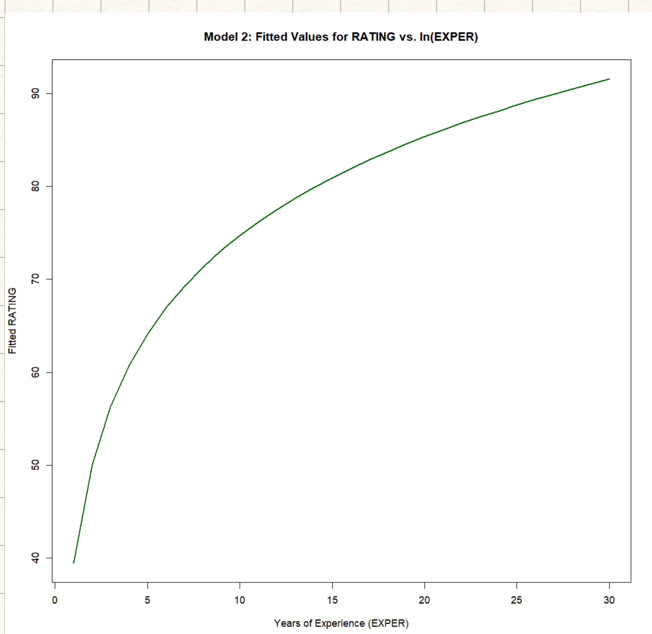
Model 2:

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

**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.




Because Model 2 uses: $\ln(EXPER)$

but: $\ln(0) = -\infty$

$\Rightarrow$ Therefore, observations with 0 years of experience cannot be included in the regression analysis, which is why the sample sizes decreases from 50 to 46.

**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.

Model 1:

$$\widehat{RATING} = 64.289 + 0.990EXPER$$

Since it is a linear model and $\dfrac{\partial \, RATING}{\partial \, EXPER} = 0.990$

$\Rightarrow$ for (i) and (ii), the marginal effect = 0.990 *

**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.

Model 2:

$$\widehat{RATING} = 39.464 + 15.312\ln(EXPER)$$

Marginal effect: $\dfrac{\partial \widehat{RATING}}{\partial \, EXPER} = \dfrac{15.312}{EXPER}$

for (i), the marginal effect $= \dfrac{15.312}{10} = 1.532$ *

(ii), the marginal effect $= \dfrac{15.312}{20} = 0.7656$ *

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

$R^2$ of Model 2 $= 0.6414 > 0.4858 = R^2$ of Model 1

$\Rightarrow$ Model 2 fits the data better *

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

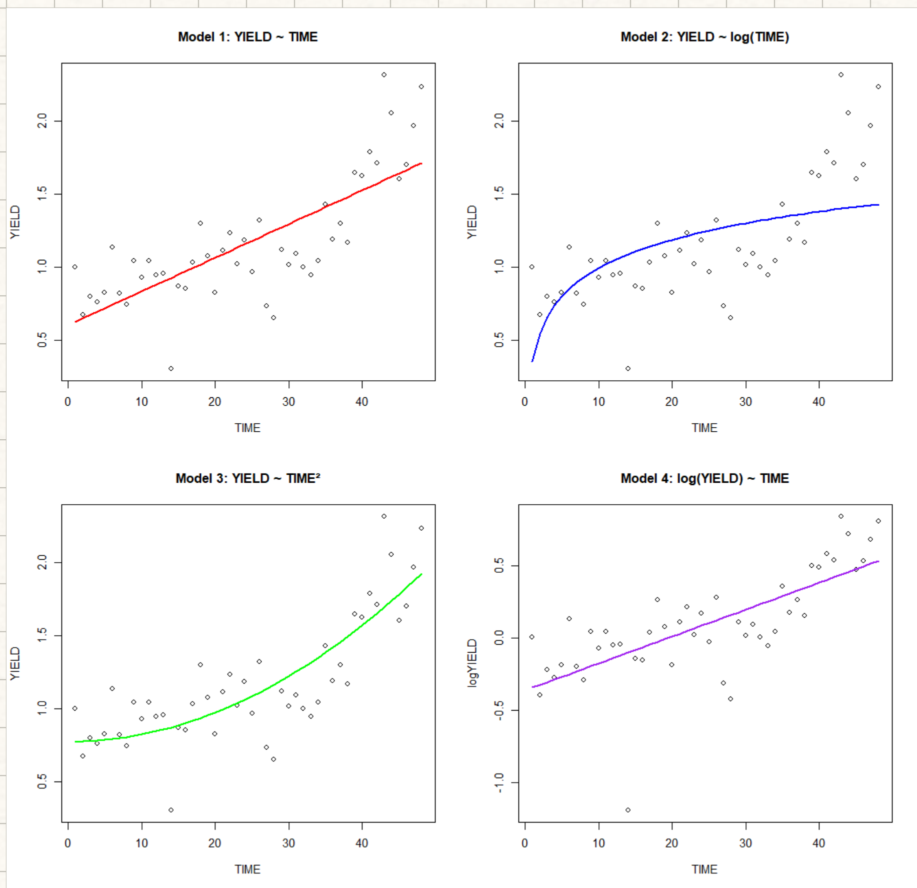
Model 2 is more reasonable because it reflects diminishing returns to experience — early years improve performance more than later years. This fits better with economic intuition than the constant effect assumed in Model 1. *

**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.

```
> # 模型比較用 R²
> summary(model1)$r.squared
[1] 0.5778369
> summary(model2)$r.squared
[1] 0.3385733
> summary(model3)$r.squared
[1] 0.6890101
> summary(model4)$r.squared
[1] 0.5073566
>
> # 常態性檢定
> shapiro.test(resid(model1))

        Shapiro-Wilk normality test

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

> shapiro.test(resid(model2))

        Shapiro-Wilk normality test

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

> shapiro.test(resid(model3))

        Shapiro-Wilk normality test

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

> shapiro.test(resid(model4))

        Shapiro-Wilk normality test

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




⇒ Model 3 (YIELD ~ TIME²) is the most appropriate model. It not only provides a good visual fit, but also exhibits randomly distributed residuals, passes the normality test of the errors, and has the highest $R^2$ value. Therefore, it most effectively captures the trend in wheat yield for the Northampton shire.

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

```
> summary(model3)

Call:
lm(formula = YIELD ~ TIME2)

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
```

$\Rightarrow$ Intercept $= 0.7737$

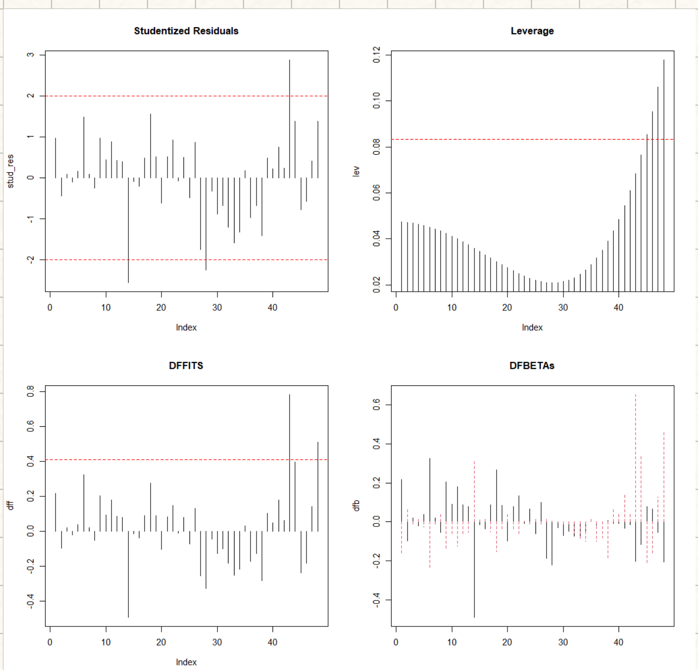
$TIME^2 = 0.0004986$

p-value $= 3.008e^{-13} < 0.01$

$R^2 = 0.689$

#

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




In Model 3, observations 46 and 47 are influential. Observation 46 shows large studentized residuals, high leverage, and exceeds the DEFITS threshold. Both 46 and 47 have high DEBETAs, indicating strong influence on the regression coefficients. These points should be further examined for data issues or structural changes.

#

**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?

```
> # 取出前 47 筆資料（1950-1996）
> YIELD_train <- YIELD[1:47]
> TIME_train <- TIME[1:47]
> TIME2_train <- TIME_train^2
>
> # 建立模型（重新估計）
> model3_train <- lm(YIELD_train ~ TIME2_train)
>
> # 建立 1997 的預測資料（TIME = 48, TIME² = 2304）
> new_data <- data.frame(TIME2_train = 48^2)
>
> # 預測 + 95% 預測區間
> predict(model3_train, newdata = new_data, interval = "prediction", level = 0.95)
       fit      lwr      upr
1 1.881111 1.372403 2.389819
>
> # 實際觀察值（1997 年）是第 48 筆
> true_value <- YIELD[48]
> true_value
[1] 2.2318
```

$\Rightarrow$ Fit $= 1.881111$

95% prediction interval : $[1.372403, 2.389819]$

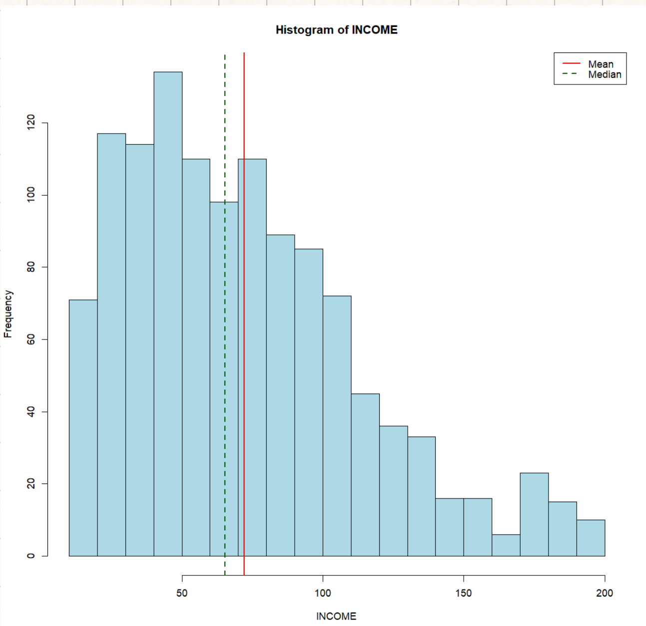
The true value in 1997 is $2.2318 \in [1.372403, 2.389819]$

$\Rightarrow$ the interval contains the true value #

**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 extended example, 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.

```
> describe(cex5_small[, c("food", "income")])
       vars    n   mean    sd median trimmed   mad  min    max  range skew kurtosis  se
food      1 1200 114.44 72.66  99.80  105.03 66.18  9.63 476.67 467.04 1.35     2.36 2.1
income    2 1200  72.14 41.65  65.29   67.94 41.65 10.00 200.00 190.00 0.84     0.32 1.2
```




```
> # Jarque-Bera 常態性檢定
> 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
```

Both are skew right, not bell-shaped, and the sample mean is larger than the median

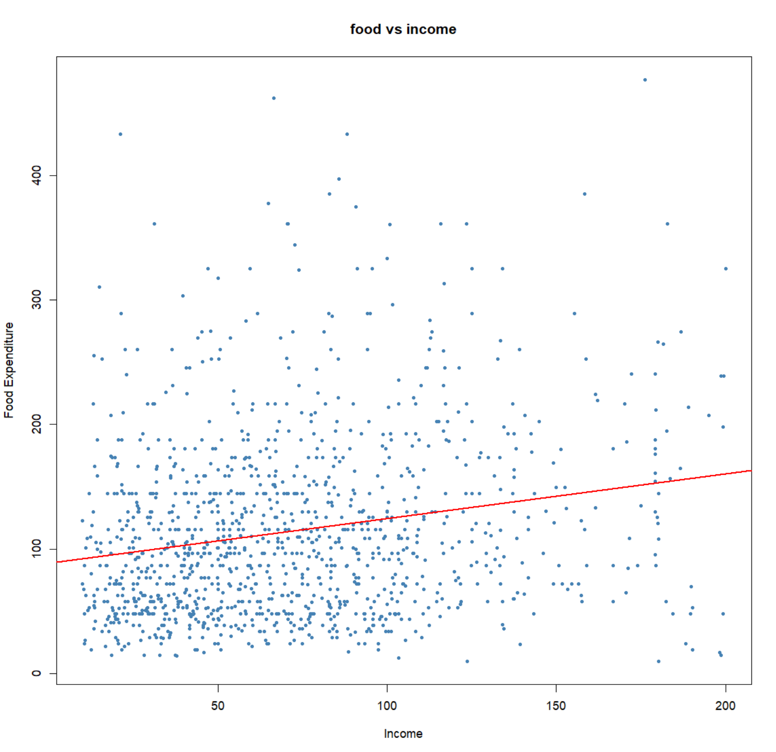
and since p-value << 0.05 for both food and income ⇒ reject the normality #

**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?

```
> model_linear <- lm(food ~ income, data = cex5_small)
> summary(model_linear)

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
```

```
> confint(model_linear, level = 0.95)
                 2.5 %    97.5 %
(Intercept) 80.5064570 96.626543
income       0.2619215  0.455452
```

95% interval = [0.2619215, 0.455452]

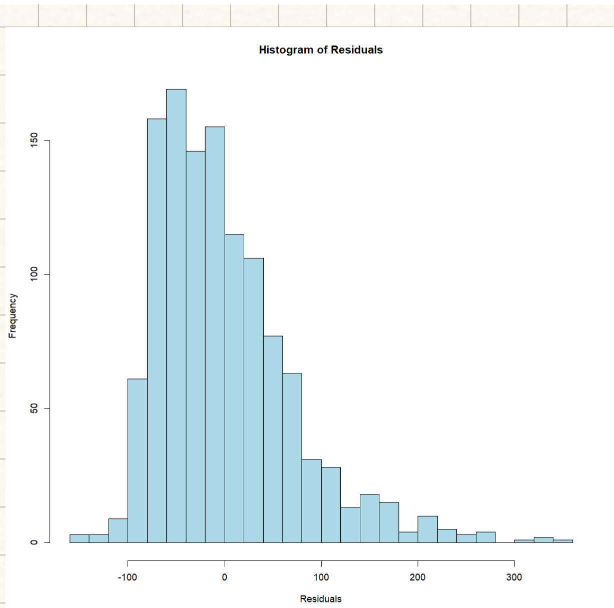



The regression shows a positive, significant effect of income on food spending: each unit increase income raises food expenditure by about 0.36 units (95% CI: [0.262, 0.455]). However, the low $R^2$ (4.2%) suggests income explains little of the variation in food spending.

**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(residuals_linear)

	Jarque Bera Test

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

⇒ Reject the null hypothesis of normally distributed error.




In linear regression, it's the normality of the error term — not the variables themselves — that matters for valid inference like t-tests and F-tests.

**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?

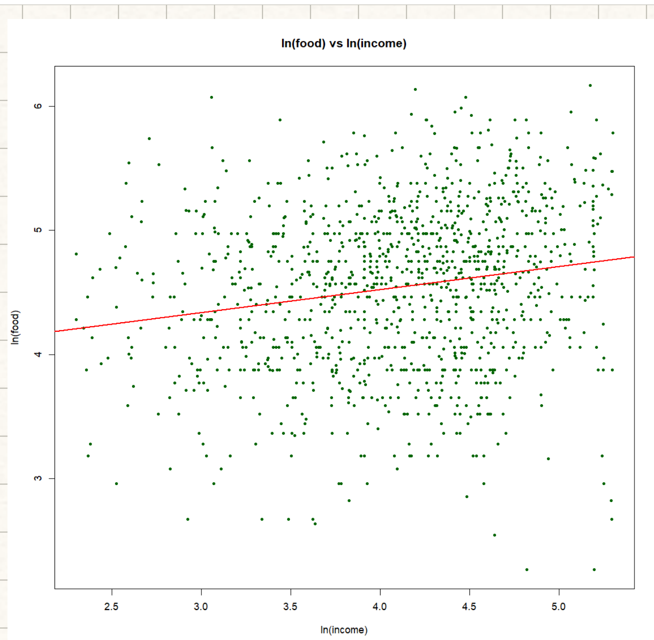
```
> result
  income fitted_food elasticity CI_lower CI_upper
1     19       95.38     0.0715   0.0522   0.0907
2     65      111.88     0.2084   0.1522   0.2646
3    160      145.96     0.3932   0.2871   0.4993
```

The elasticities are dissimilar, and their confidence interval do not fully overlap.

As income increases, the income elasticity of food should increase based on economic 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?

```
> # 建立 ln(food) 和 ln(income) 變數
> cex5_small$ln_food <- log(cex5_small$food)
> cex5_small$ln_income <- log(cex5_small$income)
>
> # 建立 log-log 模型
> model_loglog <- lm(ln_food ~ ln_income, data = cex5_small)
> summary(model_loglog)

Call:
lm(formula = ln_food ~ ln_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 ***
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

> summary(model_linear)$r.squared      # 原本線性模型 R²
[1] 0.0422812
> summary(model_loglog)$r.squared      # log-log 模型 R²
[1] 0.03322915
```




⇒ The log-log model shows a clearer linear pattern in the plot, but has a slightly lower $R^2$ (0.033 v.s. 0.042). So, the linear model fits slightly better statistically, but the log-log model shows a more well-defined relationship.

**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.

The elasticity from the log-log model is 0.1863 with a 95% CI of [0.1293, 0.2433]. In part (d), the elasticity estimates

**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.