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:

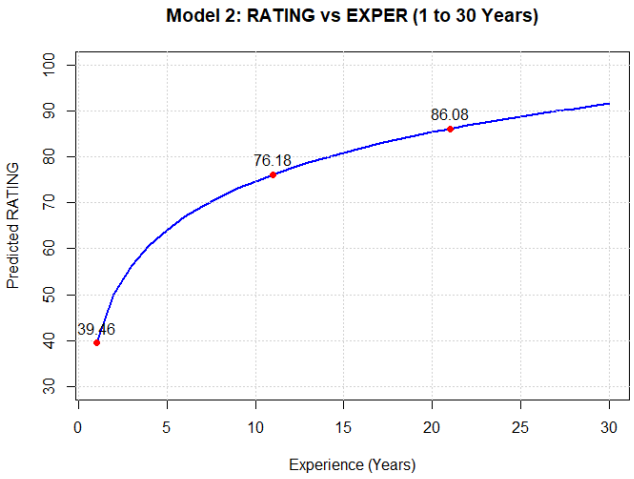
$$\widehat{RATING} = 64.289 + 0.990EXPER \quad N = 50 \quad R^2 = 0.3793$$
$$(se) \qquad (2.422) \quad (0.183)$$

Model 2:

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

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.



因為 Ln(0)無法定義故排除資料

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.

$$Marginal\ effect\ = \frac{dRATING}{dEXPER} = 0.99$$

意味著無論 10 年 20 年 marginal effect 皆為 0.99(每增加一年經驗，RATING 增加 0.99)

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.

$$Marginal\ effect\ = \frac{dRATING}{dEXPER} = \frac{15.315}{EXPER}$$

10 年經驗: marginal effect 為 1.5315

20 年經驗: marginal effect 為 0.76575

➔marginal effect 隨經驗增加而減少

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

在 EXPER 不等於 0 的藝術家中 Rsquare=0.4858, 樣本數 46，相比同樣樣本 46 的模型二，Rsquare=0.6414，因 Rsquare 較大，可見模型二有更好的解釋能力

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

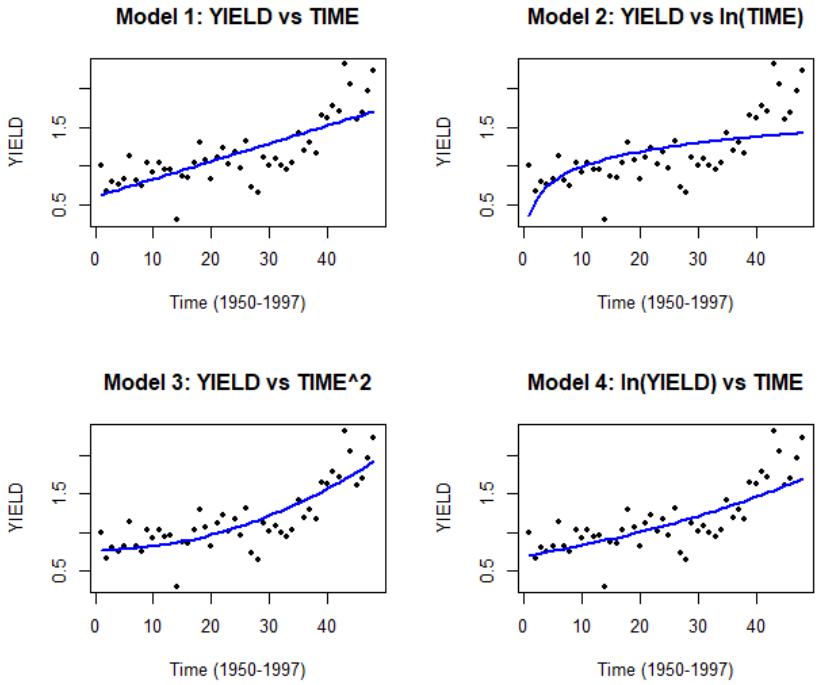
Model1 較合理，其邊際效應遞減符合經濟邏輯，對於高經驗者，多一年經驗所能帶來的提升會較小

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:

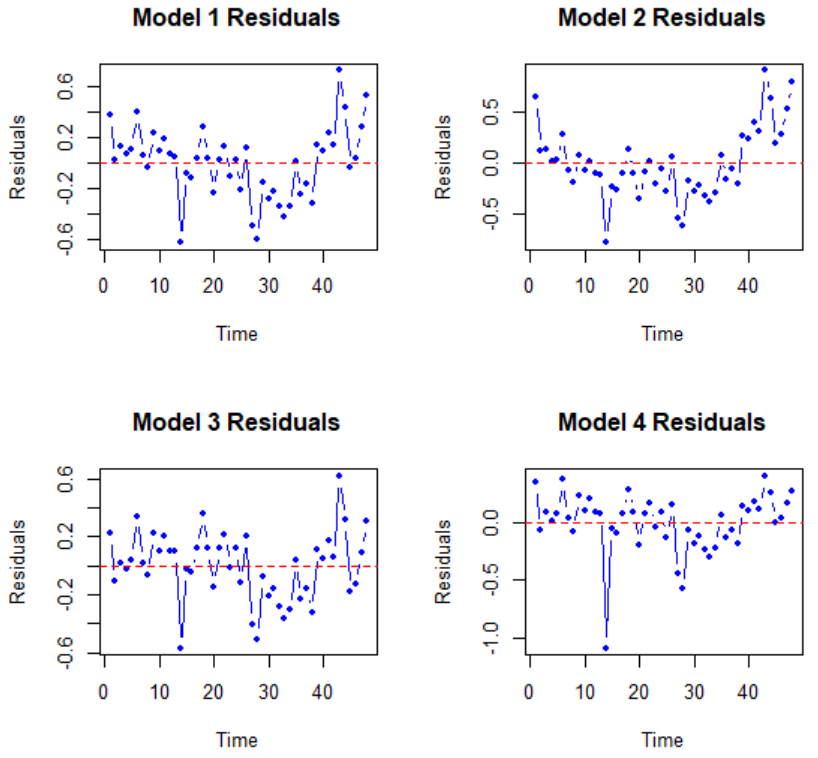
$$YIELD_t = \beta_0 + \beta_1 TIME + e_t$$
$$YIELD_t = \alpha_0 + \alpha_1 \ln(TIME) + e_t$$
$$YIELD_t = \gamma_0 + \gamma_1 TIME^2 + e_t$$
$$\ln(YIELD_t) = \phi_0 + \phi_1 TIME + e_t$$

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)


Model 1: YIELD vs TIME
YIELD
0.5
1.5
0
10
20
30
40
Time (1950-1997)
Model 2: YIELD vs ln(TIME)
YIELD
0.5
1.5
0
10
20
30
40
Time (1950-1997)
Model 3: YIELD vs TIME^2
YIELD
0.5
1.5
0
10
20
30
40
Time (1950-1997)
Model 4: ln(YIELD) vs TIME
YIELD
0.5
1.5
0
10
20
30
40
Time (1950-1997)

(ii)


Model 1 Residuals
Residuals
-0.6
-0.2
0.2
0.6
0
10
20
30
40
Time
Model 2 Residuals
Residuals
-0.5
0.0
0.5
0
10
20
30
40
Time
Model 3 Residuals
Residuals
-0.6
-0.2
0.2
0.6
0
10
20
30
40
Time
Model 4 Residuals
Residuals
-1.0
-0.5
0.0
0
10
20
30
40
Time

(iii)

```
> cat("Shapiro-Wilk Normality Test for Residuals:\n")
Shapiro-Wilk Normality Test for Residuals:
> cat("Model 1: p-value =", shapiro.test(residuals(model1))$p.value, "\n")
Model 1: p-value = 0.6792056
> cat("Model 2: p-value =", shapiro.test(residuals(model2))$p.value, "\n")
Model 2: p-value = 0.1855502
> cat("Model 3: p-value =", shapiro.test(residuals(model3))$p.value, "\n")
Model 3: p-value = 0.826645
> cat("Model 4: p-value =", shapiro.test(residuals(model4))$p.value, "\n")
Model 4: p-value = 7.205319e-05
>
```

```
R^2 Values:
> cat("Model 1:", r2_model1, "\n")
Model 1: 0.5778369
> cat("Model 2:", r2_model2, "\n")
Model 2: 0.3385733
> cat("Model 3:", r2_model3, "\n")
Model 3: 0.6890101
> cat("Model 4:", r2_model4, "\n")
Model 4: 0.5073566
>
```

選擇 model 3，其 normality test P>0.05，且 Rsquare 最高最能解釋數據，且從圖形來看也是最符合數據型態

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

```
> #b
> beta1 <- coef(model3)["I(wa_wheat$time^2)"]
> cat("Coefficient of TIME^2 in Model 3:", beta1, "\n")
Coefficient of TIME^2 in Model 3: 0.0004986181
> cat("Interpretation: The rate of change of YIELD with respect to TIME is 2 *", beta1, "* TIME units per yea
r.\n")
Interpretation: The rate of change of YIELD with respect to TIME is 2 * 0.0004986181 * TIME units per year.
> cat("For example, at TIME = 48 (1997), YIELD increases by approximately", 2 * beta1 * 48, "units per addition
year.\n")
For example, at TIME = 48 (1997), YIELD increases by approximately 0.04786734 units per additional year.
```

(解釋寫於程式檔輸出中)

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

根據 studentized residuals 判斷 unusual observations

```
    Time Year Studentized_Residual
14    14 1963            -2.560682
28    28 1977            -2.246847
43    43 1992             2.889447
```

根據其他三項判斷 unusual observations

```
High Leverage Points:
> print(data.frame(Time = wa_wheat$time[high_leverage], Y
leverage], Leverage = leverage[high_leverage]))
   Time Year   Leverage
45   45 1994 0.08542511
46   46 1995 0.09531255
47   47 1996 0.10614453
48   48 1997 0.11796846
> cat("\nHigh DFBETAS Points (for TIME coefficient):\n")

High DFBETAS Points (for TIME coefficient):
> print(data.frame(Time = wa_wheat$time[high_dfbetas], Ye
fbetas], DFBETAS_TIME = dfbetas[high_dfbetas, "I(wa_wheat
   Time Year DFBETAS_TIME
14   14 1963    0.3205200
43   43 1992    0.6521798
44   44 1993    0.3383169
48   48 1997    0.4607666
> cat("\nHigh DFFITS Points:\n")

High DFFITS Points:
> print(data.frame(Time = wa_wheat$time[high_dffits], Yea
fits], DFFITS = dffits[high_dffits]))
   Time Year     DFFITS
6     6 1955  0.3238234
14   14 1963 -0.4944002
28   28 1977 -0.3277591
43   43 1992  0.7823199
44   44 1993  0.3966661
48   48 1997  0.5077802
```

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?

```
Prediction for 1997 (TIME = 48):
> print(pred[48,])
     fit      lwr      upr
1.922482 1.412563 2.432401
>
> # 提取 1997 年的真實 YIELD
> true_yield_1997 <- wa_wheat$northampton[wa_wheat$time == 48]
> cat("True YIELD in 1997:", true_yield_1997, "\n")
True YIELD in 1997: 2.2318
>
> # 檢查真實值是否在預測區間內
> in_interval <- true_yield_1997 >= pred[48, "lwr"] && true_yield_1997 <= pred[48, "upr"]
> cat("Does the prediction interval contain the true value?", in_interval, "\n")
Does the prediction interval contain the true value? TRUE
```

真實值落在區間中

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.

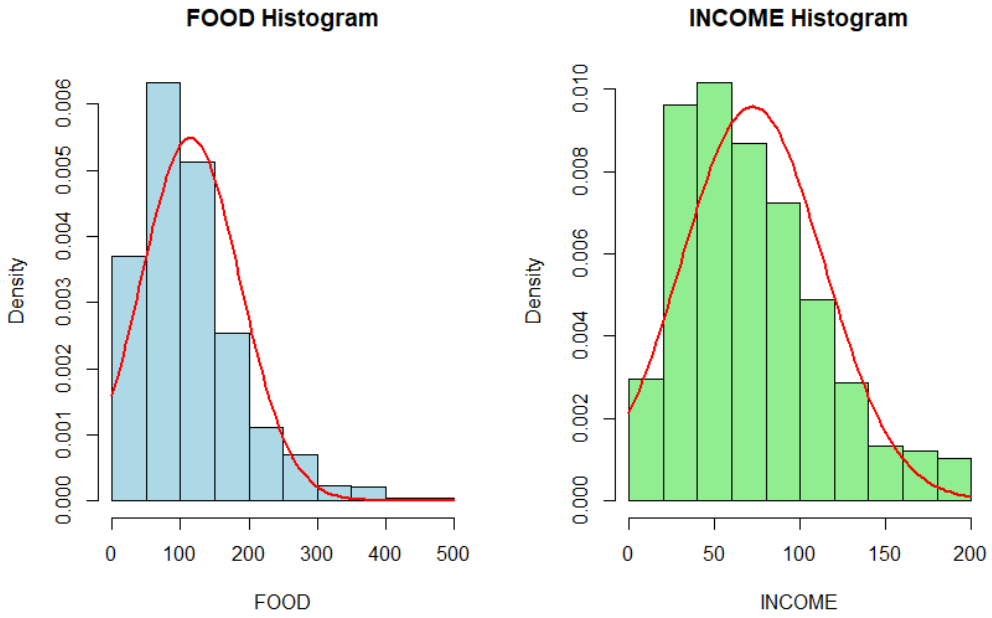
a. Calculate summary statistics for the variables: *FOOD* and *INCOME*. Report for each the sample mean, median, minimum, maximum, and standard deviation. Construct histograms for both variables. Locate the variable mean and median on each histogram. Are the histograms symmetrical and "bell-shaped" curves? Is the sample mean larger than the median, or vice versa? Carry out the Jarque–Bera test for the normality of each variable.

Food summary:

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

Income summary:

```
    mean    median       min       max        sd
72.14264  65.29000  10.00000 200.00000  41.65228
```



兩者並非對稱的鐘形分布;樣本 mean 皆大於樣本 median

Jarque-Bera test:

```
[1] "FOOD Jarque-Bera 檢驗:"
> print(food_jb)

        Jarque-Bera Normality Test

data:  cex5_small$food
JB = 648.65, p-value < 2.2e-16
alternative hypothesis: greater

> print("INCOME Jarque-Bera 檢驗:")
[1] "INCOME Jarque-Bera 檢驗:"
> print(income_jb)

        Jarque-Bera Normality Test

data:  cex5_small$income
JB = 148.21, p-value < 2.2e-16
alternative hypothesis: greater
```

➔皆拒絕符合常態分佈的假設

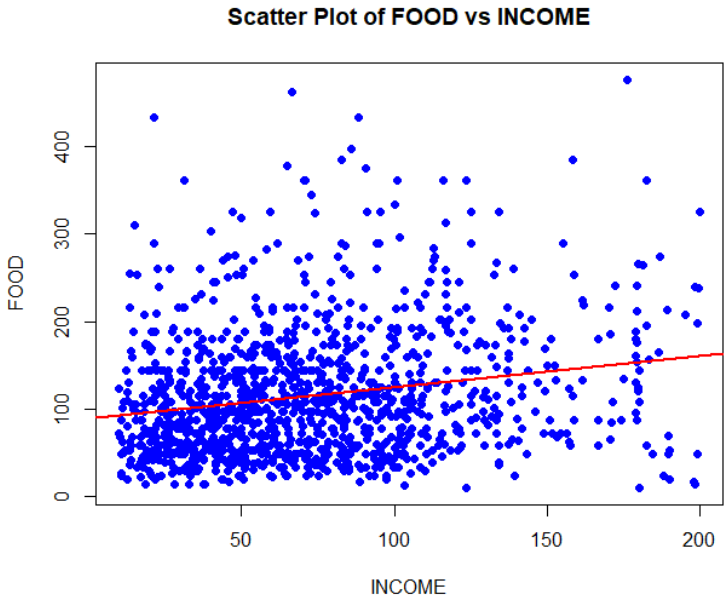
b. Estimate the linear relationship $FOOD = \beta_1 + \beta_2 INCOME + e$. Create a scatter plot *FOOD* versus *INCOME* and include the fitted least squares line. Construct a 95% interval estimate for β_2. Have we estimated the effect of changing income on average *FOOD* relatively precisely, or not?

```
Call:
lm(formula = cex5_small$food ~ cex5_small$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 ***
cex5_small$income  0.35869    0.04932   7.272 6.36e-13 ***
```

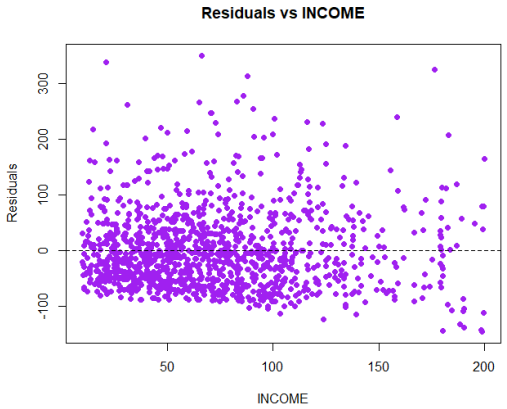
$$\beta_1 = 88.5665, \beta_2 = 0.35869$$

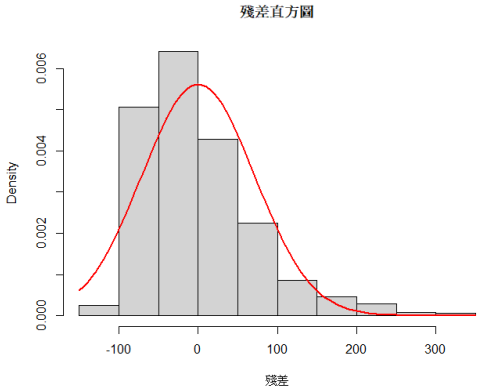


CI of β_2:

```
> # 計算 β₂ 的 95% 置信區間
> confint(model, "cex5_small$income", level = 0.95)
                          2.5 %    97.5 %
cex5_small$income 0.2619215 0.455452
```

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.





➔殘差分配有明顯的 patterns

```
        Jarque-Bera Normality Test

data:  residuals
JB = 624.19, p-value < 2.2e-16
alternative hypothesis: greater
```

➔殘差不為常態分佈

殘差為常態分布比較重要，這影響到模型是否準確

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?

```
> print("FOOD 的點估計和 95% 置信區間 (INCOME = 19.65 和 160):")
[1] "FOOD 的點估計和 95% 置信區間 (INCOME = 19.65 和 160):"
> print(predictions[19,])
     fit      lwr      upr
133.3808 126.8745 139.8872
> print(predictions[65,])
     fit      lwr      upr
112.7456 108.6908 116.8003
> print(predictions[160,])
      fit       lwr       upr
 95.89088  89.46586 102.31590
[1] "收入彈性 (INCOME = 19.65):"
> print(elasticity_19)
cex5_small$income
       0.05109466
> print("收入彈性 (INCOME = 65):")
[1] "收入彈性 (INCOME = 65):"
> print(elasticity_65)
cex5_small$income
        0.2067898
> print("收入彈性 (INCOME = 160):")
[1] "收入彈性 (INCOME = 160):"
> print(elasticity_160)
cex5_small$income
        0.5984915
```

➔income 越高彈性越高

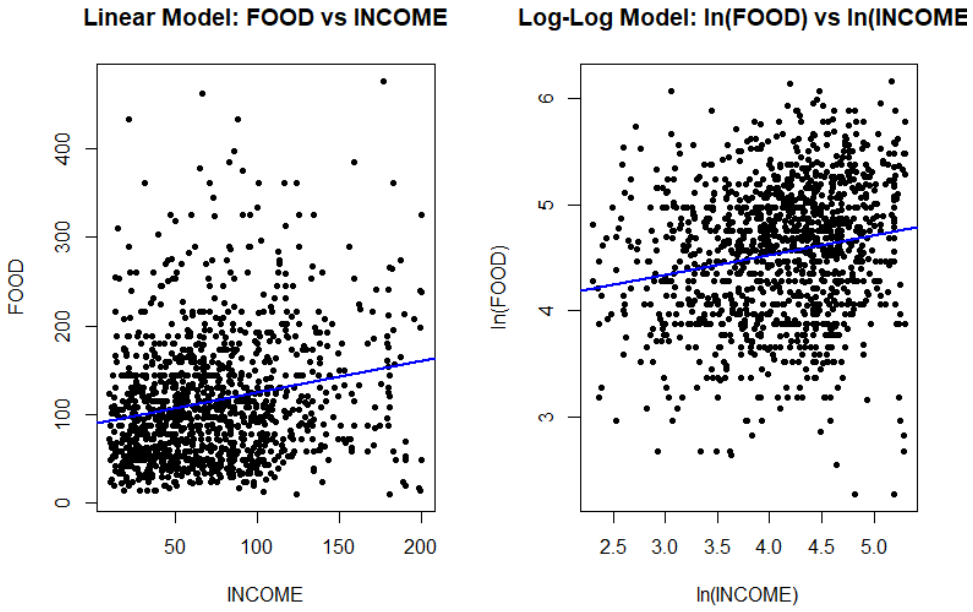
兩者區間無重疊

```
> # d. 分析收入彈性是否隨收入增加而增加或減少
> # 比較 elasticity_19.65 和 elasticity_160
> if (elasticity_160 > elasticity_19.65) {
+   print("隨著收入增加，收入彈性增加。")
+ } else {
+   print("隨著收入增加，收入彈性減少。")
+ }
[1] "隨著收入增加，收入彈性增加。"
>
```

➔隨著 income 增加，彈性增加，但根據經濟學 Engel's Law 假設，隨著收入增加，食品支出的比例減少，也就是收入彈性通常會隨著收入增加而減少，結果與經濟學不符，原因有可能是採用的模型不適合

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?

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.77893    0.12035  31.400   <2e-16 ***
log(cex5_small$income)   0.18631    0.02903   6.417    2e-10 ***
---
```



以圖形上來看 Log-Log 較好

```
> # 計算廣義 R^2
> ln_food_pred <- fitted(loglog_model)
> ln_food_actual <- log(cex5_small$food)
> generalized_r2 <- cor(ln_food_pred, ln_food_actual)^2
> cat("R^2 for Log-Log Model:", generalized_r2, "\n")
R^2 for Log-Log Model: 0.03322915
>
> # 線性模型的 R^2
> r2_linear <- summary(linear_model)$r.squared
> cat("R^2 for Linear Model:", r2_linear, "\n")
R^2 for Linear Model: 0.0422812
```

以 Rsquare 來看線性模型較好

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.

Log-log:

```
> cat("Point Estimate of Elasticity (Log-Log Model):", gamma2, "\n")
Point Estimate of Elasticity (Log-Log Model): 0.1863054
> cat("95% Confidence Interval for Elasticity:", ci_lower, "to", ci_upper, "\n")
95% Confidence Interval for Elasticity: 0.1293432 to 0.2432675
```

Linear:

```
> cat("Elasticity from Linear Model (at mean):", elasticity_linear, "\n")
Elasticity from Linear Model (at mean): 0.2261089
> cat("95% Confidence Interval for Linear Model Elasticity:", ci_lower_linear, "t
o", ci_upper_linear, "\n")
95% Confidence Interval for Linear Model Elasticity: 0.1651101 to 0.2871078
```

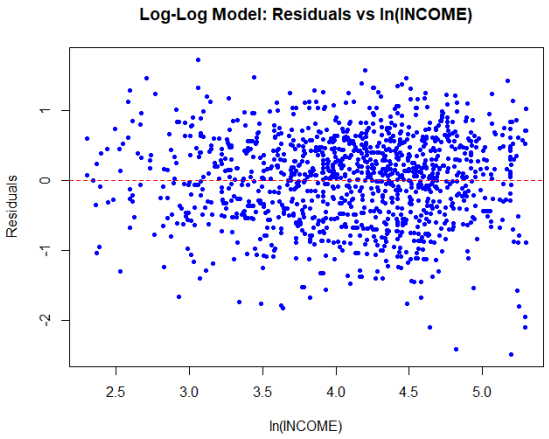
➔兩者彈性看起來差不多

t-test:

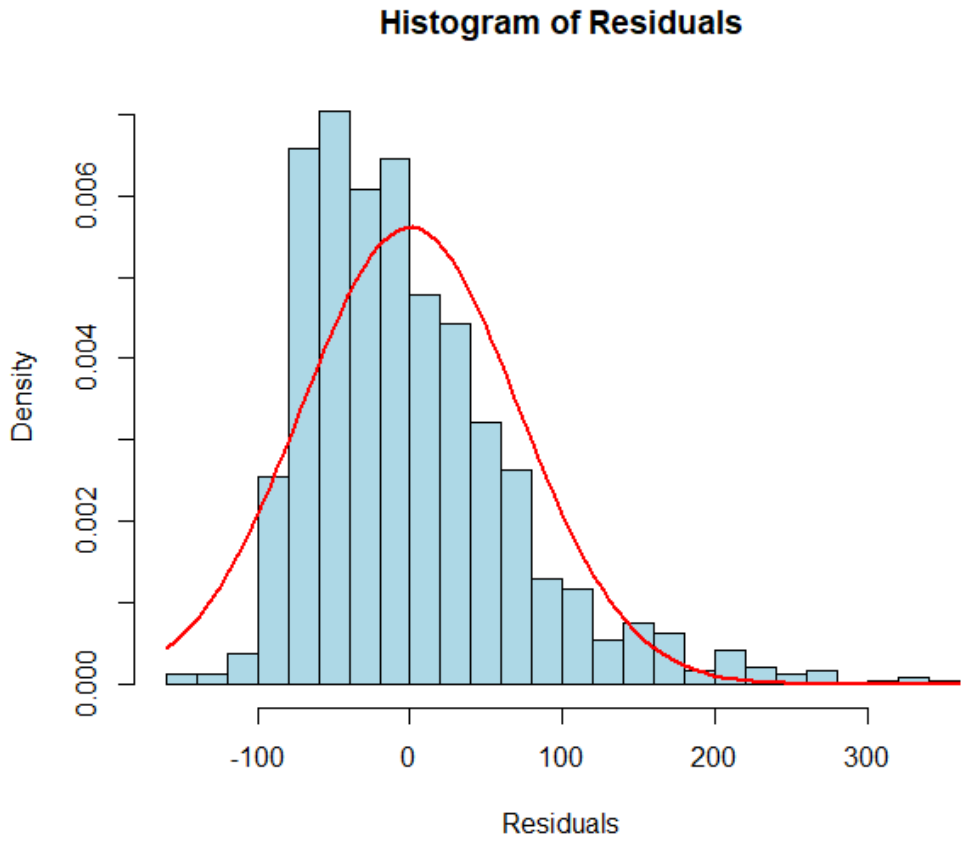
```
> t_stat <- (gamma2 - elasticity_linear) / sqrt(se_gamma2^2 + se_elasticity_linea
r^2)
> p_value <- 2 * (1 - pt(abs(t_stat), df = nrow(cex5_small) - 2))
> cat("t-statistic for difference in elasticities:", t_stat, "\n")
t-statistic for difference in elasticities: -0.935689
> cat("p-value:", p_value, "\n")
p-value: 0.3496219
> if (p_value < 0.05) {
+   cat("There's statistical evidence that the elasticities are different\n")
+ } else {
+   cat("There's no statistical evidence that the elasticities are different\n")
+ }
There's no statistical evidence that the elasticities are different
```

➔無法拒絕兩者無差異的虛無假設

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?



pattern:收入高較密集

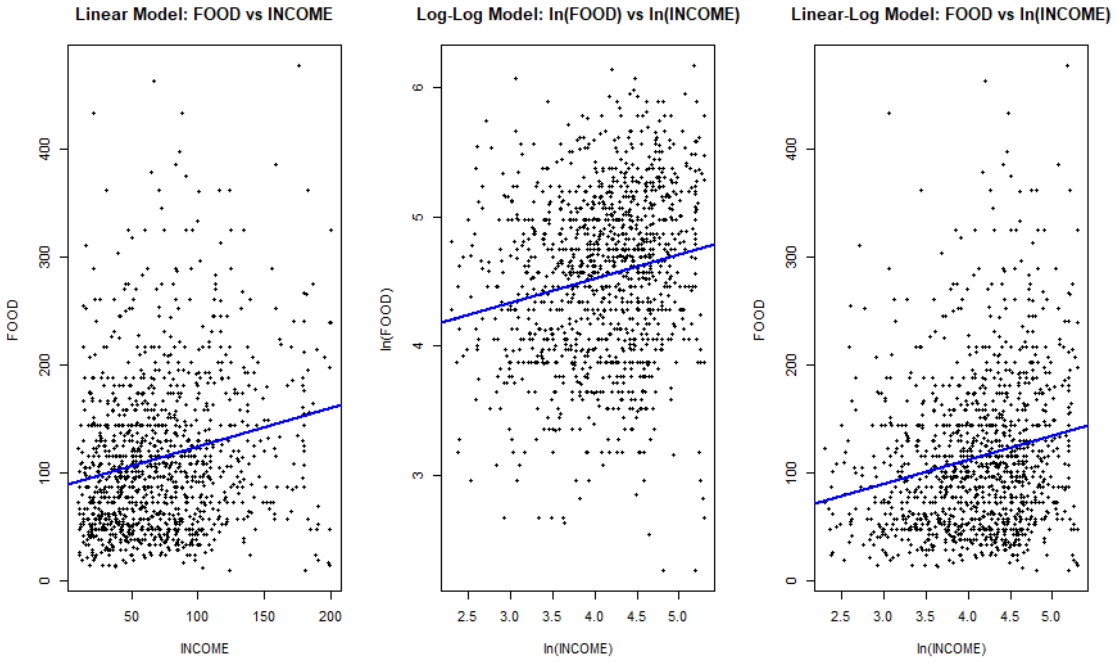


```
Jarque-Bera Test for Normality:
> cat("Test Statistic:", jb_test$statistic, "\n")
Test Statistic: 25.84998
> cat("p-value:", jb_test$p.value, "\n")
p-value: 2.436404e-06
> if (jb_test$p.value < 0.05) {
+   cat("Conclusion: Reject the null hypothesis of normality (p < 0.05). The regressio
n errors are not normally distributed.\n")
+ } else {
+   cat("Conclusion: Fail to reject the null hypothesis of normality (p >= 0.05). The
regression errors are approximately normally distributed.\n")
+ }
Conclusion: Reject the null hypothesis of normality (p < 0.05). The regression errors
are not normally distributed.
>
```

➔並非常態分佈

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?

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               23.568     13.370   1.763   0.0782 .
log(cex5_small$income)    22.187      3.225   6.879 9.68e-12 ***
```



➔Log-Log model 看起來較好

```
R^2 Comparison:
> cat("Linear Model R^2:", r2_linear, "\n")
Linear Model R^2: 0.0422812
> cat("Log-Log Model Generalized R^2:", r2_loglog, "\n")
Log-Log Model Generalized R^2: 0.03322915
> cat("Linear-Log Model R^2:", r2_linlog, "\n")
Linear-Log Model R^2: 0.03799984
```

➔以 Rsquare 來說目前仍是 Linear 的較好

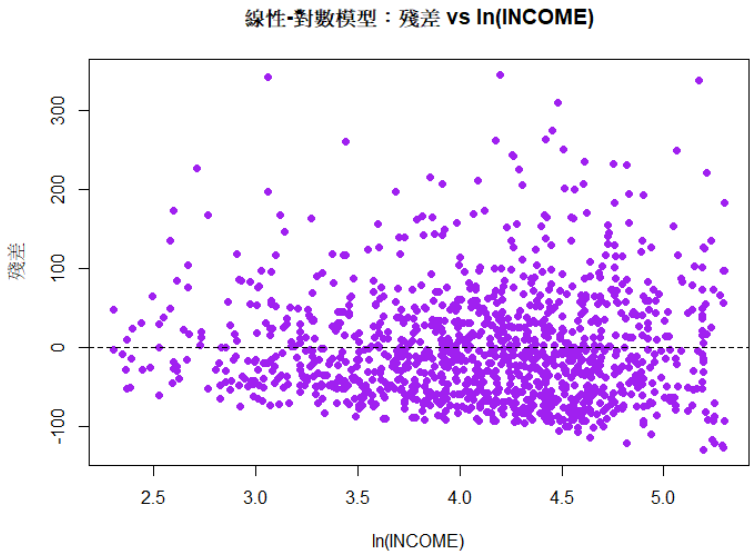
i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at *INCOME* = 19, 65, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide statistical evidence for your claim.

```
> print("FOOD 的點估計和 95% 置信區間 (INCOME = 19,65 和 160):")
[1] "FOOD 的點估計和 95% 置信區間 (INCOME = 19,65 和 160):"
> print(predictions[19,])
     fit      lwr      upr
130.6855 124.5403 136.8306
> print(predictions[65,])
     fit      lwr      upr
116.9950 112.8921 121.0979
> print(predictions[160,])
     fit      lwr      upr
90.49706 82.56320 98.43091
```

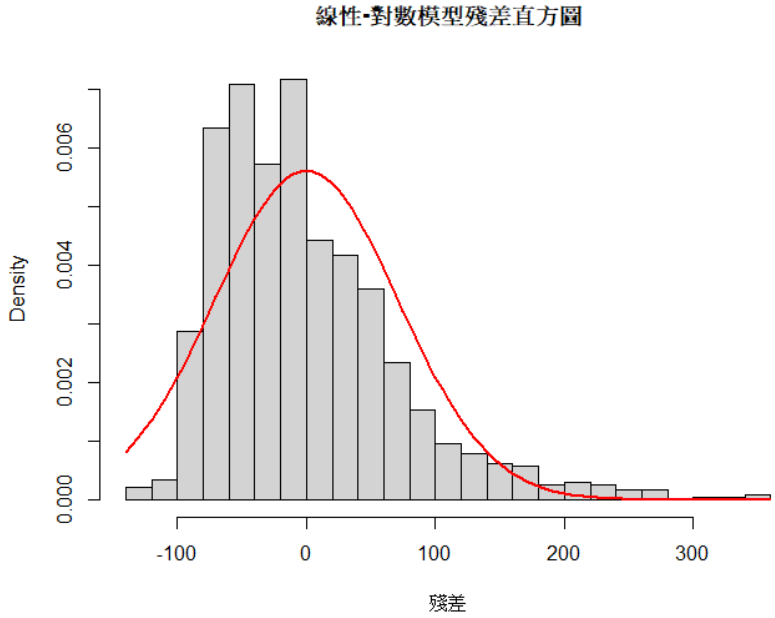
```
[1] "收入彈性 (INCOME = 19):"
> print(elasticity_19)
log(cex5_small$income)
            3.225762
> print("收入彈性 (INCOME = 65):")
[1] "收入彈性 (INCOME = 65):"
> print(elasticity_65)
log(cex5_small$income)
            12.32685
> print("收入彈性 (INCOME = 160):")
[1] "收入彈性 (INCOME = 160):"
> print(elasticity_160)
log(cex5_small$income)
            39.22759
```

t-test

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?



➔較集中於 LN(INCOME)高的地方




```
        Jarque-Bera Normality Test

data:  residuals_linlog
JB = 628.07, p-value < 2.2e-16
alternative hypothesis: greater
```

➔拒絕虛無假設(常態分佈)

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.

Prefer log-log model，因為其殘差分布表現較好