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:

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

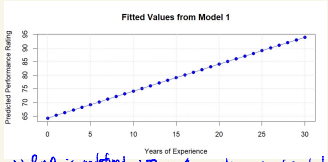
Model 2:

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

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.
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.
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.
e. Which of the two models fits the data better? Estimation of Model 1 using just the technical artists with some experience yields $R^2 = 0.4858$.
f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning? Explain.

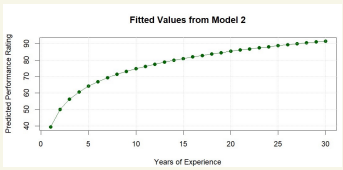
4.4

(a)



∵ ln 0 is undefined ∴ These 4 samples are not included.

(b)



(c) marginal effect (i) $\frac{\partial RATING}{\partial EXPER} = 0.990$ (ii) $\frac{\partial RATING}{\partial EXPER} = 0.990$

(d) ME (i) $\frac{\partial RATING}{\partial EXPER} = \frac{15.312}{EXPER}_{=10} = 1.5312$ (ii) $\frac{15.312}{EXPER_{=20}} = 0.7656$

(e) $R^2_{Model_2} = 0.6414 > R^2_{Model_1} = 0.3793$

⇒ Model 2 is better to capture the relationship between EXPER and RATING. (more explanatory power)

(f) Model 2 is more plausible. ∵ It suggests the "diminishing marginal effect" over time, which is more reasonable of learning curve.

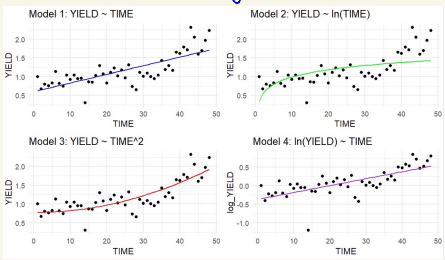
4.25 The file *wa-wheat.dat* contains observations on wheat yield in Western Australian shires. There are 48 annual observations for the years 1950–1997. For the Northampton shire, consider the following four equations:

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

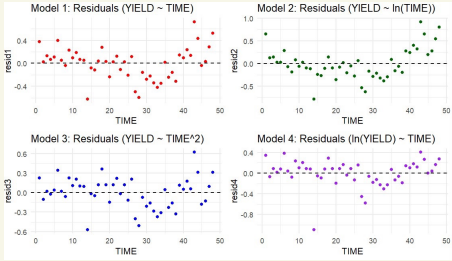
a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.
b. Interpret the coefficient of the time-related variable in your chosen specification.
c. Using your chosen specification, identify any unusual observations, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*.
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?

4.25

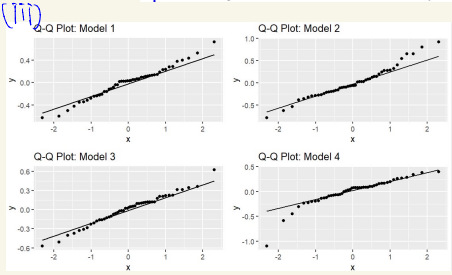
(a) (i)



(ii)



(iii)



(iv)

	Model	R2_Value
1	Linear (YIELD ~ TIME)	0.5778369
2	Log-Time (YIELD ~ ln(TIME))	0.3385733
3	Quadratic-Time (YIELD ~ TIME^2)	0.6890101
4	Log-Yield (ln(YIELD) ~ TIME)	0.5073566

Model 3 is the most preferable.
∵ The residuals plots are close to Normal Distribution and the R-square is the largest.

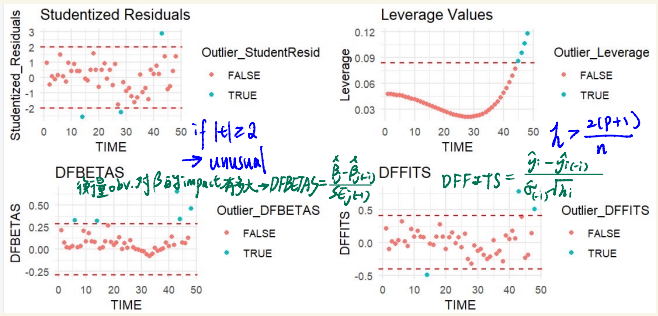
(b) Model 3: $YIELD = \gamma_0 + \gamma_1 TIME^2$

It means that as time goes by, the yield changed will be accelerated. It's reasonable since the agriculture products are usually increased non-linearly with technology improved or climate changed and so on.

```
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 ***
```

→ it's stastically significant.

(c)



if $|t| \geq 2$ → unusual

$h > \frac{2(p+1)}{n}$

衡量obv.对β的impact有多大 → $DFBETAS = \frac{\hat{\beta} - \hat{\beta}_{(-i)}}{\hat{\sigma}_{(-i)}}$

$DFFITS = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\hat{\sigma}_{(-i)}\sqrt{h_i}}$

The blue dots are unusual obvs.

(d)

```
> cat("redicted YIELD (1997):", round(pred[1], 4), "\n")
redicted YIELD (1997): 2.0079
> cat("95% Prediction Interval: [", round(pred[2], 4), ",",
round(pred[3], 4), "]\n")
95% Prediction Interval: [ 1.4994 , 2.5165 ]
> cat("Actual YIELD (1997):", round(actual_1997, 4), "\n")
Actual YIELD (1997): 2.2318
```

<note> 95% prediction interval

$= \hat{Y}_0 \pm t_{\alpha/2, n-p-1} \cdot S \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

要预测的x的坐标

The actual yield 2.2318 is in [1.4994, 2.5165], thus the interval contains the true value.

4.29

V4.29 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.

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?

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.

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. 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?

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.

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?

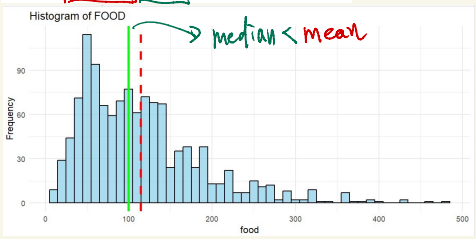
i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide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

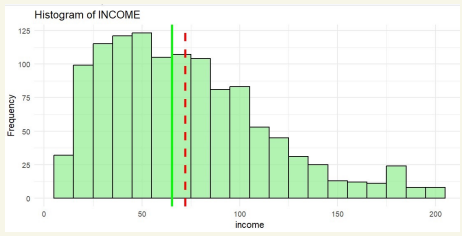
k. Based on this exercise, do you prefer the linear relationship model, or the log-log model or the linear-log model? Explain your reasoning.

(a)

	變數	平均數	中位數	最小值	最大值	標準差
1	food	114.44311	99.80	9.63	476.67	72.65750
2	income	72.14264	65.29	10.00	200.00	41.65228



median < mean



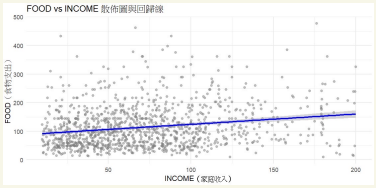
The histogram are not symmetrical and bell-shaped. Instead, they are skewed to the right.
(sample mean > median)

```
        Jarque Bera Test

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

```
        Jarque Bera Test

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

→ The Jarque Bera Test shows that χ^2 = 148.2 (income), 648.65 (food)
∴ reject H_0 → food, income are not normally distributed.

(b)

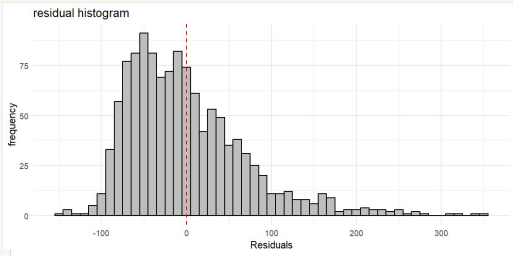


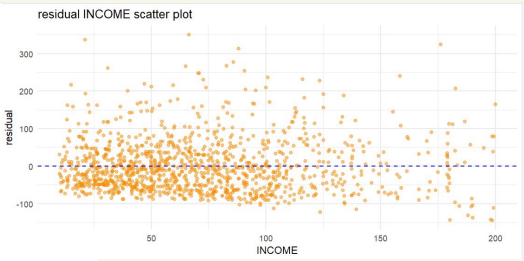

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.56650    4.10819  21.559  < 2e-16 ***
income       0.35869    0.04932   7.272 6.36e-13 ***
```

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

The interval of income = [0.3619, 0.4555] is relatively narrow.
→ precise ; not include 0 → significant.

(c)






```
        Jarque Bera Test

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

→ it indicates that it is significantly not normal distributed.

The residual (e) is more important than normal distribution of Income and Food.
normal distribution of → it's important because it might make t, p-value, C.I. unreliable.
↳ it can be adjusted by model.

(d)

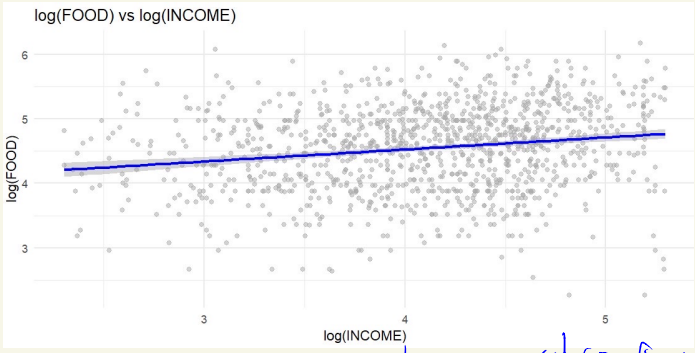
	INCOME	Fitted_FOOD	Elasticity	Lower_95_CI	Upper_95_CI
1	19	95.38	0.071	0.052	0.091
2	65	111.88	0.208	0.152	0.265
3	160	145.96	0.393	0.287	0.499

$$\varepsilon = \beta_2 \cdot \frac{INCOME}{FOOD} = \frac{\beta_2 INCOME}{\beta_1 + \beta_2 INCOME}$$

→ 結果顯示 income↑, ε↑, 是因為用線性模型, 應該用 log-log 會較正確.
→ the estimate elasticity at different levels are dissimilar.
Additionally, the C.I.s are not fully overlap.

Based on economy principle, as income↑, the ε of income-food should ↓, because food is typically considered as necessity.

(e)



It shows a clearer relationship compared to figure in (b).

	Model	R2	Generalized_R2
1	Linear	0.04228120 → better fit.	NA
2	Log-Log	0.03322915	0.01403311

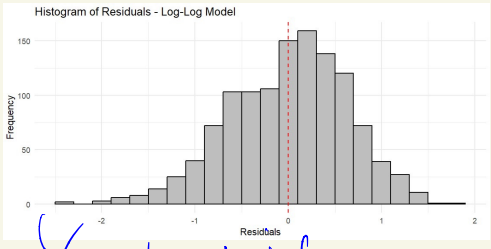
R^2_{linear} is still greater than $R^2_{Log\text{-}Log}$, but Log-Log model is still attractive due to reducing the outliers. (and better residual distribution)

(f)

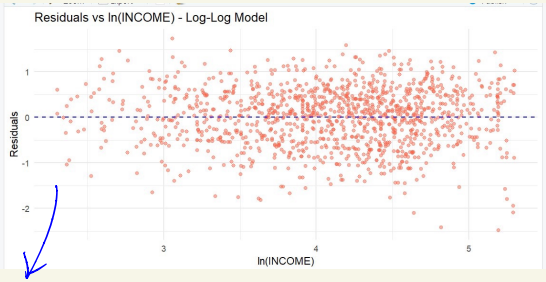
	2.5 %	97.5 %
(Intercept)	3.5428135	4.0150507
log(income)	0.1293432	0.2432675

Dissimilar. It's a constant value.

(g)



skew to the left



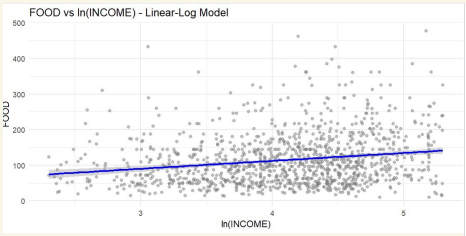
No obvious pattern.

```
        Jarque Bera Test

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

p-value < 0.05

→ the residual is not normally distributed.

(h)




```
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 **
```

	Model	R2
1	Linear	0.04228120
2	Log-Log	0.03322915
3	Linear-Log	0.03799984

→ Medium explanory power.

It's R^2 < Linear model's, and scatter plot is not interpret enough. (pattern, some outliers...)

By R^2 → Linear is a better model.

By "fit" of scatter plot, Log-Log is better.

(i) $\varepsilon = \frac{\alpha_2}{\widehat{FOOD}} = \frac{\alpha_1}{\alpha_1 \ln(Income) + \alpha_1}$?

	INCOME	Predicted_FOOD	Elasticity	Lower_95_CI	Upper_95_CI
1	19	88.90	0.2496	0.1785	0.3207
2	65	116.19	0.1910	0.1366	0.2454
3	160	136.17	0.1629	0.1165	0.2094

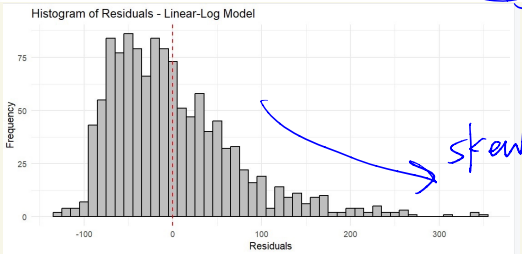
decrease

When income ↑, the ε ↓, it meets the economical intuitive.

(j)

```
        Jarque Bera Test

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

it's not Normal Distribution.



skew to the right

(K) The linear model is not reasonable because of increasing in ε.

The linear log model satisfied the economic reasoning, but the residual is far from normal distribution.

Thus, Log-Log model is a relatively better choice!