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.

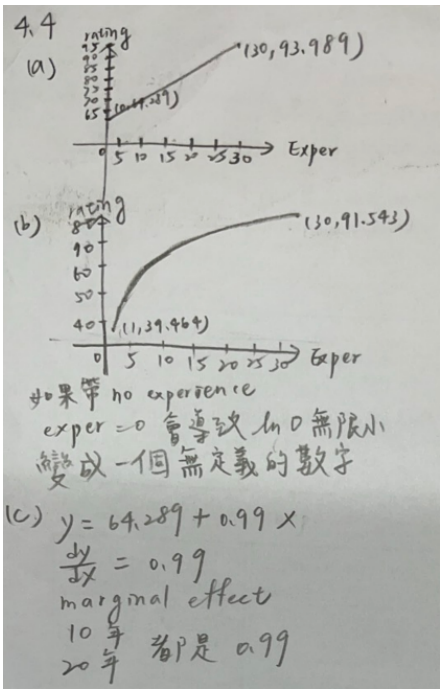
b. Sketch the fitted values from Model 2 against $EXPER = 1$ to 30 years. Explain why the four artists with no experience are not used in the estimation of Model 2.

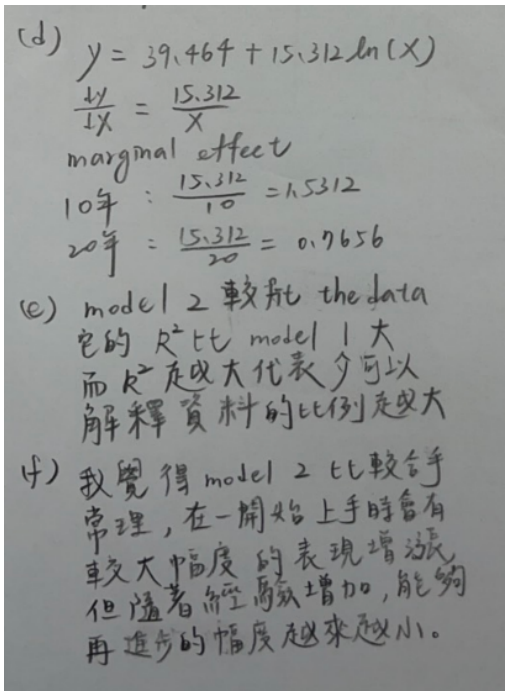
c. Using Model 1, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.

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.

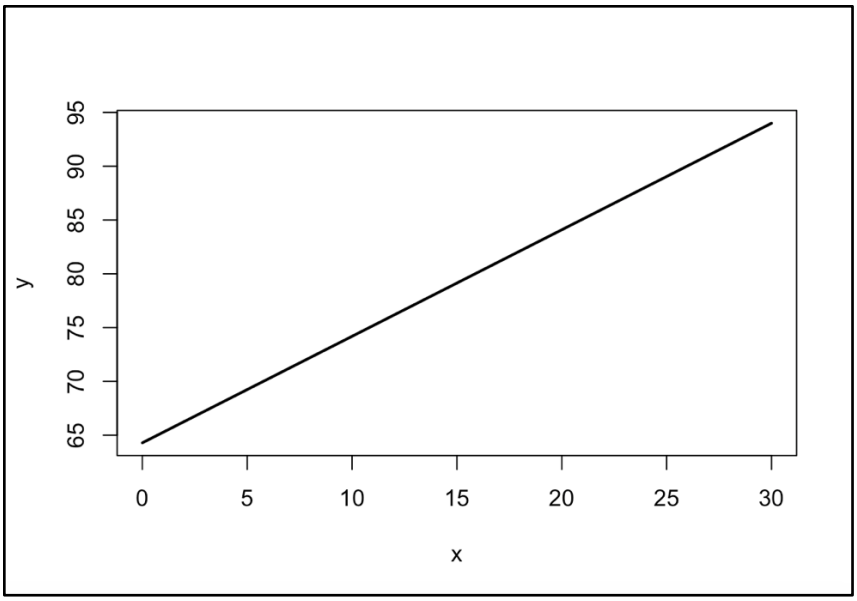
e. Which of the two models fits the data better? Estimation of Model 1 using just the technical artists with some experience yields $R^2 = 0.4858$.

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

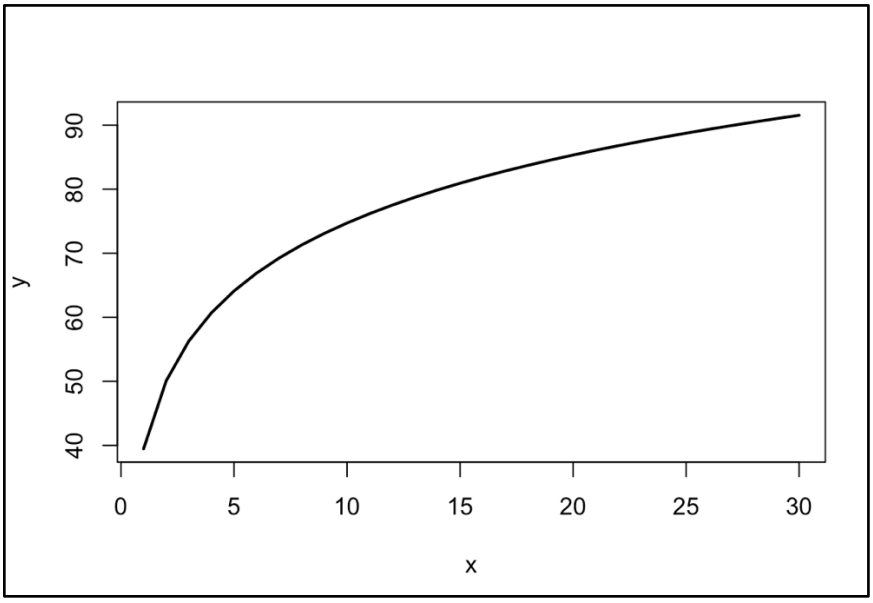




a.


y
65
70
75
80
85
90
95
0
5
10
15
20
25
30
x

b.


y
40
50
60
70
80
90
0
5
10
15
20
25
30
x

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.

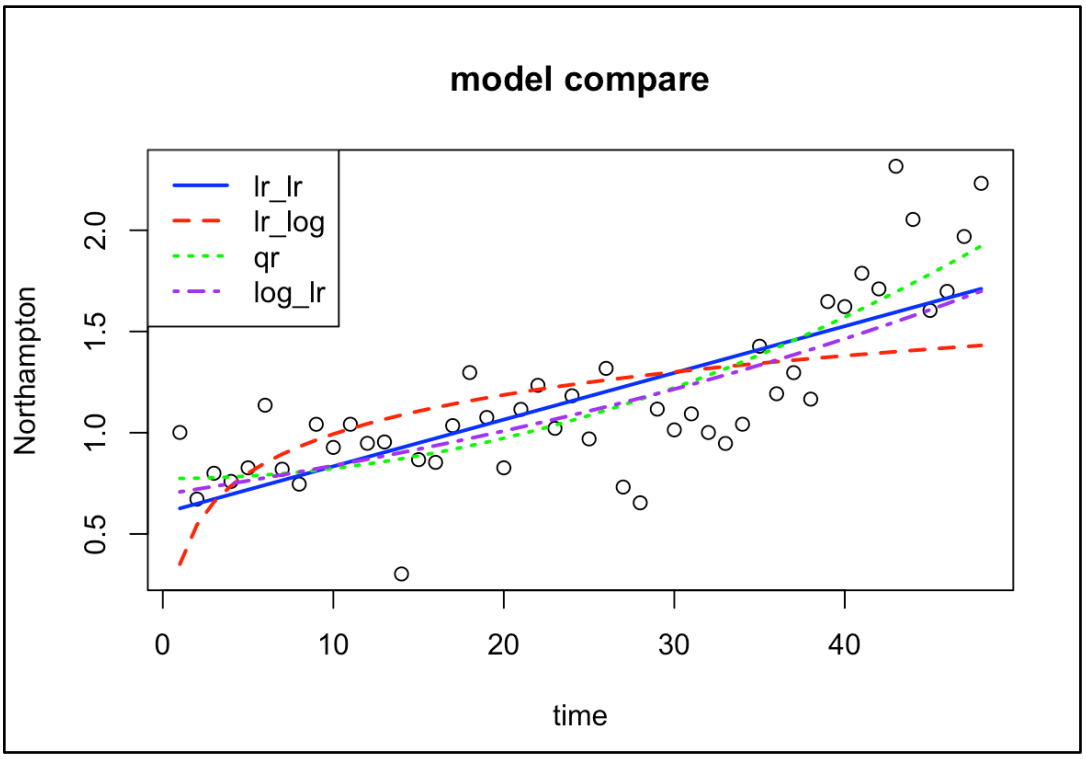
b. Interpret the coefficient of the time-related variable in your chosen specification.

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?

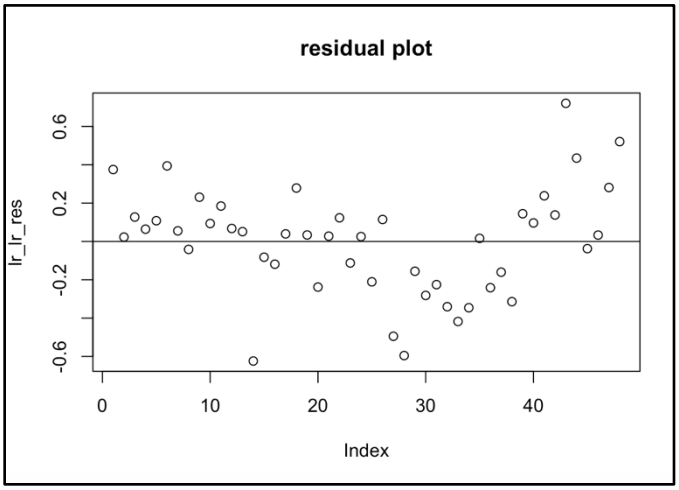
a.

(i) plots of the fitted equations



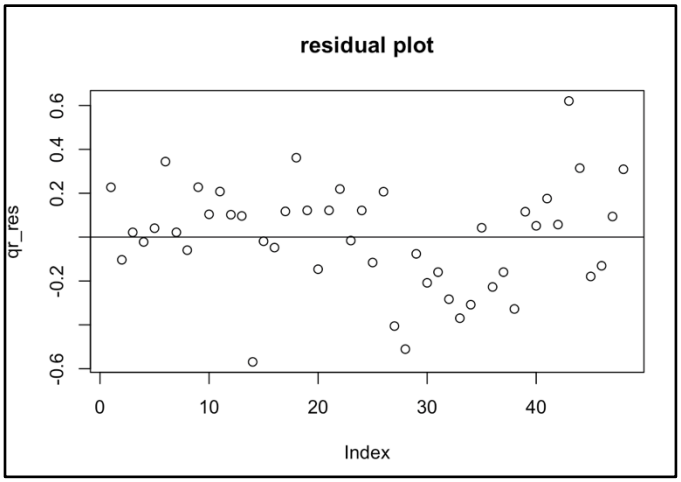
(ii) plots of the residuals

Linear-Linear model



Linear-Log model

Quadratic model



Log-Linear model

(iii) error normality tests

Model	P_Value
lr_lr	0.9358650
lr_log	0.2512080
qr	0.8504138
log_lr	0.0000000

(iii) values for R^2

Table: Multiple R-squared

Model	r_square
lr_lr	0.5778369
lr_log	0.3385733
qr	0.6890101
log_lr	0.5073566

從散佈圖觀察資料感覺用 quadratic 或是 log-linear 比較合適，接著透過殘差分析，客觀得知 quadratic 比較符合常態，加上它的 R-squared 值也不低，因此我認為使用 quadratic 較為合適。

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

選擇的 model，y 對 x 進行微分後是 2*b2*x, 代表斜率會隨著 x 上升而上升。

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

studentized residuals 異常資料

-通常超過 ±2 或 ±3，表示該點可能是異常值。

	Observation	residual_Value
14	14	-2.560682
28	28	-2.246847
43	43	2.889447

LEVERAGE 異常資料

-衡量某個觀測值對回歸模型的影響力 (若>2(k+1)/n 表示有影響力，可能為異常資料，需搭配 studentized residuals 近一步確認) k:自變數數量

	Observation	leverage_Value
45	45	0.08542511
46	46	0.09531255
47	47	0.10614453
48	48	0.11796846

DFBETAS 異常資料

-（回歸係數變動指標）測量某個觀測值 對每個回歸係數的影響

絕對值大於 2 / sqrt(n)（n 是樣本數）表示該點影響較大。

	Observation	Coefficient	DFBETAS_Value
1	6	(Intercept)	0.3237367
2	14	(Intercept)	-0.4894502
3	14	I(x^2)	0.3205200
4	43	I(x^2)	0.6521798
5	44	I(x^2)	0.3383169
6	48	I(x^2)	0.4607666

DFFITS 異常資料

-衡量某個觀測值對模型預測結果的影響力。

絕對值超過 2 * sqrt(p/n)（p 是參數數量，n 是樣本數）表示影響大。

	Observation	DFFITS_Value
6	6	0.3238234
14	14	-0.4944002
28	28	-0.3277591
43	43	0.7823199
44	44	0.3966661
48	48	0.5077802

綜合上述，第 6 筆與第 43 筆可能為異常資料。

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 interval 是 (1.372403, 2.389819)，包含實際值 2.2318

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.

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. **summary statistics for the variables: FOOD and INCOME**

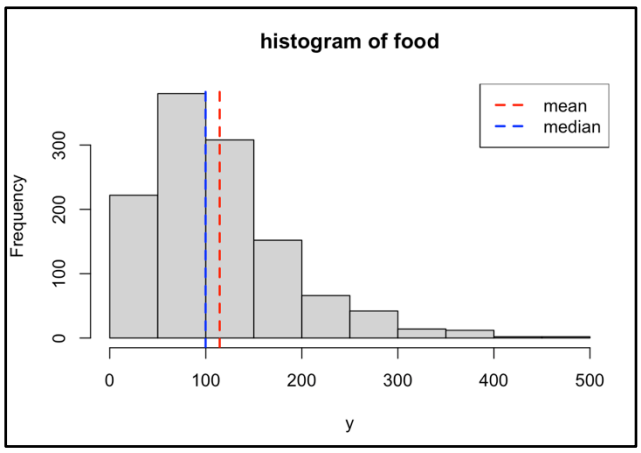
food

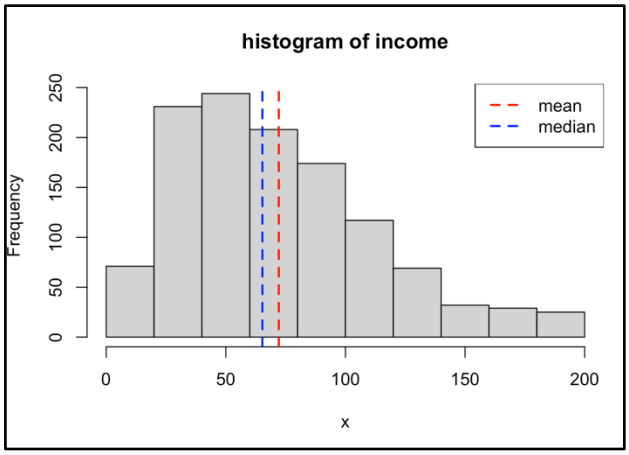
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.63	57.78	99.80	114.44	145.00	476.67

income

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.00	40.00	65.29	72.14	96.79	200.00

histograms





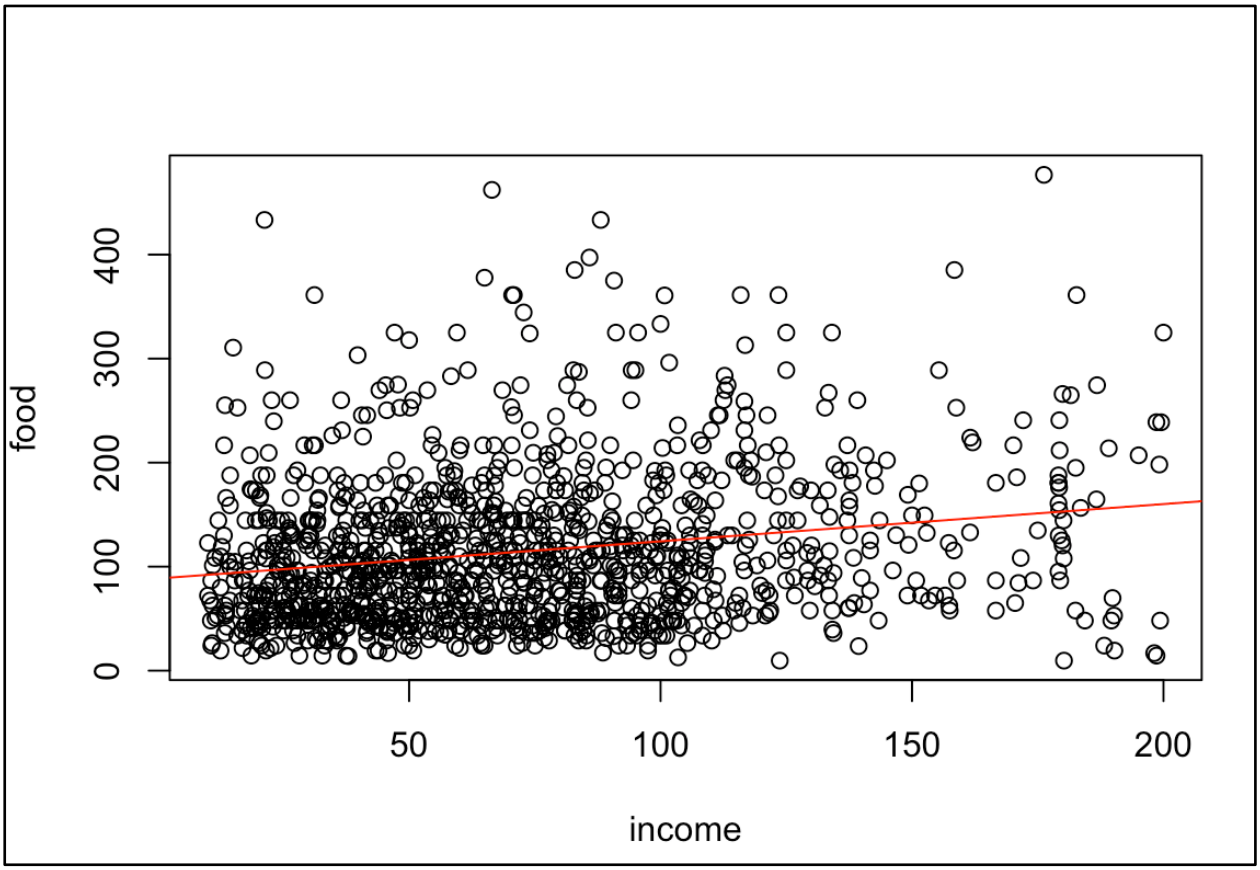
food 跟 income 資料都呈現右偏，平均數都大於中位數，也沒有左右對稱，不是鍾型分佈。

Jarque–Bera test

Model	P_Value
income	0
food	0

從上述表格可以發現 income 跟 food 資料不呈現常態分佈。

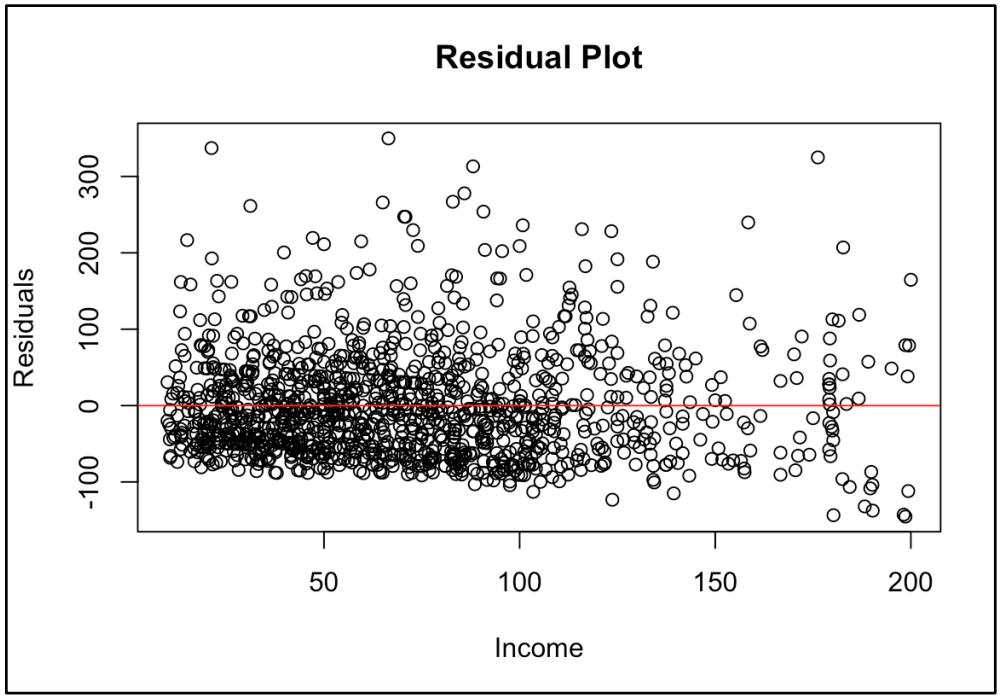
b. **scatter plot *FOOD* versus *INCOME*, Construct a 95% interval estimate for β_2**



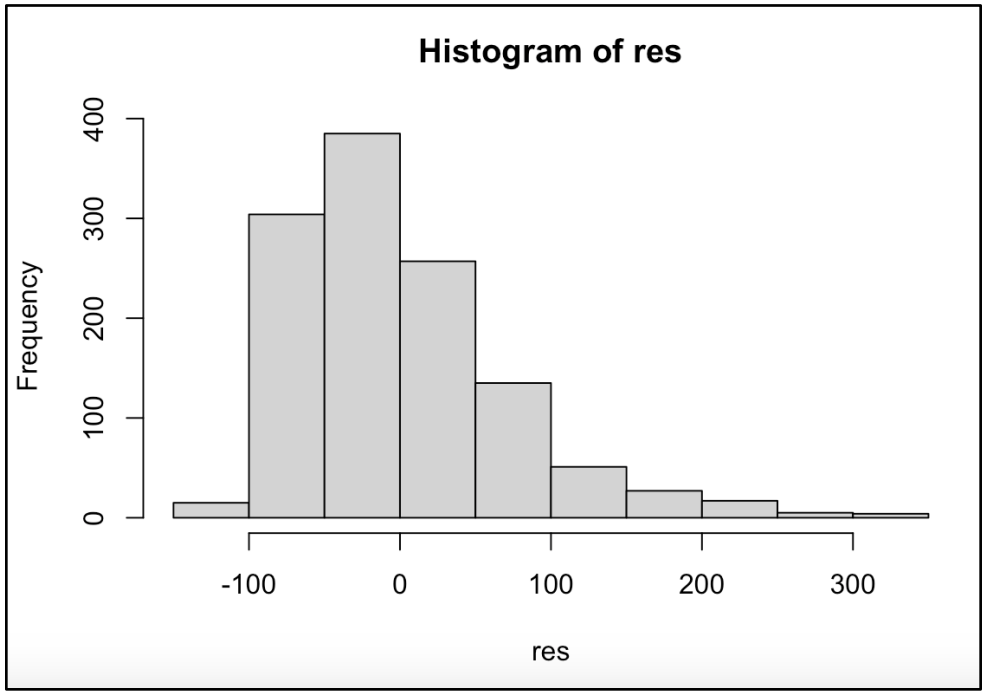
estimate for β_2: (0.2619215 0.4554520)

從圖觀察，預測線的結果跟資料沒有很適配。

c. **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?**



殘差的平均不太接近 0，而變異隨著 income 上升而上升。



```
        Jarque Bera Test

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

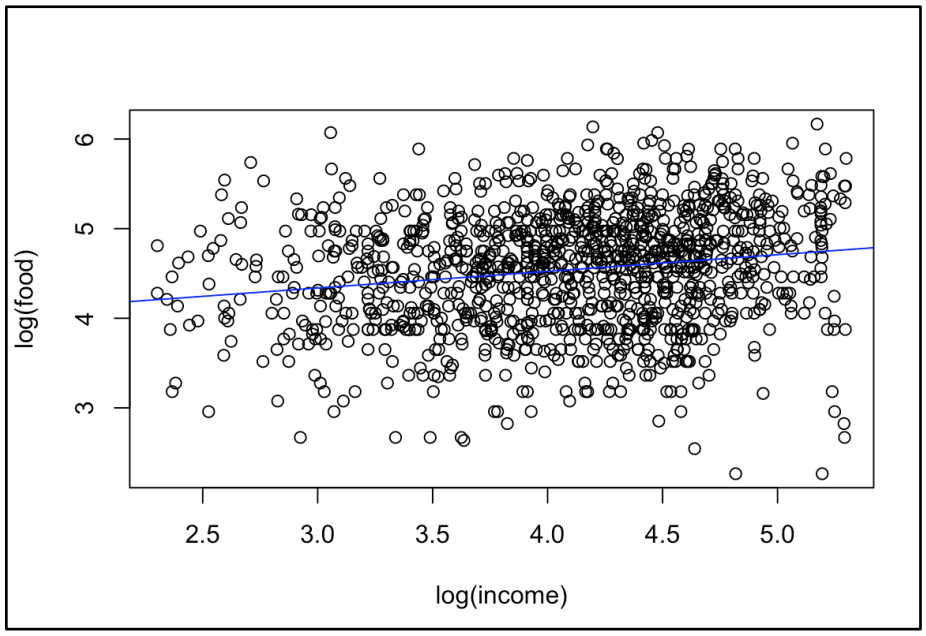
透過 Jarque Bera Test，殘差不符合常態。而比較重要的是 error 要符合常態，我們假設的是殘差符合常態，透過調整 food 跟 income 變數跟選擇合適的 model 來實現殘差常態。

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?**

Income	elas_Estimate	Lower_Bound	Upper_Bound
19	0.07145038	0.05217475	0.09072601
65	0.20838756	0.15216951	0.26460562
160	0.39319883	0.28712305	0.49927462

Interval estimate 並沒有重疊，根據經濟原理，因為食物不是奢侍品，應該隨著收入上升，花費的比例減少。也就是 elasticity 下降。

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



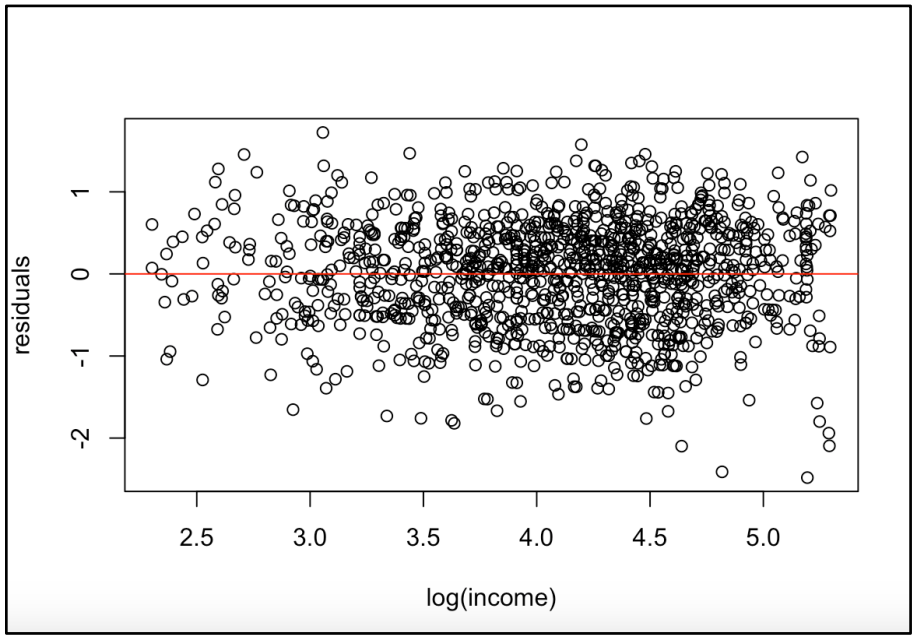
與線性模型相比，log-log 模型將後段資料轉成較集中的分佈，有比較貼近線性關係。但其 r-squared 為 0.03322915，與 linear 0.0422812 相比較小，但兩個都不盡理想。

f. **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.**

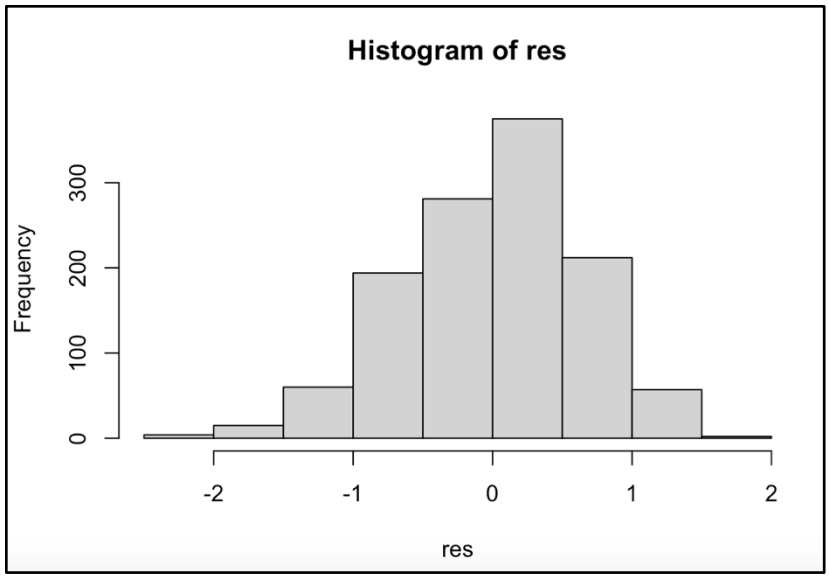
	Income	elas_Estimate	Lower_Bound	Upper_Bound
1	19	0.1863054	0.1293432	0.2432675
2	65	0.1863054	0.1293432	0.2432675
3	160	0.1863054	0.1293432	0.2432675

與 linear 相比，log-log model 的 elasticity 是固定的。

g. **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?**



從殘差圖較看不出 pattern。

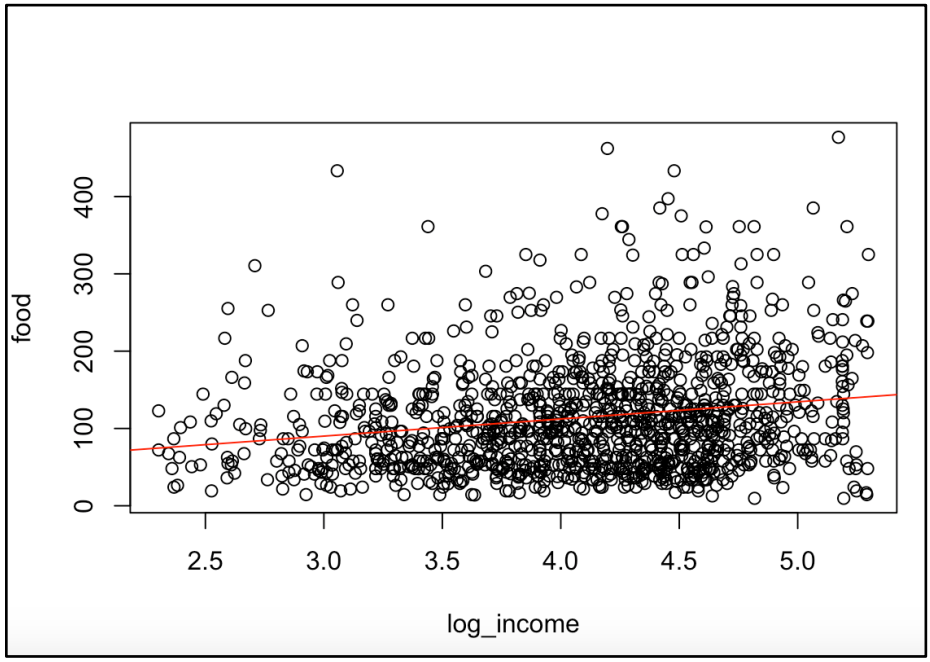



```
        Jarque Bera Test

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

Histogram 呈現殘差的分佈較 linear model 殘差常態，但透過 Jarque bera test 殘差 p-value 太小，偏向不符合常態。

h. $FOOD = \alpha_1 + \alpha_2 \ln(INCOME) + e$.



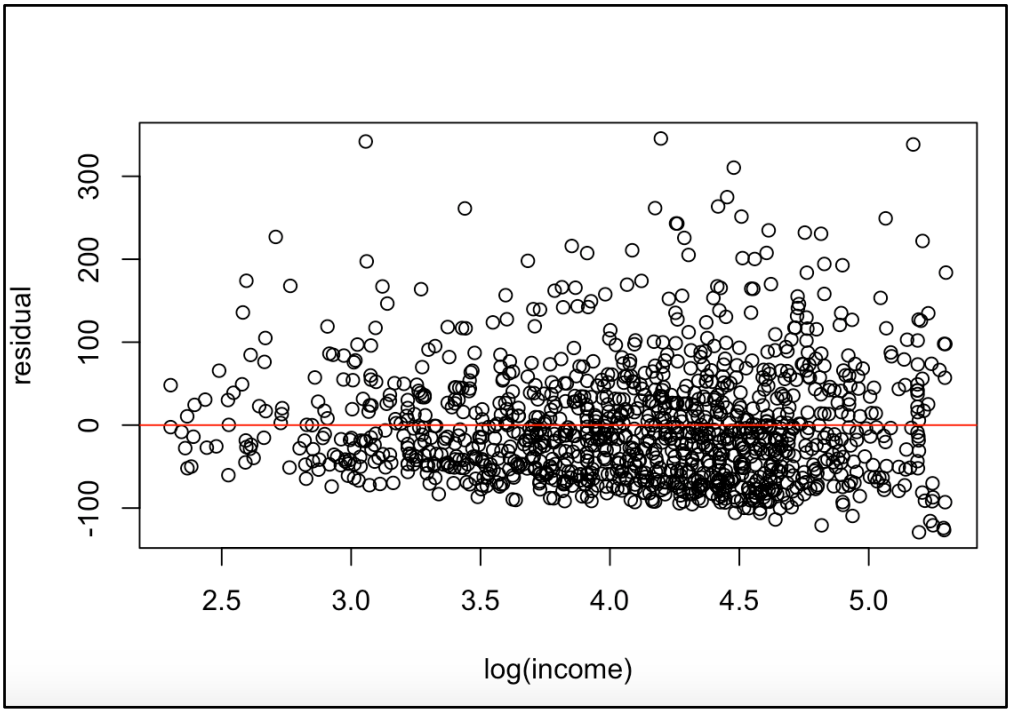
它的 R-squared 是 0.03799984，大於 log-log model，小於 linear model。線性關係也不太明顯。

i. **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.**

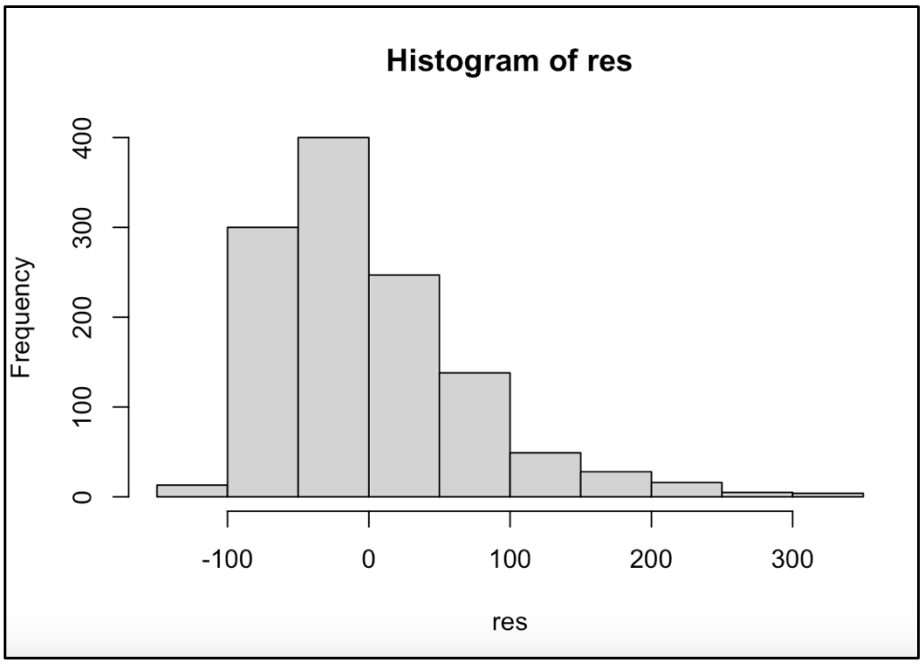
	Income	elas_Estimate	Lower_Bound	Upper_Bound
1	19	0.2495828	0.1784009	0.3207648
2	65	0.1909624	0.1364992	0.2454256
3	160	0.1629349	0.1164652	0.2094046

與其他模型相比，linear-log 的 elasticity 隨著 income 上升而下降。

j. **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?**



大多殘差值很大，殘差呈現右偏。




```
        Jarque Bera Test

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

殘差不符合常態。

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.**

我偏好 log-log model。因為 log-log model 的殘差 histogram 較其他兩者像常態分佈，雖然透過 jarque bera test 可以發現這三者都不太像是常態分佈。加上三者的 r-squared 值都很低，表示可能有其他相對於 income 更適合用來解釋 food 花費的變數，或是其他更適合的模型。