HW0317

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:

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

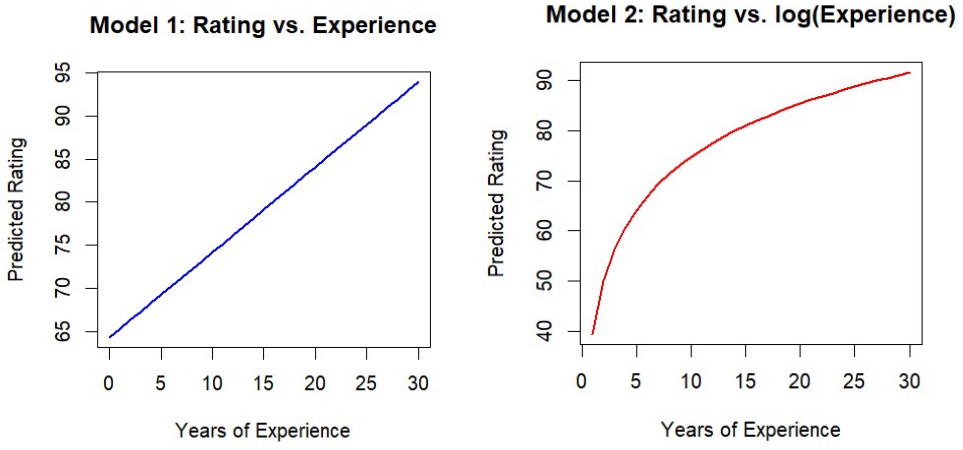
Model 2:

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

---

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.
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.



b. 由於 ln(0)未定義，因此 EXPER = 0 的藝術家無法納入此模型

c. 在 model 1 中，由於模型是線性的，因此邊際影響均相同，且等於模型斜率，即邊際影響均為 0.990。

d. Model 2 透過對 EXPER 偏微分，可知邊際影響為 15.312 / EXPER，因此當 EXPER = 10 時，邊際影響為 1.5312；EXPER = 20 時，邊際影響為 0.7656。

e. 以 R-square 來比較，model 2 的 R-square 大於只考慮有經驗的藝術家的 model 1 的 R-square，可推論 model 2 擬合的較好。

f. 我認為 model 2 較合理，因為現實中的學習曲線確實會表現出學習初期進步速度較快，而中後期進步速度較慢，較不符合 model 1 的線性表現。

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. <linear-linear>

Residuals:

Min	1Q	Median	3Q	Max
-0.62394	-0.17302	0.03342	0.12996	0.72050

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.603245	0.081858	7.369	2.55e-09	***
time	0.023078	0.002908	7.935	3.69e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2791 on 46 degrees of freedom

Multiple R-squared: 0.5778, Adjusted R-squared: 0.5687

F-statistic: 62.96 on 1 and 46 DF, p-value: 3.689e-10

<linear-log>

Residuals:

Min	1Q	Median	3Q	Max
-0.78488	-0.20711	-0.06382	0.15447	0.91573

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.3510	0.1759	1.995	0.052	.
log(time)	0.2790	0.0575	4.852	1.44e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3494 on 46 degrees of freedom
Multiple R-squared: 0.3386, Adjusted R-squared: 0.3242
F-statistic: 23.55 on 1 and 46 DF, p-value: 1.44e-05

<poly model>

Residuals:

Min	1Q	Median	3Q	Max
-0.54787	-0.13543	0.00819	0.17041	0.58048

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.16865	0.03305	35.363	< 2e-16	***
poly(time, 2)1	2.21500	0.22896	9.674	1.45e-12	***
poly(time, 2)2	1.10699	0.22896	4.835	1.59e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.229 on 45 degrees of freedom
Multiple R-squared: 0.7222, Adjusted R-squared: 0.7098
F-statistic: 58.48 on 2 and 45 DF, p-value: 3.057e-13

<log-linear>

Residuals:

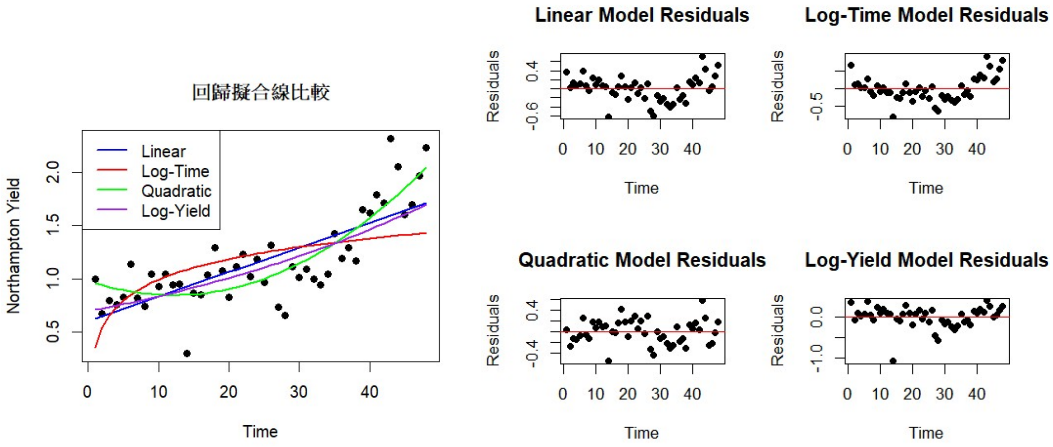
Min	1Q	Median	3Q	Max
-1.09292	-0.10049	0.07125	0.14140	0.40263

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.363938	0.076192	-4.777	1.85e-05	***
time	0.018632	0.002707	6.883	1.37e-08	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

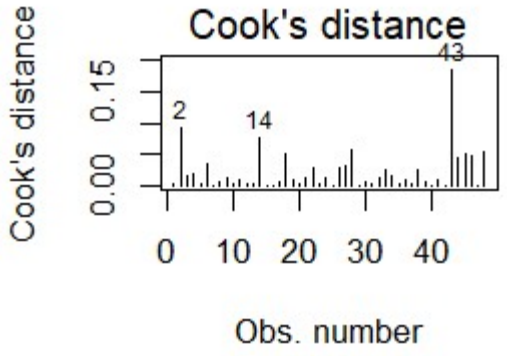
Residual standard error: 0.2598 on 46 degrees of freedom
Multiple R-squared: 0.5074, Adjusted R-squared: 0.4966
F-statistic: 47.37 on 1 and 46 DF, p-value: 1.366e-08



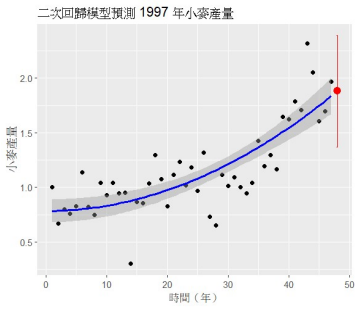
對四個模型進行 jarque bera 檢定，只有 log-linear 模型的 p-value＜0.05，表示 log-linear 的殘差不符合常態分配假設，再來以四個模型的 R-square 比較，以二次式模型有最高的 R-square，有較好的擬合度。

b. 在二次式模型中 Time 的係數表示了邊際影響非線性的加速。

c.



d. 根據二次式模型，95%信心水準下的預測區間為[1.372403, 2.389819]，而 1997 年的真實值為 2.2318，有落在預測區間內。



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?

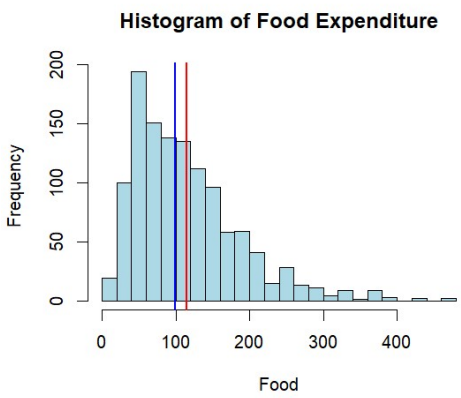
h. For expenditures on food, estimate the linear-log relationship $FOOD = \alpha_1 + \alpha_2 \ln(INCOME) + e$. Create a scatter plot for *FOOD* versus $\ln(INCOME)$ and include the fitted least squares line. Compare this to the plots in (b) and (e). Is this relationship more well-defined compared to the others? Compare the R^2 values. Which of the models seems to fit the data better?

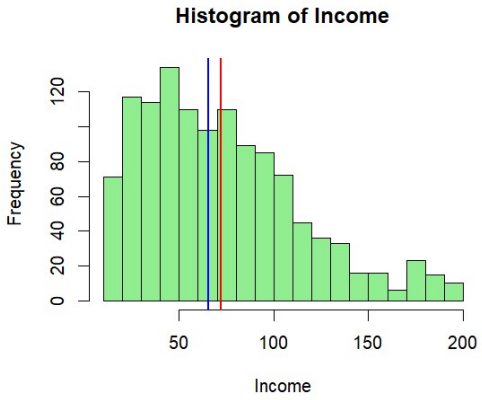
i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at *INCOME* = 19, 65, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide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






```
    Jarque Bera Test
data:  cex5_small$food
X-squared = 648.65, df = 2, p-value < 2.2e-16
```

```
    Jarque Bera Test
data:  cex5_small$income
X-squared = 148.21, df = 2, p-value < 2.2e-16
```

由於 Jarque Bera Test 兩者 p-value 都小於 0.05，因此有證據支持 Food 以及 Income 兩變數不符合常態分配。

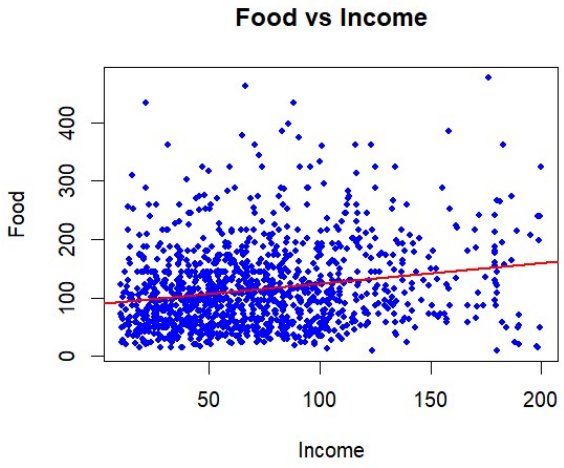
b. 迴歸估計結果 ：

```
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
```

係數在 95%信心水準下的區間估計 ：

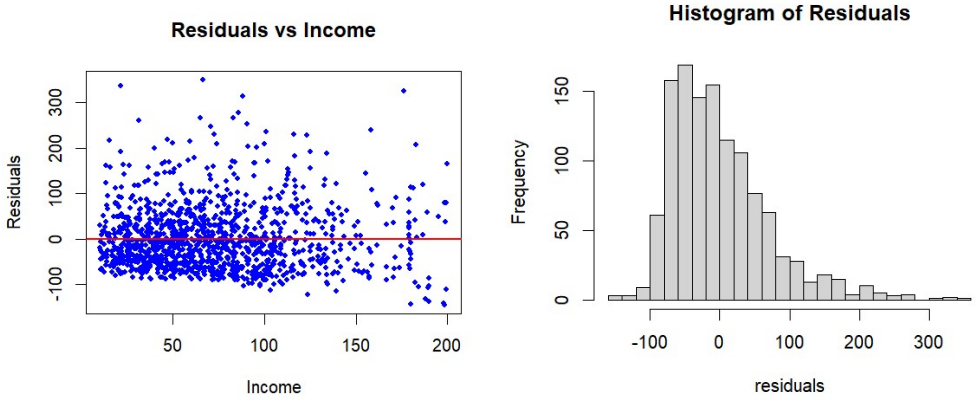
```
                2.5 %    97.5 %
income      0.2619215  0.455452
```

散佈圖以及迴歸線 ：



由散佈圖可知，隨著收入變高，食物的支出的離散程度也隨之增加，因此在面對高收入族群時，此迴歸線估計並不一定準確。

c. 殘差 v.s.收入 / 殘差直方圖 ：




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

以 Jarque Bera Test 之結果來看，並無證據支持殘差符合常態分配。

Food 以及 Income 並不需要服從常態分配也能使回歸估計具有不偏及一致性，但殘差項若不服從常態分配，會使回歸估計不具有不偏性，因此殘差項服從常態分配更為重要。

d. 彈性結果以及預測值 ：

Income = 19 65 160

Elascity = 0.0715 0.2084 0.3932

Food_hat = 95.3816 111.8811 145.9564

彈性隨著收入增加而增加，表示收入越高，食物的支出對收入的變化越敏感，反映出高收入家庭更可能購買單價更高的食物。

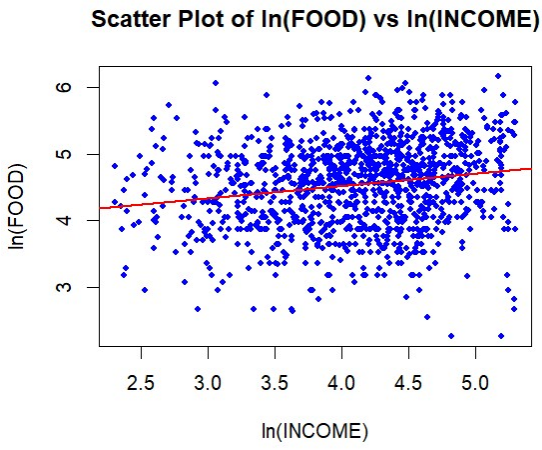
95%信心水準下彈性的區間估計 ：

Income = 19, 65, 160

CI = [0.0522, 0.0907], [0.1522, 0.2646], [0.2871, 0.4993]

e. Log-log 模型的結果、散佈圖以及迴歸線 ：

```
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
```



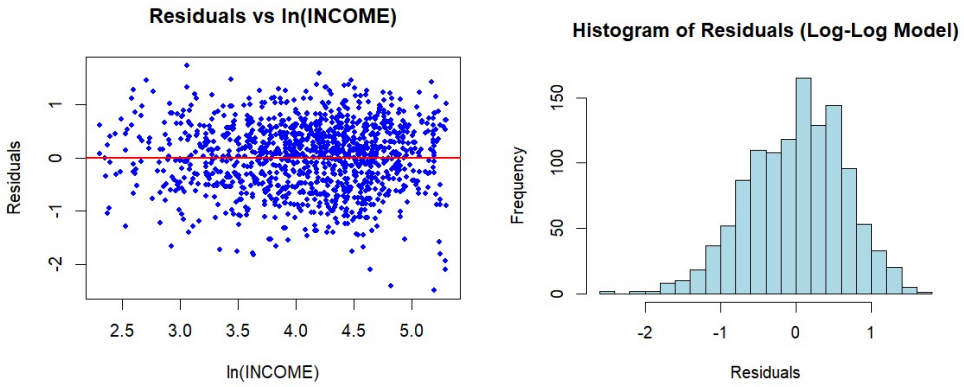
對數-對數模型的 R^2: 0.03322915

線性模型的 R^2: 0.0422812

由 R-square 可看出，相比於 log-log 模型，線性模型有較好的擬合度。

f. 在 log-log 模型中，係數就是彈性，因此彈性為 0.1863，與 d 小題不同，在 log-log 模型中，各點的彈性是固定的。
而彈性的 95% 信賴區間: [0.1293432 , 0.2432675]

g.

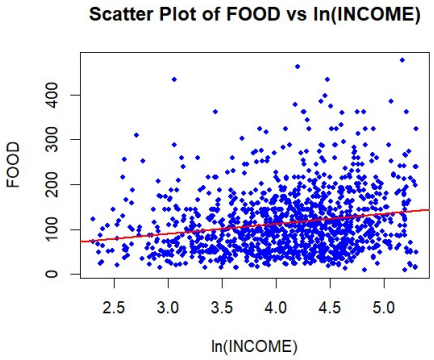



```
    Jarque Bera Test
data:  loglog_residuals
X-squared = 25.85, df = 2, p-value = 2.436e-06
```

從圖中我們可以得知殘差項可能存在異質變異數的問題，並且檢定結果有證據拒絕殘差項符合常態分配之虛無假設。

h. Linear-log 模型的結果 :

```
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 .
ln_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
```



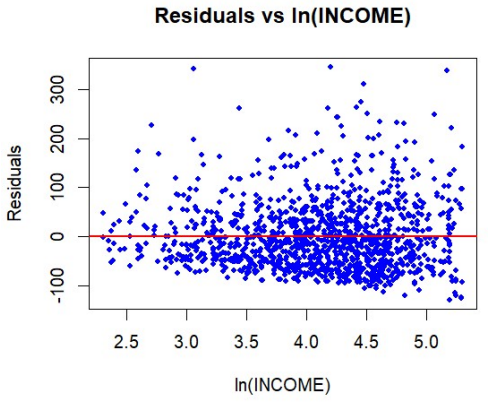
Linear-log 模型的 R-square 為 0.0380，比 log-log 模型略大，但仍舊小於線性模型，因此線型模型仍是三個模型中擬合程度最好的。

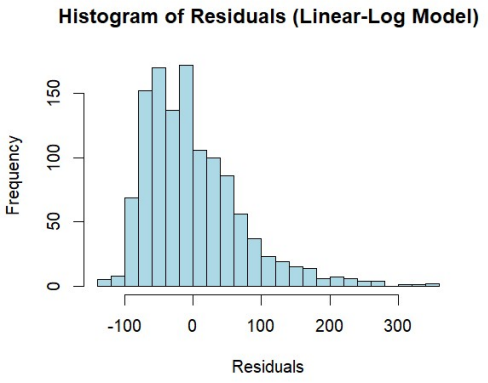
i.

	income	predicted_food	elasticity	CI_lower	CI_upper
1	19	88.89788	0.2495828	0.1784009	0.3207648
2	65	116.18722	0.1909624	0.1364992	0.2454256
3	160	136.17332	0.1629349	0.1164652	0.2094046

與前兩個模型相比，linear 模型的信賴區間是逐漸擴大，而 log-log 模型的彈性是固定值，linear-log 模型的信賴區間則是有逐漸縮小的趨勢。

j. 由殘差圖以及檢定結果均說明此模型的殘差項不服從常態分配假設，模型係數的估計上會出現偏誤。






```
        Jarque Bera Test

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

k. 儘管以上三個模型的殘差項均不符合常態分配假設，會使估計出現偏誤，不過若僅依照 R-square 來選擇模型，則單純的線性模型能提供最多的解釋力。