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:

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

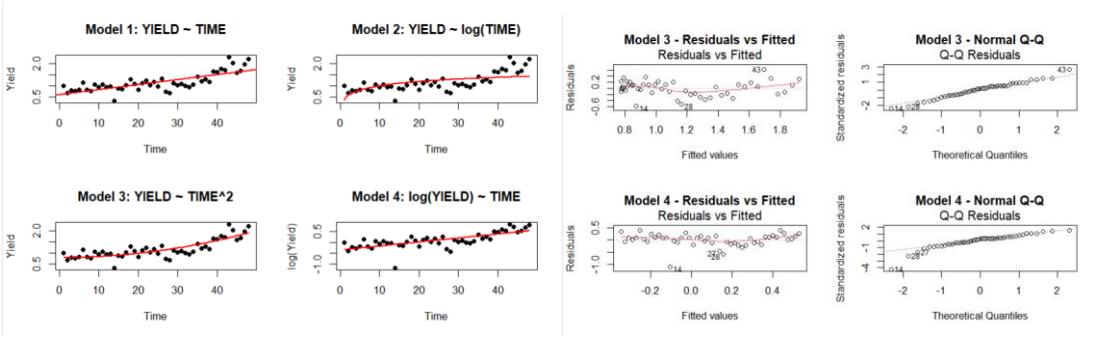
a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.

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).(ii)



(iii)

	Shapiro-Wilk W 值	p-value	conclusion
Model 1: YIELD ~ TIME	W = 0.98236	p-value = 0.6792	Not reject H0，殘差可能服從常態
Model 2: YIELD ~ log(TIME)	W = 0.96657	p-value = 0.1856	Not reject H0，殘差可能服從常態
Model 3: YIELD ~ TIME²	W = 0.98589	p-value = 0.8266	Not reject H0，殘差可能服從常態
Model 4: log(YIELD) ~ TIME	W = 0.86894	p-value = 7.205e-05	Reject H0，殘差可能不服從常態

(iv)

	R square
Model 1: YIELD ~ TIME	0.5778
Model 2: YIELD ~ log(TIME)	0.3386
Model 3:	0.6890

YIELD ~ TIME²	
Model 4: log(YIELD) ~ TIME	0.5074

根據以上資訊可看出 model 3 比較好

b.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.737e-01  5.222e-02   14.82  < 2e-16 ***
I(TIME^2)   4.986e-04  4.939e-05   10.10 3.01e-13 ***
```

隨著時間增加，產量變化呈現加速成長，時間每增加一年，對於產量增加的邊際影響= 0.0009972 × TIME。

c.

```
> print(unusual_obs)
   Obs Studentized_Residual   Leverage      DFFITS DFBETA_TIME2
14  14           -2.5606825 0.03593775 -0.4944002  0.320519995
28  28           -2.2468473 0.02083617 -0.3277591  0.003822742
43  43            2.8894474 0.06829921  0.7823199  0.652179762
44  44            1.3788286 0.07643579  0.3966661  0.338316939
45  45           -0.7794852 0.08542511 -0.2382270 -0.207150911
46  46           -0.5694893 0.09531255 -0.1848466 -0.163400687
47  47            0.4111600 0.10614453  0.1416857  0.127022369
48  48            1.3884647 0.11796846  0.5077802  0.460766575
```

d.

Using the quadratic model 3 estimated on data from 1950 to 1996, we forecasted the wheat yield in 1997 (time = 48).

The predicted yield is 1.8811, with a 95% prediction interval of [1.3724, 2.3898].

The actual observed yield in 1997 was 2.2318, which is within the prediction interval.

Therefore, the model performs reasonably well in capturing the yield trend, although the prediction is slightly lower than the observed 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 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.

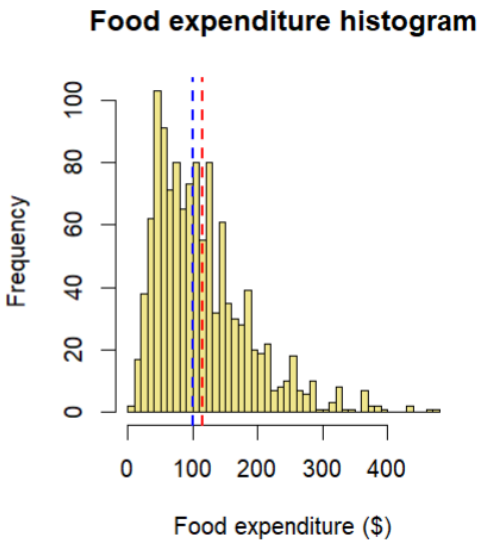
b. Estimate the linear relationship $FOOD = \beta_1 + \beta_2 INCOME + e$. Create a scatter plot *FOOD* versus *INCOME* and include the fitted least squares line. Construct a 95% interval estimate for β_2. Have we estimated the effect of changing income on average *FOOD* relatively precisely, or not?

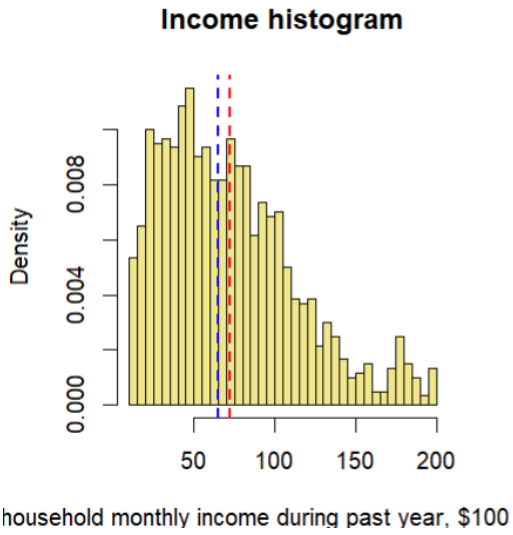
c. Obtain the least squares residuals from the regression in (b) and plot them against *INCOME*. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. Is it more important for the variables *FOOD* and *INCOME* to be normally distributed, or that 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?

(a.)



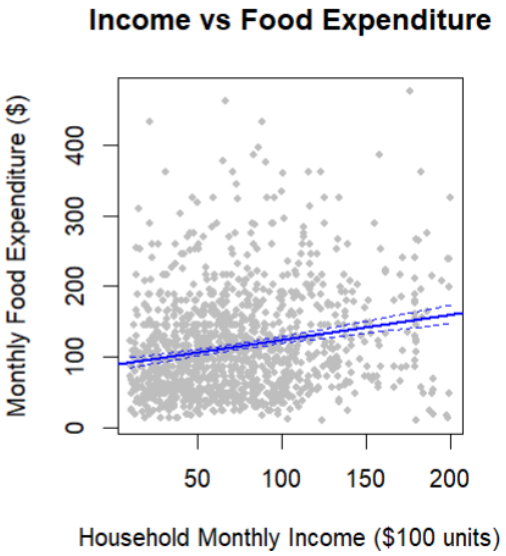



```
> summary(cex5_small$food)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.63   57.78   99.80  114.44  145.00  476.67
> summary(cex5_small$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   40.00   65.29   72.14   96.79  200.00
>
> sd(cex5_small$food, na.rm = TRUE)
[1] 72.6575
> sd(cex5_small$income, na.rm = TRUE)
[1] 41.65228
```

```
        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
```

食物支出和收入的平均都大於中位數，由直方圖可看出分配為右偏分配，並不是對稱分配。收入的常態分配檢定為 148.21，食物支出的常態分配檢定為 648.65，均拒絕虛無假設，有證據說明兩者都不服從常態分配。

(b.)

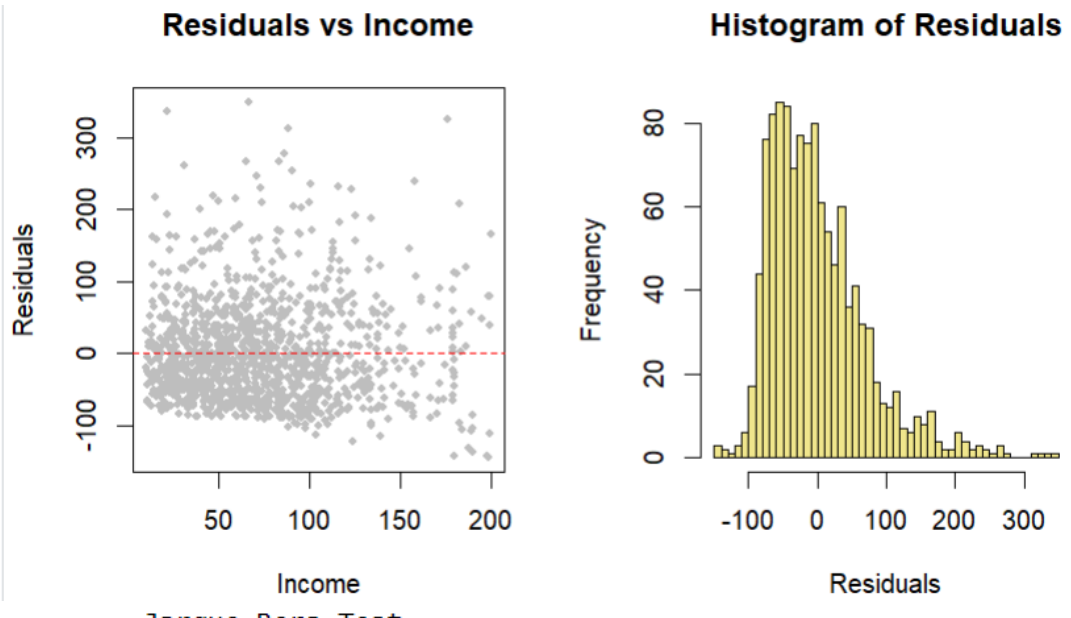



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.56650    4.10819  21.559  < 2e-16 ***
income       0.35869    0.04932   7.272 6.36e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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, "income", level = 0.95)
             2.5 %    97.5 %
income 0.2619215 0.455452
```

收入與食物支出之間存在正相關，迴歸斜率的 95% 信賴區間為 [0.2619, 0.4555]，每增加 100 美元的家庭月收入，將會使每人平均的食物支出增加約 0.26 至 0.46 美元。係數為顯著的但信賴區間寬度較寬仍不確定。

(c.)




```
Jarque Bera Test

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

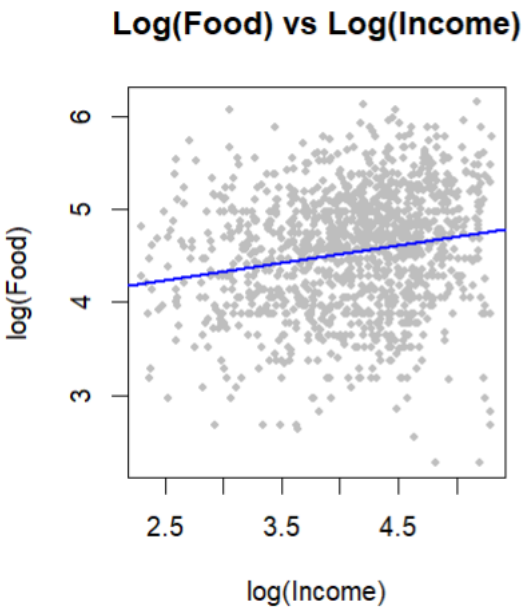
殘差呈現右偏的分布，在散布圖中僅高收入區有發散的現象，其他區域都是集中的，且很少殘差位於負 100 以下，殘差的常態分配檢定為 624.186，p 值也小於 0.05，因此拒絕殘差為常態分配的假設，誤差項是否為常態分配更重要。

(d.)

	INCOME	Predicted_FOOD	Elasticity	CI_Lower	CI_Upper
1	19	95.38	0.0715	0.0522	0.0907
2	65	111.88	0.2084	0.1522	0.2645
3	160	145.96	0.3932	0.2872	0.4992

三個收入下的彈性估計值差異大，且信賴區間沒有重疊，表示彈性隨收入上升而增加。根據經濟學，食物屬於必要財，收入提高時，其支出佔比上升幅度會趨緩，因此**理論上彈性應下降**，但實證結果顯示彈性上升，可能與高收入家庭選擇較高價食物有關。

(e.)




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.77893    0.12035  31.400   <2e-16 ***
log_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
```

經過 log 轉換後，兩者的散布圖分配的較平均，但其 r square 較線性模型小，因此線性模型較適合。

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?

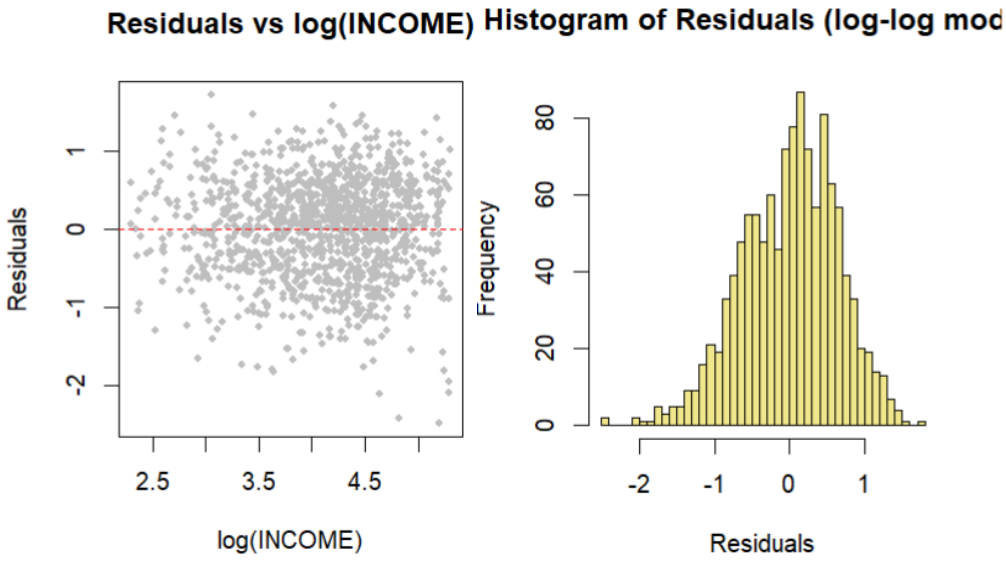
k. Based on this exercise, do you prefer the linear relationship model, or the log-log model or the linear-log model? Explain your reasoning.

(f.)

```
> summary(model_loglog)$coefficients
              Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 3.7789321 0.12034918 31.399733 2.113225e-158
log(income) 0.1863054 0.02903348  6.416915  1.998516e-10
> confint(model_loglog, level = 0.95)
                2.5 %    97.5 %
(Intercept) 3.5428135 4.0150507
log(income) 0.1293432 0.2432675
```

彈性為 0.1863，95%信賴區間為[0.1293, 0.2433]，和 INCOME = 65 時的線性模型信賴區間大致相似。表示若我們假設 log-log 模型的彈性等於線性模型在 INCOME = 19 或 160 下的任何一個區間值，那麼在 5% 顯著水準下，我們將會拒絕虛無假設。

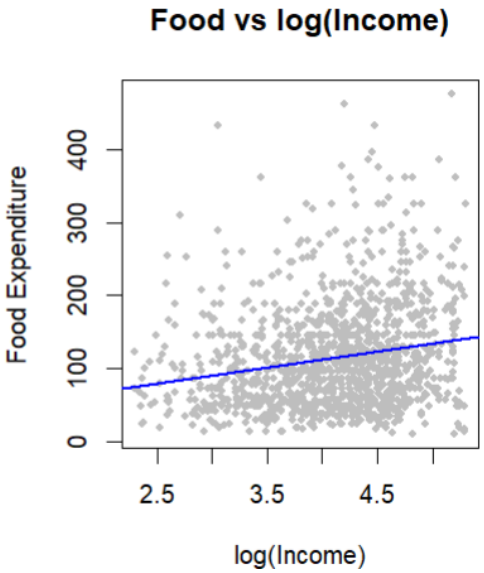
(g.)



```
        Jarque Bera Test

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

殘差圖分配為左偏分配，常態分配檢定 p 值小於 0.05，可拒絕虛無假設，此模型殘差非常態分配。

(h.)




```
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 ***
---
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
```

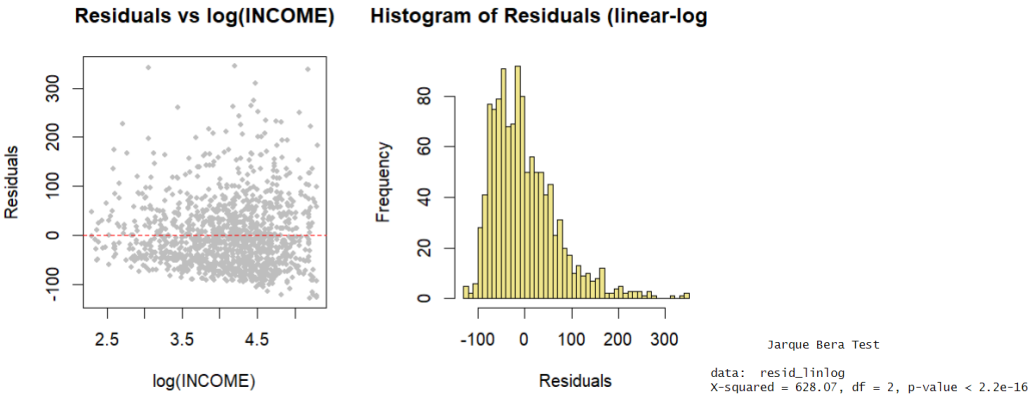
此圖與線性模型類似，資料點均集中於某些區域，不像 log-log 模型分布較均勻，且此模型 r square 最低因此線性模型仍為最適合的模型。

(i.)

	INCOME	Predicted_FOOD	Elasticity	CI_Lower	CI_Upper
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

在較高的收入下，彈性會下降，最低與最高收入的彈性信賴區間是有重疊的。這些彈性估計值與 log-log 模型的彈性（0.1863）相當接近。由於 linear-log 模型的彈性會隨收入下降，所以它與線性模型所估計的彈性區間不相似。

(j.)




```
        Jarque Bera Test

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

殘差的分配為右偏分配，由常態分配檢定可看出 p 值小於 0.05，無法拒絕虛無假設，無證據支持殘差為常態分配。

(k.)

Log-log 模型的殘差散布圖最均勻檢定統計量也最接近常態分配，雖然他的彈性都是常數違反真實狀況，但這是折衷下來較好的選擇。