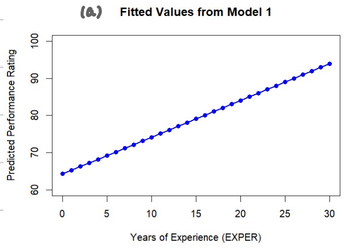
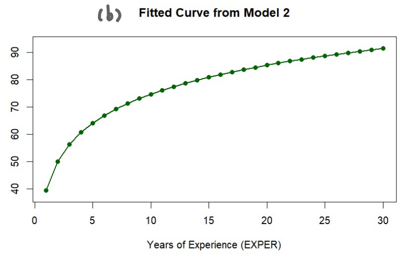
Chapter 4

4.

(a) $\widehat{RATING} = 64.289 + 0.99\,EXPR$

$\begin{cases} \text{When } EXPR = 0,\ \widehat{RATING} = 64.289 \\ \text{When } EXPR = 10,\ \widehat{RATING} = 74.189 \end{cases}$





(b) $\widehat{RATING} = 39.464 + 15.312\ln(EXPER)$

$\begin{cases} \text{When } EXPR = 0,\ \widehat{RATING} = \\ \text{When } EXPR = 10,\ \widehat{RATING} = \end{cases}$

→ Because Model 2 use the natural logarithm of EXPR,
it will cause a math error in the regression when EXPR = 0.
Therefore, 4 artists with no experience are not used.

(c) $\widehat{RATING} = 64.289 + 0.99\,EXPR$

$\frac{d(RATING)}{d(EXPR)} = 0.99$ → No matter what the value of EXPR is, the marginal effect is always 0.99.
Therefore, with 10y and with 20y are both 0.99

(d) $\widehat{RATING} = 39.464 + 15.312\ln(EXPER)$

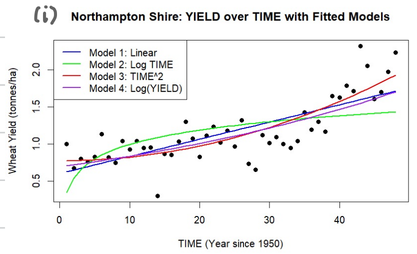
$\frac{d(RATING)}{d(EXPR)} = \frac{35.12}{EXPR}$ → (i) When EXPR = 10, marginal effect = 1.5312
(ii) When EXPR = 20, marginal effect = 0.7656

(e) Model 2 fits the data better.
→ With larger value of R^2, the data can explain more variance.
Since 0.3793 (model 1) < 0.4858 (model 2), Model 2 fits better.

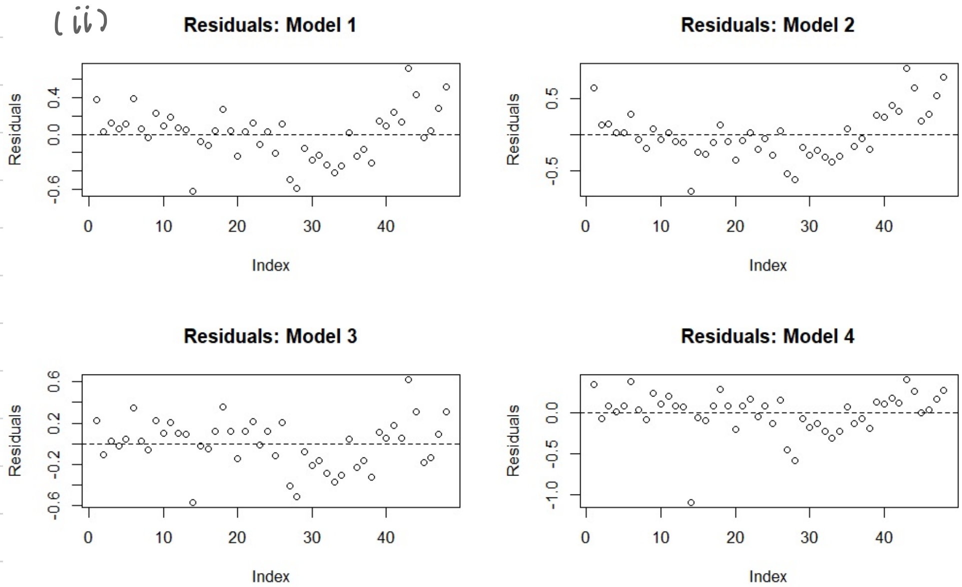
(f) Model 2 is more plausible.
→ Learning curves in real life are typically steep in the beginning and flatten over time.

28.

(a) (i)



(ii)



(iii) (iv)

Model	Description	p-value	R^2
Model 1	Linear	0.9359	0.578
Model 2	Log(TIME)	0.2512	0.339
Model 3	TIME²	0.8504	**0.689**
Model 4	Log(YIELD)	0.0000	0.507

→ Model 3 is preferable.
It has the highest R^2 with
p-value = 0.8504 > 0.5.
Besides, its residuals spread more randomly, meaning that it has better goodness of fit.

(b)

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.56899 -0.14970  0.03119  0.12176  0.62049 

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

Residual standard error: 0.2396 on 46 degrees of freedom
Multiple R-squared:  0.689,	Adjusted R-squared:  0.6822 
F-statistic: 101.9 on 1 and 46 DF,  p-value: 3.008e-13
```

→ $\widehat{YIELD} = 0.7737 + 0.0004986\, TIME^2$

→ The coefficient is 0.0004986, meaning that for each additional unit increase in time, the wheat yield is expected to increase by $0.0004986 \times (2 \cdot TIME)$.
Hence, the marginal effect of time increase over time.

(c)

```
no observations flagged by Studentized Residual > 2:
Obs TIME  YIELD Studentized_Residual Leverage DFBETA_TIME2  DFFITS
 14   14 0.3024              -2.5607   0.0359       0.3205 -0.4944
 28   28 0.6539              -2.2468   0.0208       0.0038 -0.3278
 43   43 2.3161               2.8894   0.0683       0.6522  0.7823
no observations flagged by Leverage > 0.0833:
Obs TIME  YIELD Studentized_Residual Leverage DFBETA_TIME2  DFFITS
 45   45 1.6040              -0.7795   0.0854      -0.2072 -0.2382
 46   46 1.6980              -0.5695   0.0953      -0.1634 -0.1848
 47   47 1.9691               0.4112   0.1061       0.1270  0.1417
 48   48 2.2318               1.3885   0.1180       0.4608  0.5078
no observations flagged by DFFITS > 0.4082:
Obs TIME  YIELD Studentized_Residual Leverage DFBETA_TIME2  DFFITS
 14   14 0.3024              -2.5607   0.0359       0.3205 -0.4944
 43   43 2.3161               2.8894   0.0683       0.6522  0.7823
 48   48 2.2318               1.3885   0.1180       0.4608  0.5078
no observations flagged by DFBETAS > 0.2887:
Obs TIME  YIELD Studentized_Residual Leverage DFBETA_TIME2  DFFITS
 14   14 0.3024              -2.5607   0.0359       0.3205 -0.4944
 43   43 2.3161               2.8894   0.0683       0.6522  0.7823
 44   44 2.0534               1.3788   0.0764       0.3383  0.3967
 48   48 2.2318               1.3885   0.1180       0.4608  0.5078
```

Leverage: $\frac{2(k+1)}{n} = \frac{2(1+1)}{48} \approx 0.0833$

DFFITS: $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{48}} \approx 0.2887$

DFBETAS: $2 \cdot \sqrt{\frac{k+1}{n}} = 2 \cdot \sqrt{\frac{2}{48}} \approx 0.4082$

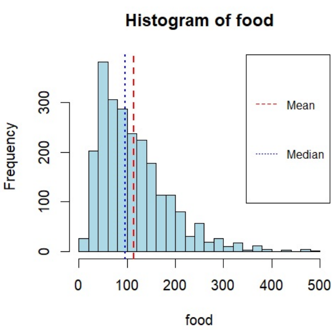
(d) 95% prediction interval = [1.3724, 2.3898].
The actual yield in 1997 is 2.2318, which is located in the interval.

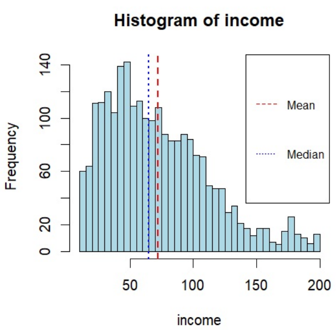
29.

(a) (i)

```
Summary for food :
Mean:     113.4269
Median:   96.3
Min:      4.81
Max:      494.44
SD:       71.31282

Summary for income :
Mean:     71.89751
Median:   65
Min:      10
Max:      200
SD:       40.8618
```

(ii)



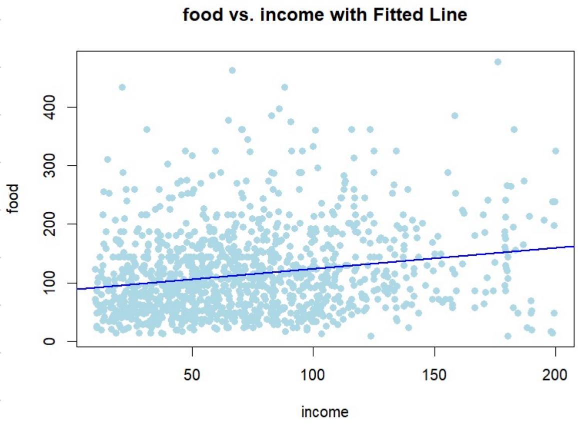


→ Both are not symmetrical and bell-shaped.
Instead, they are right-skewed, the sample mean > median.

(iii)

Variable	JB Statistic	Degrees of Freedom	p-value	Normality Conclusion
FOOD	1280.1	2	< 2.2e-16	Not normal (reject H_0)
INCOME	284.44	2	< 2.2e-16	Not normal (reject H_0)

(b)



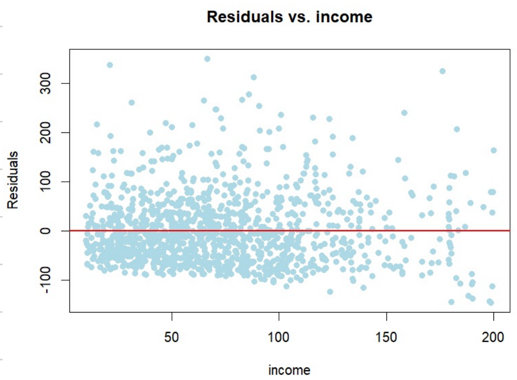

```
Residuals:
    Min      1Q  Median      3Q     Max
-145.37  -51.48  -13.52   35.50  349.81

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
```

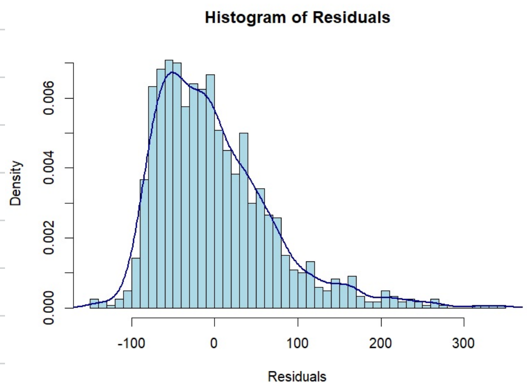
→ $\widehat{FOOD} = 88.5665 + 0.35869\, INCOME$

→ 95% interval for $\beta_2 = [0.2619, 0.4555]$

→ It does NOT relatively precise.
Though the coefficient is concluded in the interval, adjust-R^2 is low.
Hence, income alone explains very little of the variation in food spending.

(c)





→ The residuals are NOT randomly scattered.
Instead, when income income increases, the variance of residuals increases, suggesting "heteroskedasticity".

→ Jarque–Bera test: JB statistic = 624.19 with df = 2, p-value < 2.2×10^{-16} → reject H_0 of normality.

→ It is more important for the error term e to be normally distributed since it influences the correction of the suppose in OLS, what we assume that e should be normally distributed.

(d)

	income	Fitted_FOOD	Elasticity	Elasticity_Lower	Elasticity_Upper
1	19	95.3815	0.0715	0.0522	0.0907
2	65	111.8811	0.2084	0.1522	0.2646
3	160	145.9564	0.3932	0.2871	0.4993

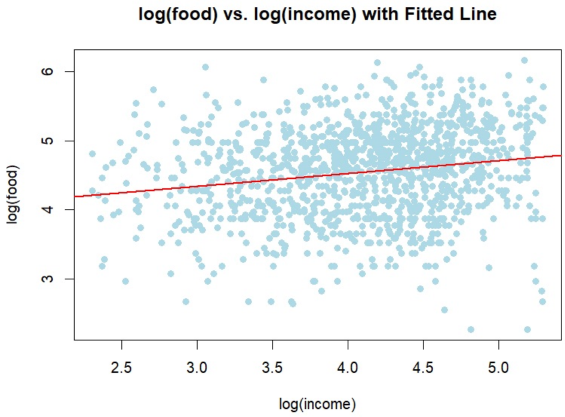
→ The elascity increases with income.

→ Interval estimates overlap slightly at low levels, indicating elasticities are statistically different at higher income levels.

→ Food is a necessity.
However, income elasticity of demand for food should decline as income rises according to Engle's law.
Hence, it is not consistent with economics.

(e)

```
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
```



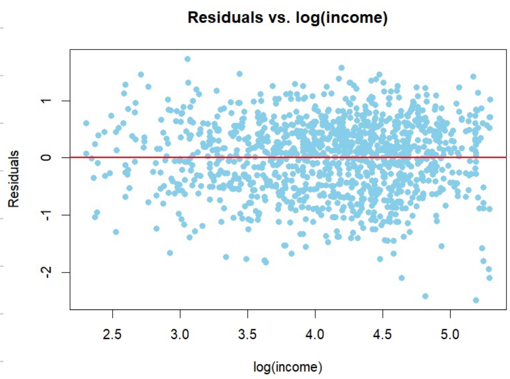
→ $\ln(FOOD) = 3.7789 + 0.1863 \ln(INCOME)$

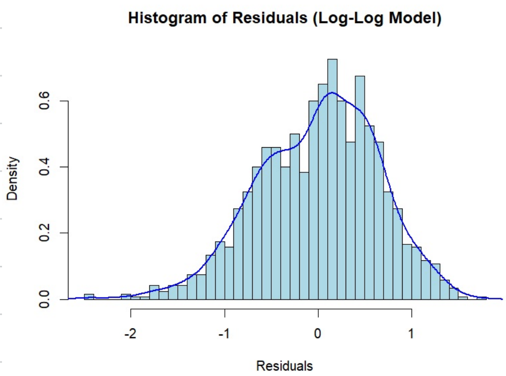
→ Consider the plot, in the log-log plot, the points appear more evenly spread around the fitted line, making the model is more well-defined.

→ As for adjusted-R^2, both models have lower value of adjusted-R^2, meaning that both models explain little of the variation in food spending. However, the linear model has higher adjusted-R^2, it seems fit the model better.

(f) The point elasticity is 0.1863 and the 95% interval estimate is [0.1294, 0.2432]. The elasticity of log-log model is fixed instead of increasing with income. Therefore, it is dissimilar with that in part (d).

(g)





→ Scatter Plot: It does NOT have clear plot.

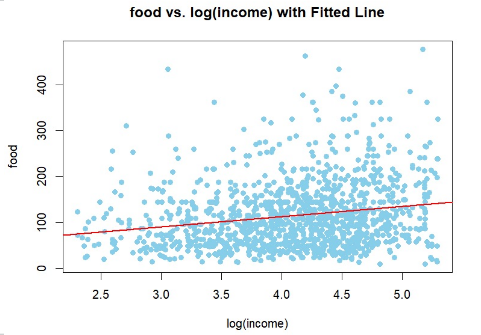
→ Histogram: It is roughly bell-shaped, but slightly left-skewed, showing that the residuals of log-log model is more normally distributed than that of the linear model.

→ Jarque-Bera test: Since p-value < 0.05, we reject H_0: the residuals of log-log model is not normally distributed.

```
	Jarque Bera Test

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

(h)

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



→ $FOOD = 23.568 + 22.187 \ln(INCOME) + e$

→ Based on plots, the spread of points around the fitted line looks more stable. The relationship is more well-defined in the linear-log model.

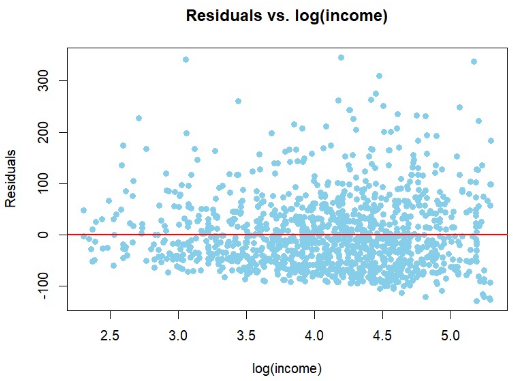
→ Based on R^2, 0.0332 (log-log) < 0.038 (linear-log) < 0.0423 (linear), the linear model fits better.

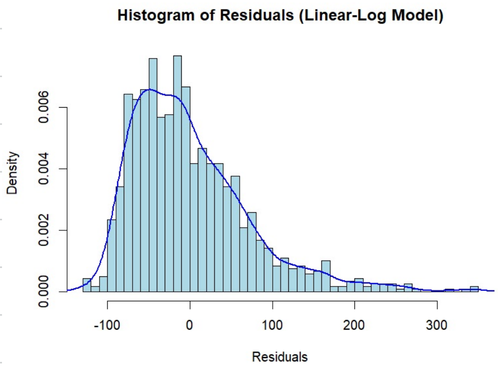
(i)

	income	Fitted_food	Elasticity	Lower_95CI	Upper_95CI
1	19	88.90	4.7421	3.3910	6.0932
2	65	116.19	12.4126	8.8760	15.9491
3	160	136.17	26.0696	18.6418	33.4974

→ It is dissimilar with other models.
Linear-log : the elasticity decreases as income increases
Linear : the elasticity increases with income.
Log — log : fixed

(j)





→ Scatter Plot : It does NOT have clear plot.

→ Histogram : It is right—skewed.

→ Jarque — Bera test : Since p—value < 0.05, we reject H_0 : the residuals of linear-log model is not normally distributed.

```
        Jarque Bera Test

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

(k) From the point of R^2, all three models have lower power to explain the data. However, the log-log model has more normally distributed residuals. Thus, I prefer the log-log model.