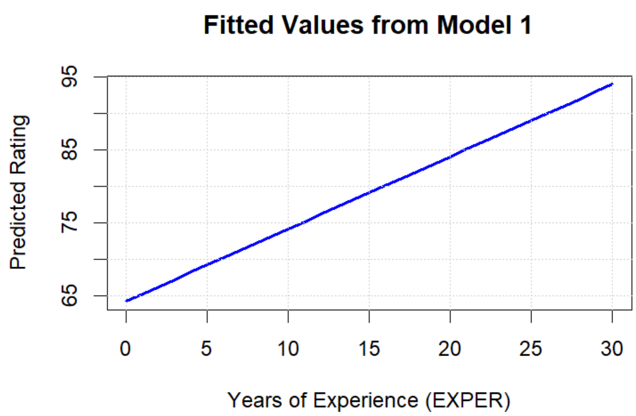
HW0317 Pinyo – 312712017

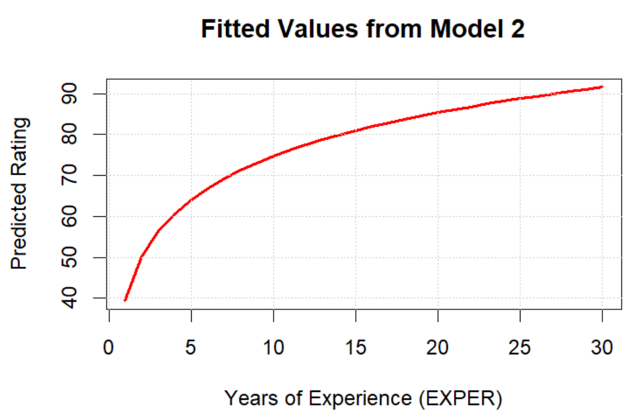
HW0317Q1

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:

a.



b.



0 experience artists information is not used in model 2 sine it is a log equation and ln(0) is undefined

c.

Since the model is **linear**, the marginal effect is constant across all values of EXPER. Therefore:

1. **For an artist with 10 years of experience**

d(RATING^) / d(EXPER) = 0.990

So, the marginal effect is **0.990**.

2. **For an artist with 20 years of experience**

d(RATING^) / d(EXPER) = 0.990

Again, the marginal effect is **0.990**.

d.

The marginal effect of another year of experience is found by differentiating the equation with respect to EXPER:

d(RATING^) / d(EXPER) = 15.312 × (1 / EXPER)

Thus, the marginal effect of an additional year of experience is:

ME = 15.312 / EXPER

For an artist with 10 years of experience (EXPER = 10)

ME = 15.312 / 10 = 1.5312

So, the marginal effect is **1.5312**.

For an artist with 20 years of experience (EXPER = 20)

ME = 15.312 / 20 = 0.7656

So, the marginal effect is **0.7656**.

Unlike Model 1, the marginal effect in Model 2 **decreases** as experience increases. This means that additional years of experience have a diminishing effect on **RATING** as artists gain more experience e.

- **Model 1 (All 50 artists, including those with zero experience):** $R^2 = 0.3793$
- **Model 1 (Only artists with experience, N=46N = 46N=46):** ^2 = 0.4858
- **Model 2 (Only artists with experience, N=46N = 46N=46):** $R^2 = 0.6414$

When considering **all 50 artists**, Model 1 has a relatively low $R^2 = 0.3793$, meaning it does not explain much variation in **RATING**.

- When **excluding** the four artists with **zero experience**, Model 1 improves slightly ($R^2 = 0.4858$), but it is still lower than Model 2.
- **Model 2 ($R^2 = 0.6414$)** has the highest R2R^2R2, meaning it explains the most variation in **RATING** among artists with experience.

f.

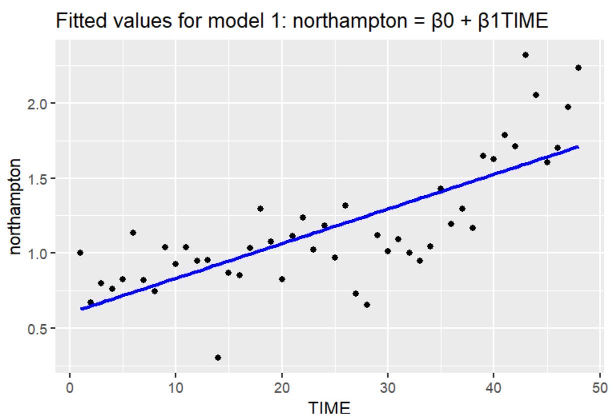
In most professions, work experience tends to improve performance but at a diminishing rate. Initially, gains from experience are large (learning curve effect), but as experience increases, the additional benefits become smaller. This suggests a non-linear relationship rather than a strictly linear one

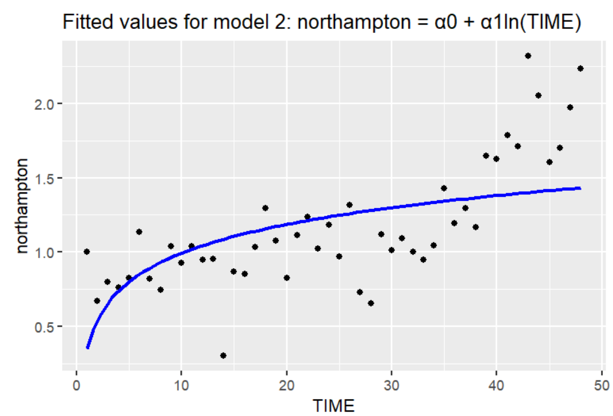
HW0317Q2

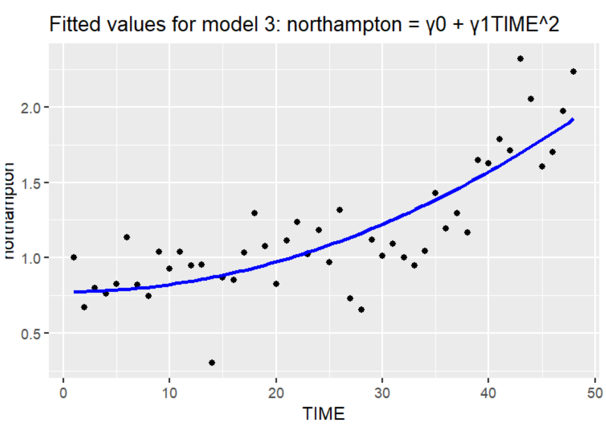
4.28 The file *wa-wheat.dat* contains observations on wheat yield in Western Australian shires. There are 48 annual observations for the years 1950–1997. For the Northampton shire, consider the following four equations:

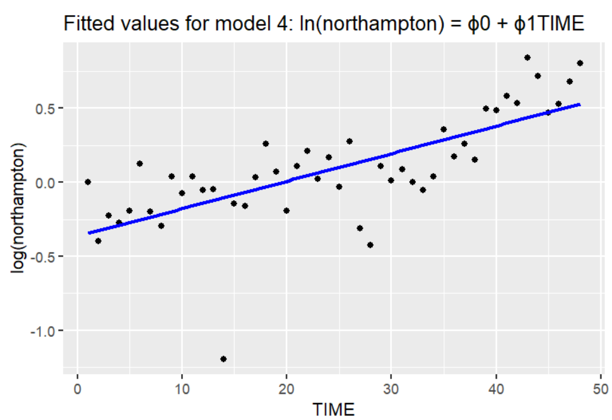
$$YIELD_t = \beta_0 + \beta_1 TIME + e_t$$
$$YIELD_t = \alpha_0 + \alpha_1 \ln(TIME) + e_t$$
$$YIELD_t = \gamma_0 + \gamma_1 TIME^2 + e_t$$
$$\ln(YIELD_t) = \phi_0 + \phi_1 TIME + e_t$$

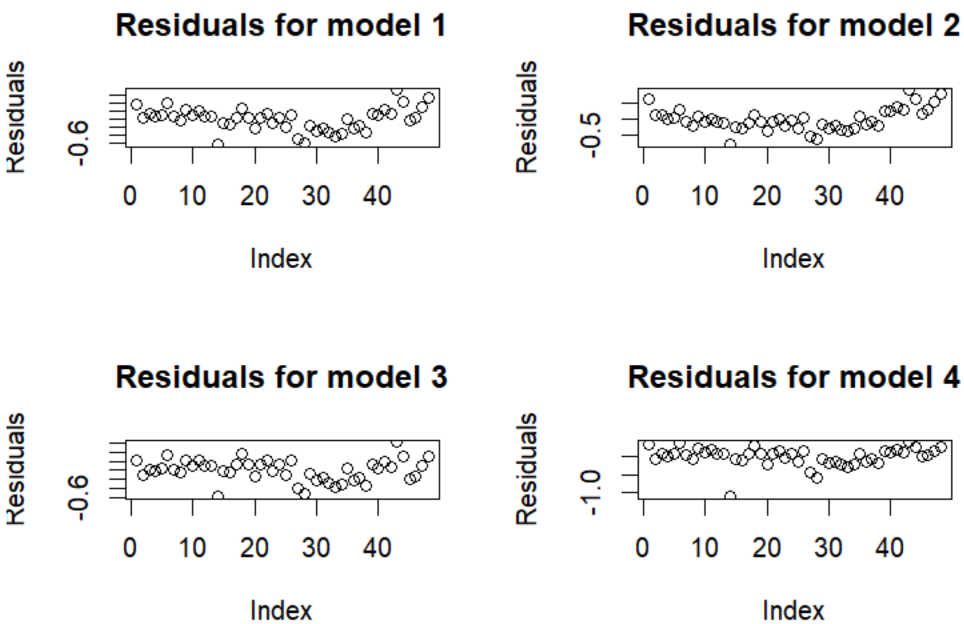
a.











Model 1: R^2 = 0.5687

Model 2: R^2 = 0.3242

Model 3: R^2 = 0.6822

Model 4: R^2 = 0.4966

From R^2 value, Model 3 is the suitable model to use year to explain yield in Northampton

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
```

Coefficient of model 3, $\gamma 0$ = 7.737e-01, $\gamma 1$ = 4.986e-04

c.

Observation	Studentized_Residuals		Leverage	DFBETAS_Intercept		DFBETAS_TIME2	DFFITS
1	1	0.97117127	0.04743473	1	0.216718802	-0.162293335	0.21671884
2	2	-0.43892315	0.04723338	2	-0.097727849	0.073063005	-0.09772816
3	3	0.09154376	0.04689948	3	0.020306565	-0.015139023	0.02030689
4	4	-0.09362102	0.04643560	4	-0.020658634	0.015340433	-0.02065970
5	5	0.17150978	0.04584531	5	0.037589949	-0.027768566	0.03759473
6	6	1.48946942	0.04513318	6	0.323736684	-0.237608499	0.32382335
7	7	0.09207526	0.04430484	7	0.019814787	-0.014429582	0.01982480
8	8	-0.25121369	0.04336691	8	-0.053440095	0.038555438	-0.05348720
9	9	0.97031031	0.04232704	9	0.203696093	-0.145364271	0.20399098
10	10	0.43928373	0.04119390	10	0.090847178	-0.064013954	0.09105340
11	11	0.88308253	0.03997718	11	0.179588800	-0.124702495	0.18020490
12	12	0.43133925	0.03868759	12	0.086097858	-0.058783799	0.08653117
13	13	0.40663499	0.03733687	13	0.079509348	-0.053242206	0.08008229
14	14	-2.56068246	0.03593775	14	-0.489450177	0.320519995	-0.49440017
15	15	-0.07921998	0.03450401	15	-0.014769753	0.009426586	-0.01497595
16	16	-0.20039139	0.03305043	16	-0.036357010	0.022524814	-0.03704808
17	17	0.49312413	0.03159284	17	0.086845864	-0.051978455	0.08906797
18	18	1.55776314	0.03014805	18	0.265587279	-0.152662876	0.27464917
19	19	0.51140018	0.02873391	19	0.084160844	-0.046123387	0.08796079
20	20	-0.61544215	0.02736930	20	-0.097452600	0.050450814	-0.10323931
21	21	0.51116364	0.02607410	21	0.077606439	-0.037496828	0.08363762
22	22	0.92505667	0.02486923	22	0.134136097	-0.059512311	0.14772978
23	23	-0.06616263	0.02377660	23	-0.009122960	0.003632899	-0.01032555
24	24	0.50898647	0.02281918	24	0.066410121	-0.022945195	0.07778015
25	25	-0.48508765	0.02202093	25	-0.059553461	0.016903995	-0.07279025
26	26	0.87138263	0.02140683	26	0.100005566	-0.021094711	0.12887955
27	27	-1.74863798	0.02100290	27	-0.186175532	0.023013433	-0.25612316
28	28	-2.24684727	0.02083617	28	-0.219908713	0.003822742	-0.32775913
29	29	-0.31870520	0.02093468	29	-0.028358757	-0.003242530	-0.04660328
30	30	-0.87713750	0.02132750	30	-0.069984024	-0.019709984	-0.12948485
31	31	-0.66971694	0.02204473	31	-0.047073157	-0.023570795	-0.10055048
32	32	-1.20147489	0.02311746	32	-0.072666378	-0.058096993	-0.18482622
33	33	-1.58783937	0.02457783	33	-0.079964118	-0.098380081	-0.25204730
34	34	-1.31340954	0.02645899	34	-0.052431496	-0.099841197	-0.21652582
35	35	0.17862729	0.02879511	35	0.005207561	0.016173251	0.03075755
36	36	-0.96321184	0.03162137	36	-0.017387057	-0.101664741	-0.17405621
37	37	-0.67396104	0.03497398	37	-0.004450857	-0.081583039	-0.12830329
38	38	-1.40993533	0.03889017	38	0.007330872	-0.193256227	-0.28361722
39	39	0.48980475	0.04340819	39	-0.008509056	0.075244047	0.10433872
40	40	0.21784595	0.04856730	40	-0.006520055	0.037193451	0.04921896
41	41	0.75037258	0.05440780	41	-0.032182354	0.141393550	0.17999300
42	42	0.24372124	0.06097100	42	-0.013713793	0.050388324	0.06210342
43	43	2.88944743	0.06829921	43	-0.202525494	0.652179762	0.78231995
44	44	1.37882863	0.07643579	44	-0.116349648	0.338316939	0.39666614
45	45	-0.77948519	0.08542511	45	0.077302871	-0.207150911	-0.23822701
46	46	-0.56948934	0.09531255	46	0.065201030	-0.163400687	-0.18484656
47	47	0.41115999	0.10614453	47	-0.053605470	0.127022369	0.14168569
48	48	1.38846474	0.11796846	48	-0.203926321	0.460766575	0.50778020

d.

```
> print(prediction)
       fit      lwr      upr
1 1.881111 1.372403 2.389819
>
> # Compare the prediction interval with the actual observed value in 1997
> actual_yield_1997 <- subset(northampton_data, TIME == 48)$northampton
> cat("Actual Yield in 1997:", actual_yield_1997, "\n")
Actual Yield in 1997: 2.2318
```

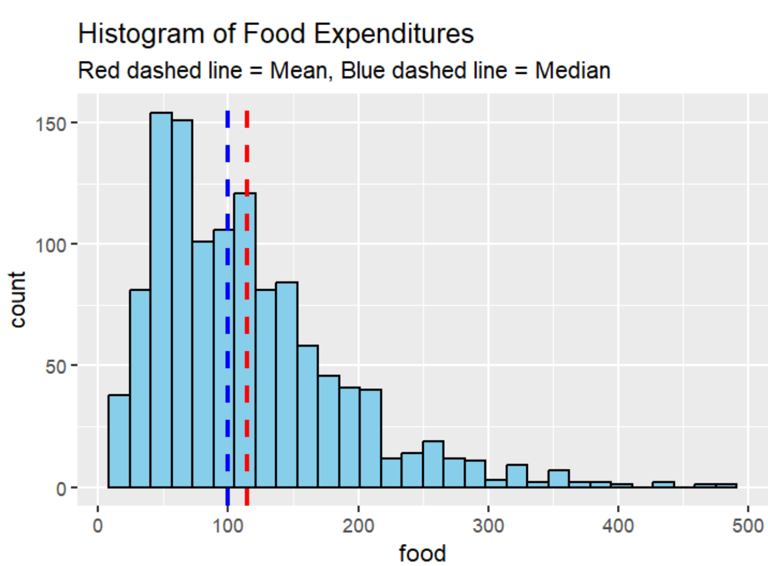
Actual yield fall in the prediction interval

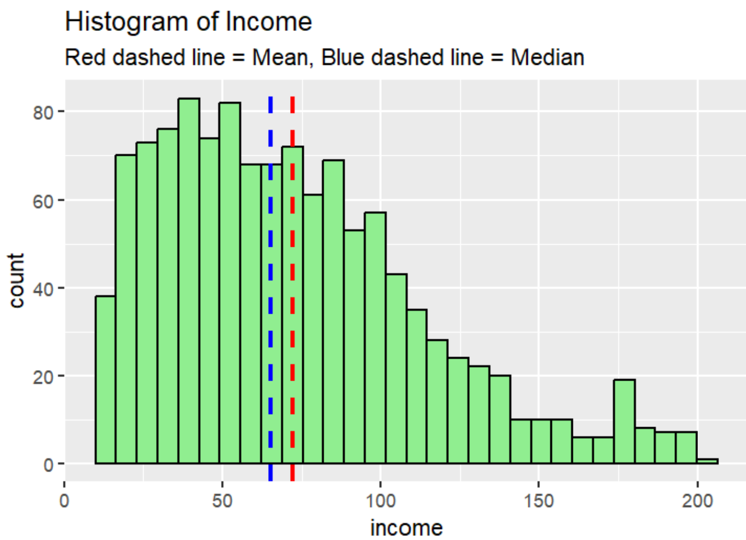
HW0317Q3

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.

```
> print(summary_stats)
  food_mean food_median food_min food_max food_sd income_mean income_median income_min
1  114.4431        99.8     9.63   476.67 72.6575    72.14264         65.29         10
  income_max income_sd
1        200  41.65228
```





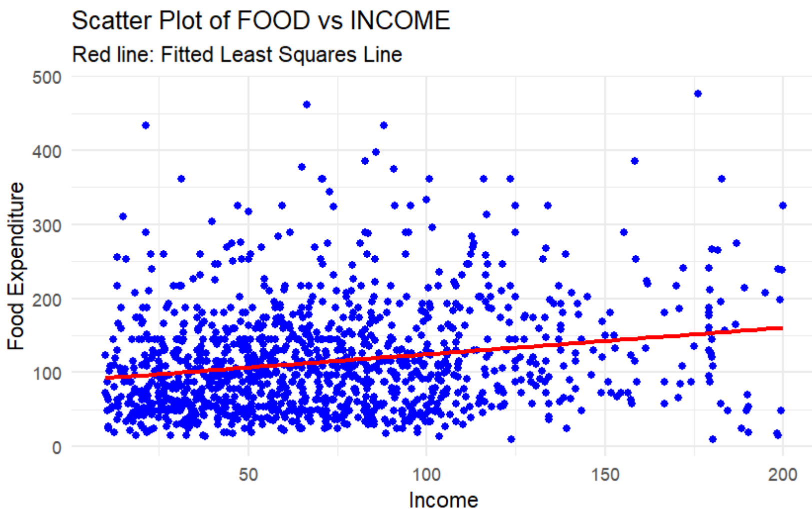
Both histograms are asymmetry (most data are in the left side of histogram. Also, both data have mean value to be greater than median value

```
Jarque-Bera Test Results:
> cat("food:   Test Statistic =", jb_test_food$statistic, "  p-value
=", jb_test_food$p.value, "\n")
food:   Test Statistic = 648.6476   p-value = 0
> cat("income: Test Statistic =", jb_test_income$statistic, "  p-value
=", jb_test_income$p.value, "\n")
income: Test Statistic = 148.2112   p-value = 0
>   ggtitle("Fitted values for model 4: ln(YIELD) = ϕ0 + ϕ1TIME")
```

Since the p-values for both "food" and "income" are 0 (which is generally interpreted as $p < 0.05$, or some other chosen significance level), we reject the null hypothesis in both cases. This means:

- The data for "food" is not normally distributed.
- The data for "income" is not normally distributed.

b.

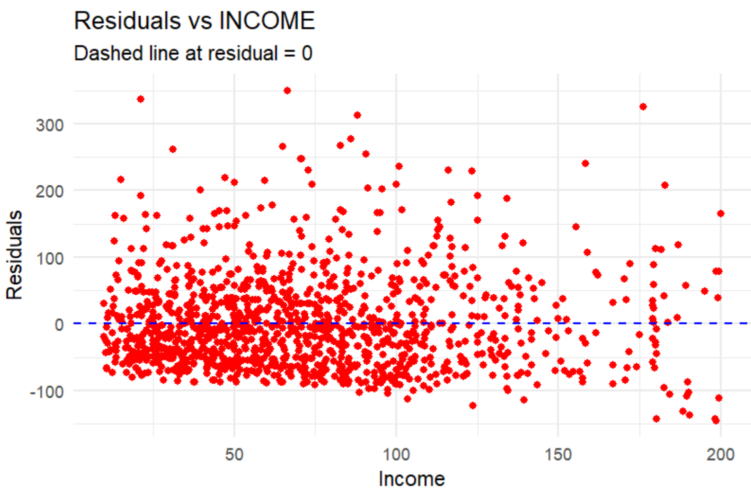



```
> cat("95% Confidence Interval for β2 (income):\n")
95% Confidence Interval for β2 (income):
> print(beta2_confidence_interval)
    2.5 %    97.5 %
0.2619215 0.4554520
```

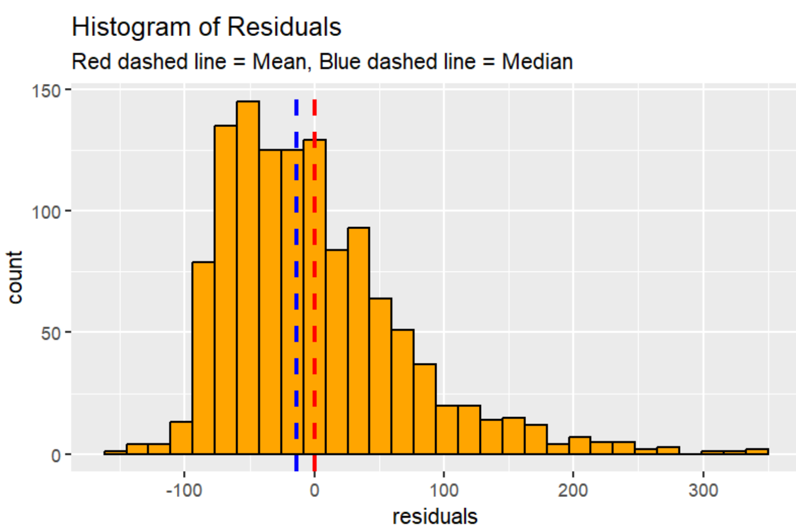
```
Residuals:
    Min      1Q  Median      3Q     Max
-145.37  -51.48  -13.52   35.50  349.81

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
```

From the value of adjust R^2, it shows that the linear regression model cannot explain the relationship between food and income precise enough.

c.



There is a pattern that the residual will be scattered wider when the income is higher.



From histogram, the residual is not normal distribution

```
Jarque-Bera Test for Residuals:
> cat("Test Statistic =", jb_test_residuals$statistic, "  p-valu
e =", jb_test_residuals$p.value, "\n")
Test Statistic = 624.186   p-value = 0
> # 5. Interpretation of normality importance
> cat("\nNormality Importance Explanation:\n")
```

It is more important that the random error (e) be normally distributed rather than the variables FOOD and INCOME themselves.

Assumption of Normality in Regression

In regression analysis, particularly Ordinary Least Squares (OLS), the normality assumption applies to the residuals (errors), not necessarily to the independent (INCOME) or dependent (FOOD) variables.

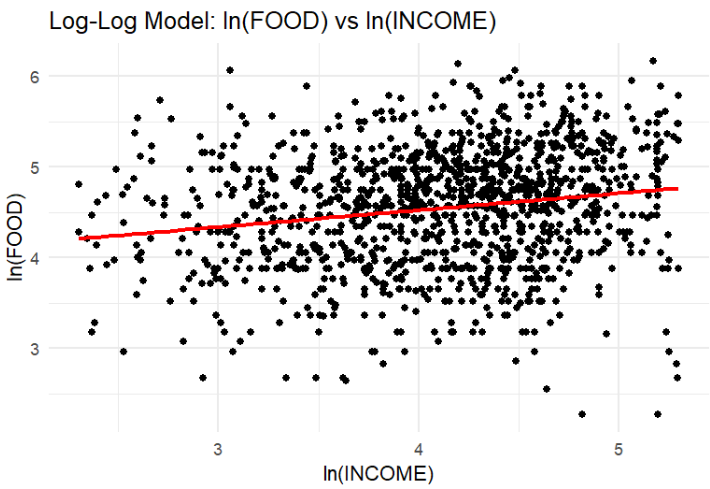
- Why residuals should be normal:
  - Ensures valid hypothesis testing (e.g., t-tests, F-tests).
  - Confidence intervals and p-values are accurate.
  - Prediction intervals are reliable.
- Why variables don't need to be normal:
  - OLS regression does not require independent or dependent variables to be normally distributed.
  - The Central Limit Theorem (CLT) states that as sample size increases, the distribution of the estimated coefficients approaches normality regardless of the original variable distributions.

d.

```
> print("Elasticity Estimates:")
[1] "Elasticity Estimates:"
> print(results)
  Income Fitted_Food Elasticity_Point_Estimate
1     19    95.38155                0.07145038
2     65   111.88114                0.20838756
3    160   145.95638                0.39319883
  Elasticity_Lower_Bound Elasticity_Upper_Bound
1             0.05217475             0.09072601
2             0.15216951             0.26460562
3             0.28712305             0.49927462
```

- Estimated elasticity is not similar from these 3 level of incomes
- Intervals of estimated elasticity are not overlapped
- Based on economic theory, elasticity for food will be increased when the household's income increases since they tend to consume more when they have more income

e.




```
Call:
lm(formula = log(food) ~ log(income), data = data)

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 ***
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
```

To compare between model in (b) and log model, R^2 of linear regression from (b) is 0.04148, while R^2 of log model in (e) is 0.03242. This means that linear regression model is better fit than log model

f.

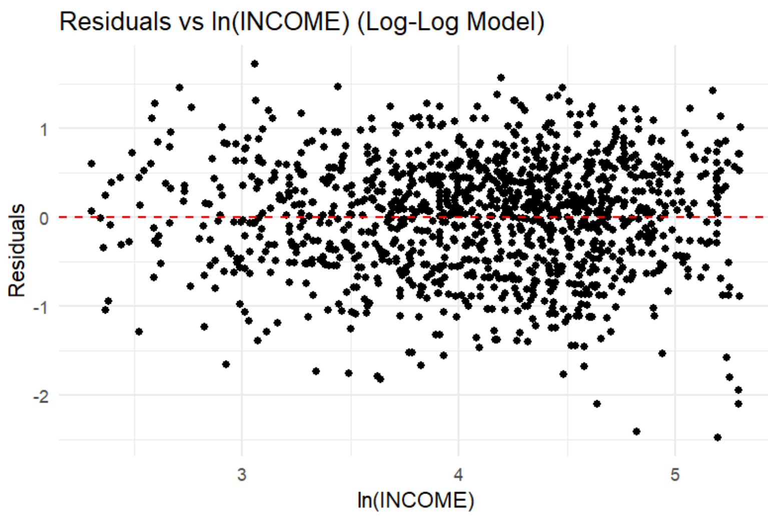
```
Point Estimate of Elasticity (Log-Log Model): 0.1863054
> cat("95% Confidence Interval for Elasticity (Log-Log Model):\n")
95% Confidence Interval for Elasticity (Log-Log Model):
> print(elasticity_confidence_interval)
    2.5 %    97.5 % 
0.1293432 0.2432675 
```

Statistics data from (d) is

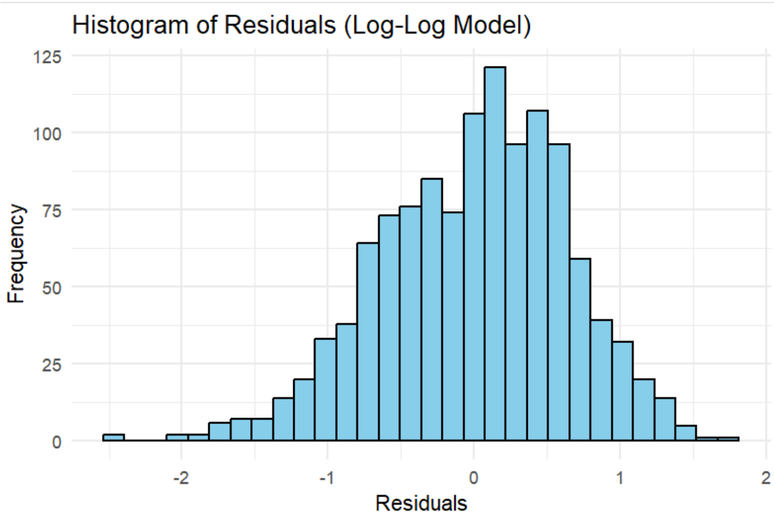
```
> print("Elasticity Estimates:")
[1] "Elasticity Estimates:"
> print(results)
  Income Fitted_Food Elasticity_Point_Estimate
1     19    95.38155                0.07145038
2     65   111.88114                0.20838756
3    160   145.95638                0.39319883
  Elasticity_Lower_Bound Elasticity_Upper_Bound
1             0.05217475             0.09072601
2             0.15216951             0.26460562
3             0.28712305             0.49927462
```

The confident interval is not similar

g.



There is a pattern that the residual will scatter more when ln(INCOME) increases.

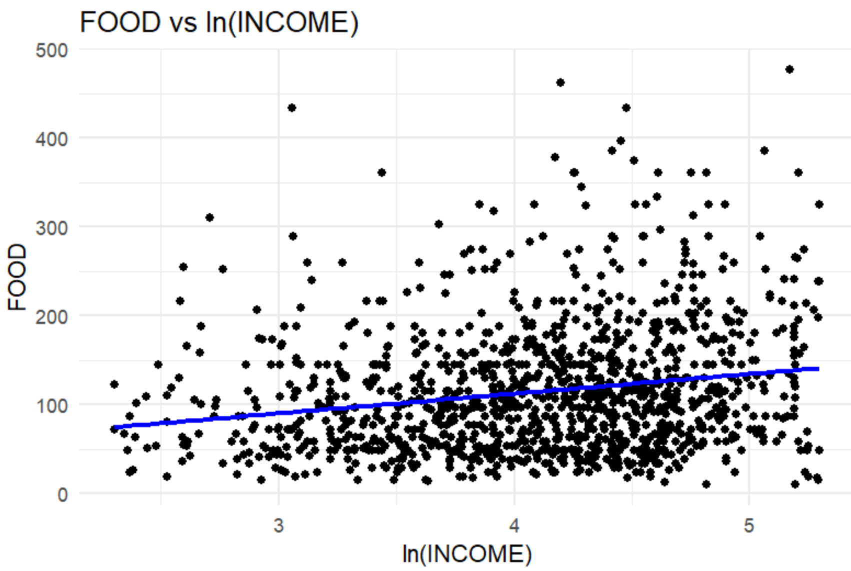


- **Data:** residuals log_log - This indicates that the test was performed on the residuals from your log-log regression model. This is the correct thing to do to assess whether the *errors* in your model are normally distributed.
- **X-squared = 25.85:** This is the test statistic. The Jarque-Bera test statistic follows a Chi-squared distribution. A larger value suggests a greater departure from normality.
- **df = 2:** This is the degrees of freedom for the Chi-squared distribution. The Jarque-Bera test has 2 degrees of freedom because it tests both skewness and kurtosis.

- **p-value = 2.436e-06:** This is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated (25.85) *if* the null hypothesis (normality) were true. 2.436e-06 is scientific notation for 2.436 x 10^-6 which is equal to 0.000002436

Since the p-value (0.000002436) is much smaller than the conventional significance levels (e.g., 0.05, 0.01), we **reject the null hypothesis**. This means there is strong statistical evidence to conclude that the residuals from your log-log regression model are *not* normally distributed. The deviations from normality, as measured by skewness and kurtosis, are statistically significant.

h.




```
Call:
lm(formula = food ~ log_income, data = data)

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
```

To compare between model in (b), and (e), R^2 of linear regression from (b) is 0.04148, R^2 of log model in (e) is 0.03242, while R^2 of linear-log in (h) is 0.0372. This means that linear regression model is better fit than log model i.

The elasticity interval from linear-log model provides the result as below:

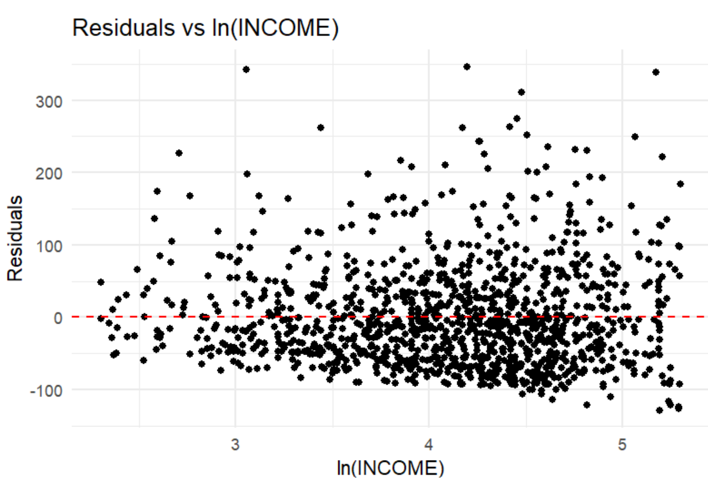
```
  INCOME Elasticity    Lower_CI  Upper_CI
1     19  1.1677571 0.83504479 1.5004694
2     65  0.3413444 0.24409001 0.4385988
3    160  0.1386712 0.09916157 0.1781807
> 
```

The elasticity interval from linear model provides the result as below:

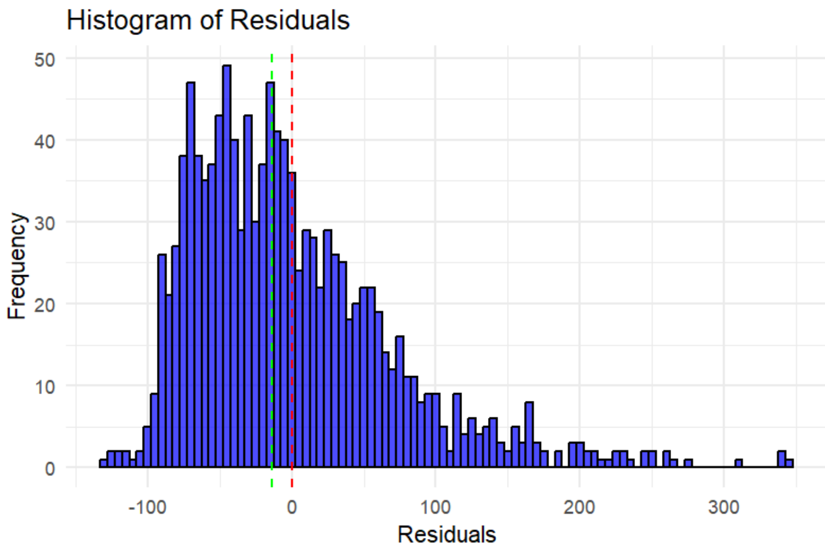
```
> print("Elasticity Estimates:")
[1] "Elasticity Estimates:"
> print(results)
  Income Fitted_Food Elasticity_Point_Estimate
1     19    95.38155                0.07145038
2     65   111.88114                0.20838756
3    160   145.95638                0.39319883
  Elasticity_Lower_Bound Elasticity_Upper_Bound
1             0.05217475             0.09072601
2             0.15216951             0.26460562
3             0.28712305             0.49927462
```

The estimate elasticity from both models are different.

j.



The spread of the residuals appears to increase as the value of ln(INCOME) increases. This is a classic sign of heteroscedasticity, meaning the variance of the errors is not constant across all levels of the independent variable




```
       Jarque Bera Test

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

The null hypothesis of the Jarque-Bera test is that the data is normally distributed. The extremely small p-value (much less than 0.05 or 0.01) provides overwhelming evidence to reject this null hypothesis. Therefore, you can conclude that the residuals are significantly non-normal

k.

From the statistic result, I prefer linear regression since it provides the highest R^2. In this case, normal distribution of residue is not the most significant factor since all models provide non-distribution of residuals.