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.

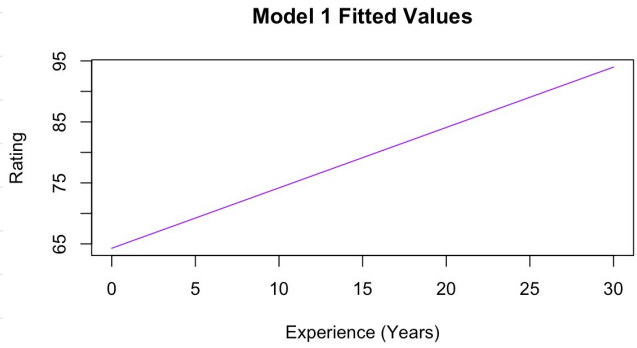
d. Using Model 2, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.

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

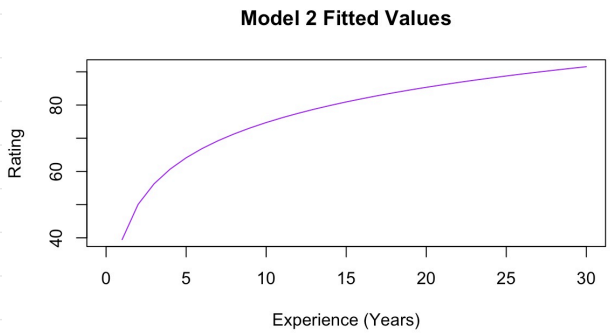
f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning? Explain.

(508)

a.



b.



The fitted values of Model 2 will be shown for the range of EXPER from 1 to 30 years, and there are no data points for EXPER = 0 because the logarithmic function cannot handle EXPER = 0. This explains why the four artists with no experience are excluded from the estimation of Model 2.

c. $\frac{\partial(\widehat{RATEING})}{\partial EXPER} = 0.990$

for an artist with 10 years or 20 years the margin effect also 0.990

d. $\frac{\partial \widehat{(RATEING)}}{\partial EXPER} = \frac{15.312}{EXPER}$

(i) $EXPER = 10 \Rightarrow \frac{\partial \widehat{(RATEING)}}{\partial EXPER} = \frac{15.312}{10} = 1.5312$

(ii) $EXPER = 20 \Rightarrow \frac{\partial \widehat{(RATEING)}}{\partial EXPER} = \frac{15.312}{20} = 0.7656$

e. To compare the goodness of fit between the two models, we can use the R^2 value as a criterion. R^2 is a measure of goodness of fit in regression models, indicating the proportion of variance explained by the model. Based on the R^2 values, Model 2 has a higher R^2 (0.6414), which means that Model 2 explains the data better and has a superior fit compared to Model 1.

f. Based on economic reasoning, Model 2 is the more reasonable and likely choice. It reflects the diminishing marginal effect of experience on job performance, which aligns better with the way experience typically affects performance in most real-world situations. While Model 1 is simple and easy to understand, it fails to capture the non-linear impact of experience on job performance. Therefore, from an economic perspective, Model 2 is more convincing.

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:

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

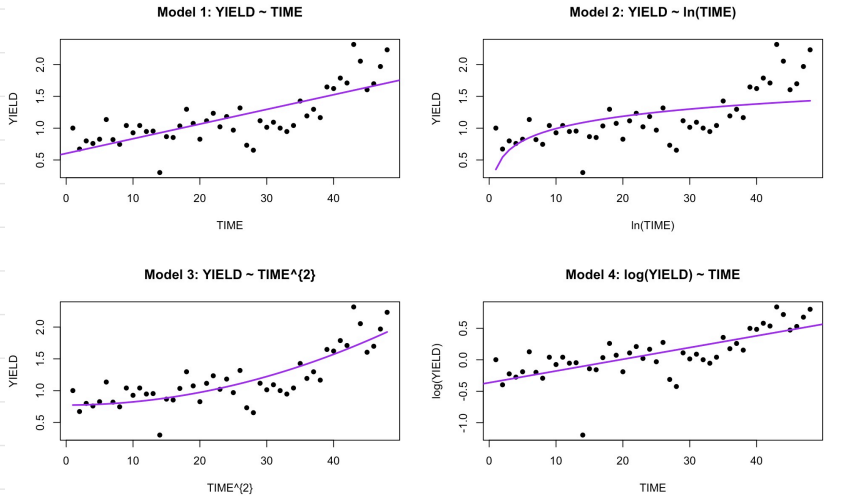
a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.

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

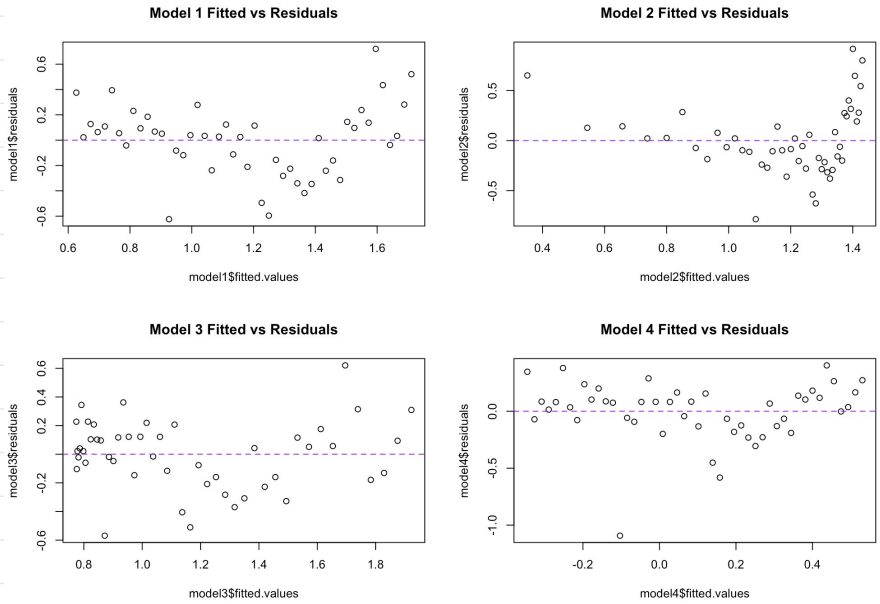
c. Using your chosen specification, identify any unusual observations, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*.

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?

(SOL) a. (i)



(ii)



(iii)

```
Model 1:

        Shapiro-Wilk normality test

data:  resid(model1)
W = 0.98236, p-value = 0.6792
```

> 0.05

```
Model 2:

        Shapiro-Wilk normality test

data:  resid(model2)
W = 0.96657, p-value = 0.1856
```

> 0.05

```
Model 3:

        Shapiro-Wilk normality test

data:  resid(model3)
W = 0.98589, p-value = 0.8266
```

> 0.05

```
Model 4:

        Shapiro-Wilk normality test

data:  resid(model4)
W = 0.86894, p-value = 7.205e-05
```

< 0.05

Only the error in Model 4 not satisfy the normality assumption.

(iv)

```
> cat("R2 for Model 1: ", summary(model1)$r.squared, "\n")
R2 for Model 1:  0.5778369
> cat("R2 for Model 2: ", summary(model2)$r.squared, "\n")
R2 for Model 2:  0.3385733
> cat("R2 for Model 3: ", summary(model3)$r.squared, "\n")
R2 for Model 3:  0.6890101
> cat("R2 for Model 4: ", summary(model4)$r.squared, "\n")
R2 for Model 4:  0.5073566
```

Model 3, because it has highest R^2

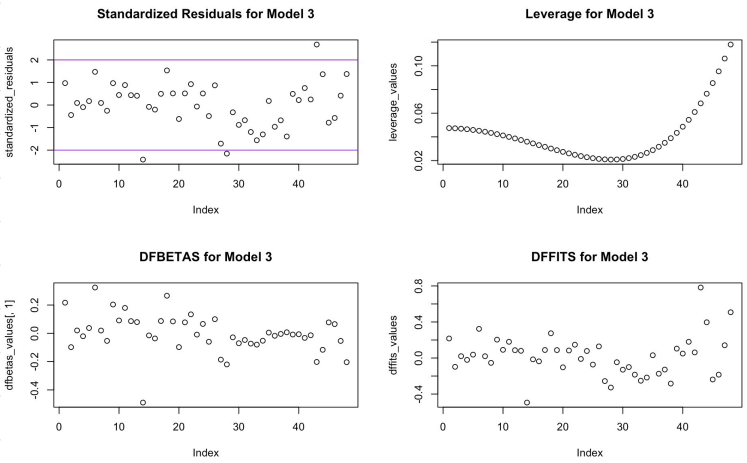
and its error satisfy normality assumption.

b、 In Model 3, the time variable is a quadratic term, indicating that wheat yield exhibits a certain nonlinear change over time.

If $\gamma_1 > 0$, the increase in wheat yield accelerates over time (showing positive accelerated growth).

If $\gamma_1 < 0$, the growth of wheat yield slows down over time and may even reach saturation or show a decline.

c、



Based on the analysis of standardized residuals, leverage values, DFBETAS, and DFFITS, Model 3 does not show any obvious influential observations. Although some data points exhibit relatively high influence in the leverage and DFFITS plots, there are no points that clearly exceed the conventional thresholds. Therefore, it can be concluded that these observations are unlikely to have a significant impact on the results of the model.

d、1997年小麥產量的 95% C.I. 為 [1.412563, 2.432401]

1997年小麥產量預測：1.922482 ∈ [1.412563, 2.432401]

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.

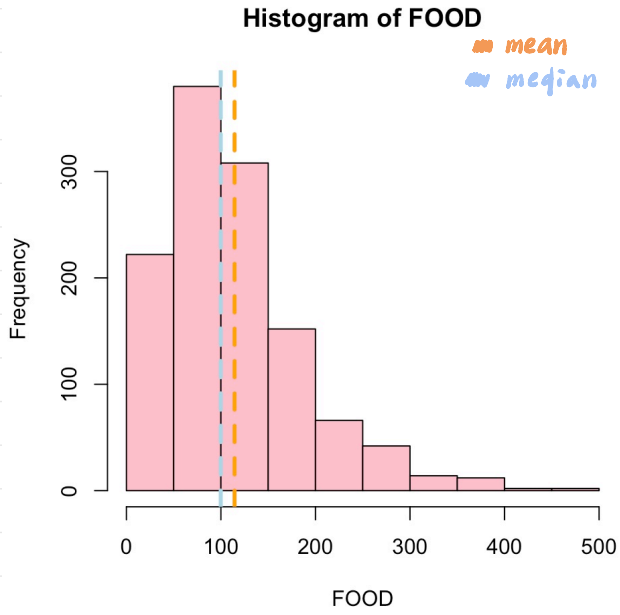
j. Obtain the least squares residuals from the linear-log model and plot them against $\ln(INCOME)$. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

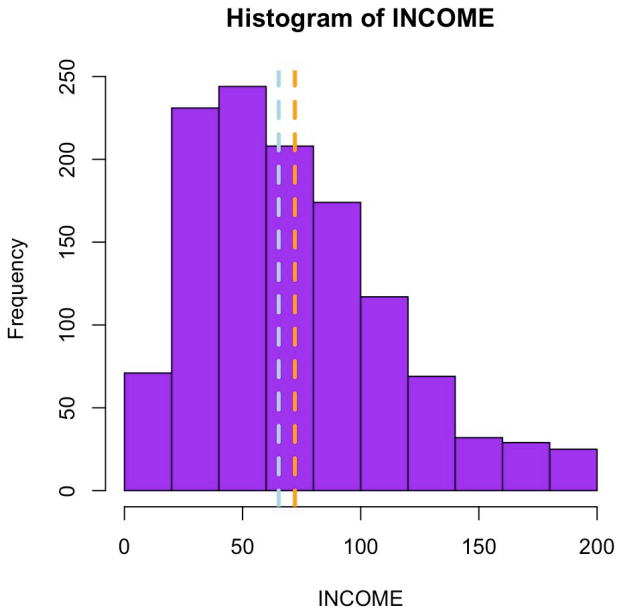
k. Based on this exercise, do you prefer the linear relationship model, or the log-log model or the linear-log model? Explain your reasoning.

isolja a.

```
> cat("FOOD summary:\n")
FOOD summary:
> cat("Mean: ", food_mean, "\n")
Mean:  114.4431
> cat("Median: ", food_median, "\n")
Median:  99.8
> cat("Min: ", food_minimum, "\n")
Min:  9.63
> cat("Max: ", food_maximum, "\n")
Max:  476.67
> cat("Standard Deviation: ", food_standard_deviation, "\n\n")
Standard Deviation:  72.6575
```

```
> cat("INCOME summary:\n")
INCOME summary:
> cat("Mean: ", income_mean, "\n")
Mean:  72.14264
> cat("Median: ", income_median, "\n")
Median:  65.29
> cat("Min: ", income_minimum, "\n")
Min:  10
> cat("Max: ", income_maximum, "\n")
Max:  200
> cat("Standard Deviation: ", income_standard_deviation, "\n\n")
Standard Deviation:  41.65228
```





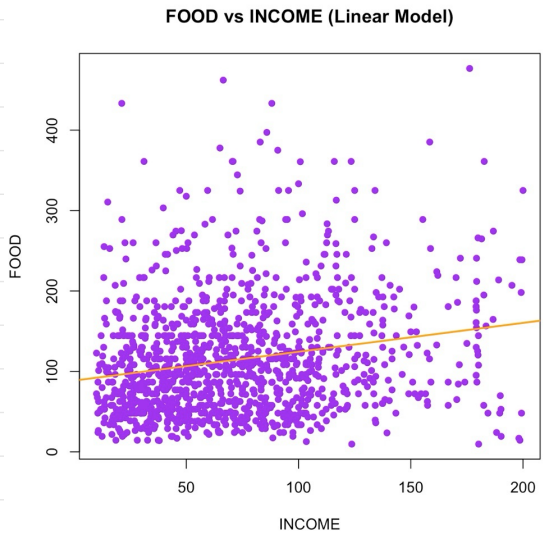
mean > median

None of them looks like bell-shaped.

```
> print(jarque.bera.test(cex5_small$food))

        Jarque Bera Test

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

> print(jarque.bera.test(cex5_small$income))

        Jarque Bera Test

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

Based on the results of the Jarque-Bera test, for both the FOOD and INCOME variables, the p-value is less than 0.05. Therefore, we reject the null hypothesis, indicating that the residuals of both variables do not follow a normal distribution.

b.



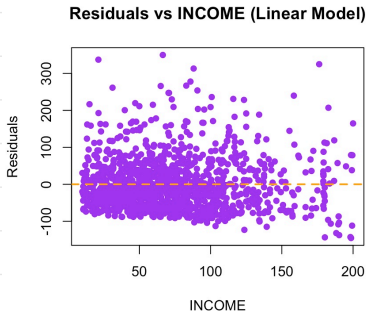

```
Call:
lm(formula = food ~ income, data = cex5_small)

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
```

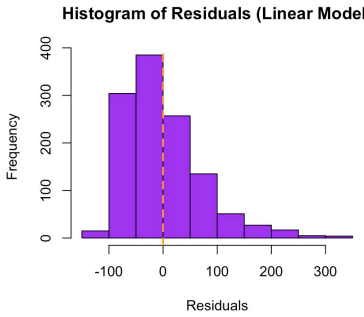
```
> confint(model_linear, level = 0.95)
                 2.5 %    97.5 %
(Intercept) 80.5064570 96.626543
income       0.2619215  0.455452
```

We estimated the linear model: FOOD = 88.56650 + 0.35869 * INCOME

The coefficient of the income variable, β_2, is 0.35869. At the 95% confidence level, its confidence interval is [0.2619, 0.4555]. The p-value for this coefficient is 6.36e-13, indicating that income has a significant positive effect on food expenditure. Since the confidence interval is relatively narrow, we consider the estimate of the effect of income changes on food expenditure to be relatively precise. However, the model's R^2 is only 0.04228, which means income explains only a small portion of the variation in food expenditure. Therefore, the overall explanatory power of the model is limited.

c.






```
        Jarque Bera Test

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

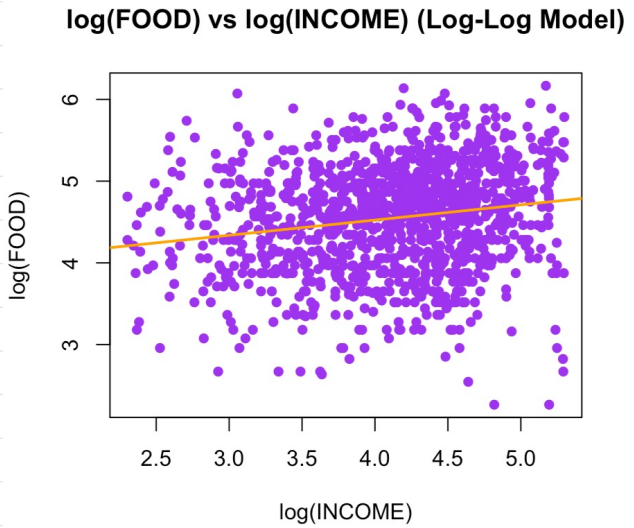
In the scatter plot of residuals versus INCOME, we observe that as income increases, the variability of the residuals also increases, indicating heteroskedasticity. The histogram of the residuals shows a skewed and asymmetric distribution. The result of the Jarque-Bera test, with a p-value less than 2.2e-16, indicates that the residuals do not follow a normal distribution. This suggests that the model violates the fundamental assumptions of homoscedasticity and normality of the error terms in regression analysis. Although the distribution of explanatory variables is important, the normality of the error term e is the key condition for conducting valid inferential analysis in regression models.

d.

```
> elasticity_table
  INCOME FOOD_Predicted Elasticity Lower_95_CI Upper_95_CI
1     19          95.38     0.0715      0.0522      0.0907
2     65         111.88     0.2084      0.1522      0.2646
3    160         145.96     0.3932      0.2871      0.4993
```

According to the estimated results of the model, at different income levels (INCOME = 19, 65, 160), the elasticities of food expenditure are 0.0715, 0.2084, and 0.3932, respectively. The 95% confidence intervals for these elasticity estimates are [0.0522, 0.0907], [0.1522, 0.2646], and [0.2871, 0.4993]. The elasticity increases as income rises, indicating that higher-income households are more responsive in their food expenditures to changes in income. This trend aligns with economic theory regarding the demand elasticity of necessities: for low-income groups, food is a necessity with relatively low demand elasticity; as income increases, the demand elasticity for food expenditure also increases, although it remains less than 1 overall. This implies that food expenditure reacts relatively steadily to changes in income, consistent with the characteristics of basic necessity goods.

e.




```
Call:
lm(formula = log_food ~ log_income, data = cex5_small)

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
```

```
> cat("R-squared of Linear Model: ", r2_linear, "\n")
R-squared of Linear Model:  0.0422812
> cat("R-squared of Log-Log Model: ", r2_loglog, "\n")
R-squared of Log-Log Model:  0.03322915
```

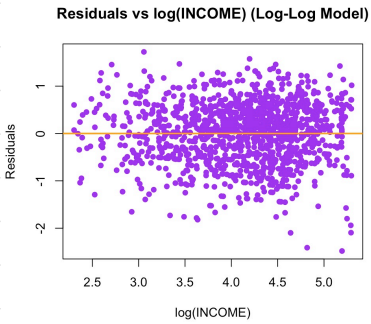
We estimated the log-log model, which shows that the coefficient of ln(INCOME) is 0.18631. This indicates that for every 1% increase in income, food expenditure increases by approximately 0.186%. However, the R^2 of the log-log model is only 0.03323, which is lower than the R^2 of the linear model at 0.04228. This suggests that the overall explanatory power of the log-log model is weaker and it cannot explain the variation in food expenditure more effectively. Although the log-log model provides an intuitive interpretation of elasticity, based on the goodness of fit and explanatory power, the linear model is slightly better than the log-log model. Nevertheless, both models have relatively low R^2 values, indicating that it may be necessary to include more explanatory variables or adjust the model specification to improve predictive ability.

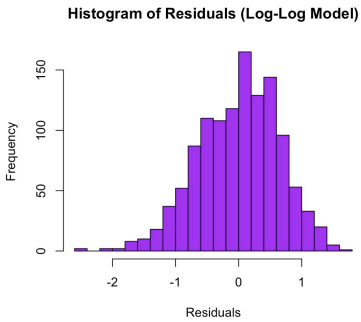
f.

```
> confint(model_loglog, level = 0.95)
                2.5 %    97.5 %
(Intercept) 3.5428135 4.0150507
log_income  0.1293432 0.2432675
```

In the log-log model, the income elasticity of food expenditure is 0.18631, with a 95% confidence interval of [0.12934, 0.24327]. This indicates that food is a necessity with an income elasticity of less than 1, meaning that as income increases, food expenditure increases by a relatively small proportion. Compared to the linear model, the log-log model provides a fixed elasticity value that does not vary with different income levels. However, the linear model shows higher income elasticity for higher-income groups, suggesting different interpretations of elasticity variation between the two models. Overall, the log-log model offers a stable and easy-to-interpret result but may underestimate the responsiveness of food expenditure to income increases among high-income households.

g.





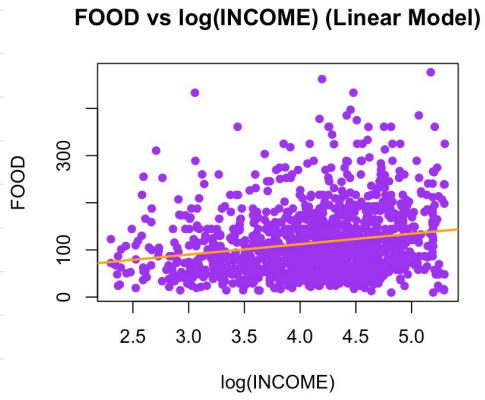

```
        Jarque Bera Test

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

In the log-log model, the scatter plot of residuals against ln(INCOME) shows that, although the residuals are mostly centered around 0, there is still heteroscedasticity, meaning that the model has not perfectly captured the variation in the data, and the variance of the residuals changes with ln(INCOME). The histogram of the residuals exhibits a negatively skewed distribution, indicating that most of the residuals are concentrated in the negative region, with a long tail in the positive region.

Further performing the Jarque-Bera test, the result shows a p-value < 0.05, so we reject the null hypothesis, indicating that the residuals do not follow a normal distribution. This suggests that the error terms in the model do not satisfy the normality assumption required for regression analysis.

In summary, although the log-log model provides some explanatory power, its error terms do not meet the normality assumption and exhibit heteroscedasticity. Therefore, the model may need further adjustments or the inclusion of additional variables to improve the accuracy of predictions.

h.




```
> cat("R-squared of Linear Model with log(INCOME): ", r2_linear_log, "\n")
R-squared of Linear Model with log(INCOME):  0.03799984
> cat("R-squared of Log-Log Model: ", r2_loglog, "\n")
R-squared of Log-Log Model:  0.03322915
```

In the linear regression model of FOOD and ln(INCOME), the R^2 value is 0.03799, indicating that the model has weak explanatory power, explaining only 3.8% of the variation. Compared to the log-log model, the linear model has a slightly higher R^2 of 0.03323, suggesting that the linear model has a slight advantage in explaining the variation in food expenditure. Nevertheless, both models have relatively low R^2 values, which implies that further improvements may be needed, such as incorporating more explanatory variables or adjusting the model structure to enhance predictive accuracy.

i.

```
> logelasticity_table
  INCOME FOOD_Predicted Elasticity Lower_95_CI Upper_95_CI
1     19          88.90     0.2496      0.1784      0.3208
2     65         116.19     0.1910      0.1365      0.2454
3    160         136.17     0.1629      0.1165      0.2094
```

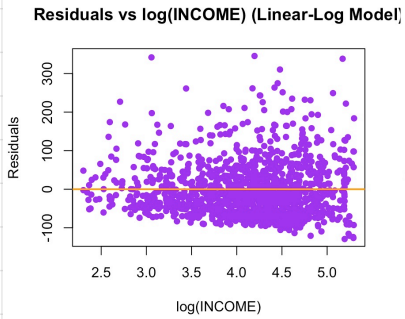
Based on the linear-log model, we estimated the food expenditure elasticity at three income levels: INCOME = 19, 65, and 160. For INCOME = 19, the elasticity is 0.2496; for INCOME = 65, it is 0.1910; and for INCOME = 160, it is 0.1629. This shows that as income increases, the response of food expenditure to income decreases.

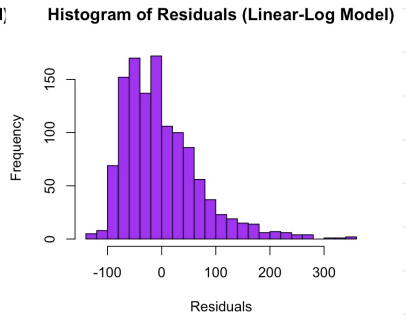
Each elasticity estimate is accompanied by a 95% prediction interval. For example, for INCOME = 19, the elasticity is 0.2496 with a prediction interval of [0.1784, 0.3208], indicating some uncertainty in the estimate.

When comparing these results to other models (such as the log-log model), it is evident that the linear-log model provides an elasticity estimate that varies with income, while the log-log model offers a more stable and fixed elasticity estimate.

Overall, the linear-log model is better suited for describing food expenditure behavior in low-income households, while the log-log model is more stable and applicable to a wider range of incomes. These results align with the economic theory that the elasticity of demand for necessities varies with income.

j.






```
        Jarque Bera Test

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

Based on the results of the linear-log model, we analyzed the scatter plot and histogram of the residuals against log(INCOME). In the scatter plot, although most of the residuals are concentrated around zero, we notice the presence of heteroskedasticity, meaning that the residuals show varying levels of volatility as log(INCOME) changes.

The histogram of the residuals shows a right-skewed distribution, indicating that most residuals are concentrated in the negative region, with a longer tail in the positive region. This suggests that the variability of the model's errors is greater in certain intervals.

According to the Jarque-Bera test, the result shows a p-value of less than 2.2e-16, which leads us to reject the null hypothesis, concluding that the residuals do not follow a normal distribution. This further indicates that the model's error terms do not meet the normality assumption required in regression analysis, which may affect the accuracy of the predictions.

In summary, while the linear-log model provides some explanatory power in describing the relationship between income and food expenditure, the non-normal distribution of residuals and heteroskedasticity suggest that the model requires further adjustment. It is recommended to modify the model or incorporate additional explanatory variables in subsequent analyses to improve prediction accuracy.

k.

Based on the results from sections a to j of question 4.29, we observe that the models show different strengths and limitations in explaining the relationship between household income and food expenditure. The linear model and the log-log model both exhibit relatively low R-squared values, indicating that neither model fully captures the variation in food expenditure. In particular, the residuals in both models display heteroskedasticity and fail the normality assumption, as shown by the Jarque-Bera test.

The linear-log model provides more flexibility by allowing the elasticity of food expenditure to vary with income, whereas the log-log model offers a more stable elasticity estimate, which is consistent across income levels. However, both models suggest that income has a positive impact on food expenditure, with elasticity estimates decreasing as income increases, particularly for the linear-log model.

Given the limitations of both models, it would be advisable to consider alternative model specifications, potentially incorporating additional variables or adjusting for heteroskedasticity, to improve prediction accuracy and better capture the dynamics of food expenditure.