Full name: Nguyen Nhut Vu Truong

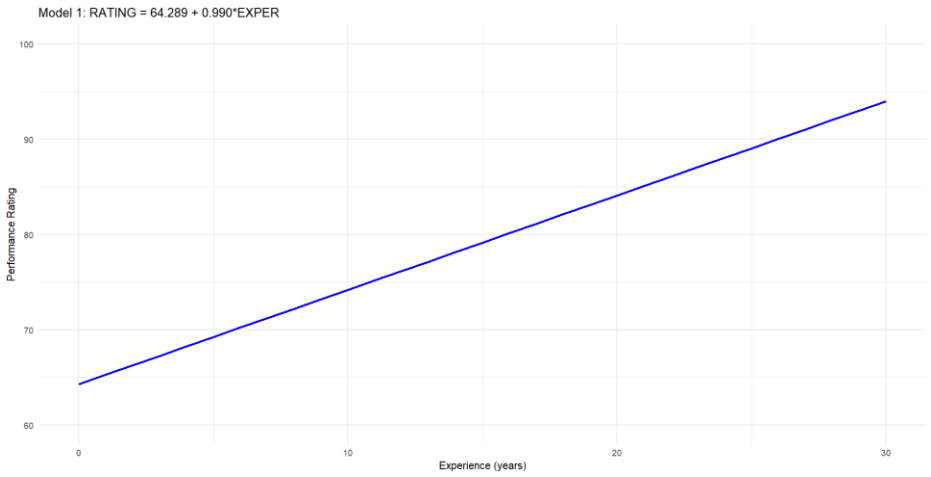
Student ID: 413707008

Course: Financial Econometrics

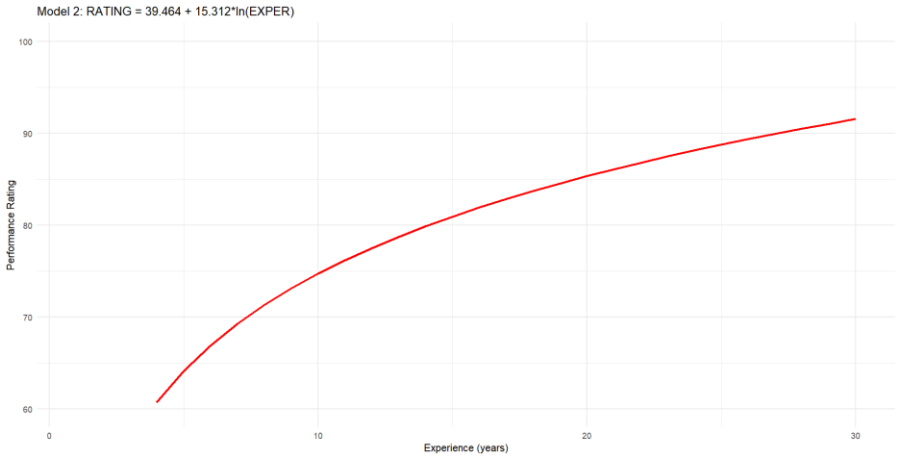
HW0317

Question 4

a.



b.



Why four artists with no experience are excluded from Model 2: The natural logarithm of zero (ln(0)) is undefined. Since Model 2 uses ln(EXPER), artists with zero experience cannot be included in the estimation.

c.

In Model 1, the marginal effect is the coefficient of EXPER, which is constant at 0.990 regardless of experience level:

(i) For an artist with 10 years of experience: 0.990

(ii) For an artist with 20 years of experience: 0.990

This means each additional year of experience increases the expected rating by 0.990 points, regardless of how much experience the artist already has.

d.

(i) For an artist with 10 years of experience: $15.312 \times (1/10) = 1.5312$

(ii) For an artist with 20 years of experience: $15.312 \times (1/20) = 0.7656$

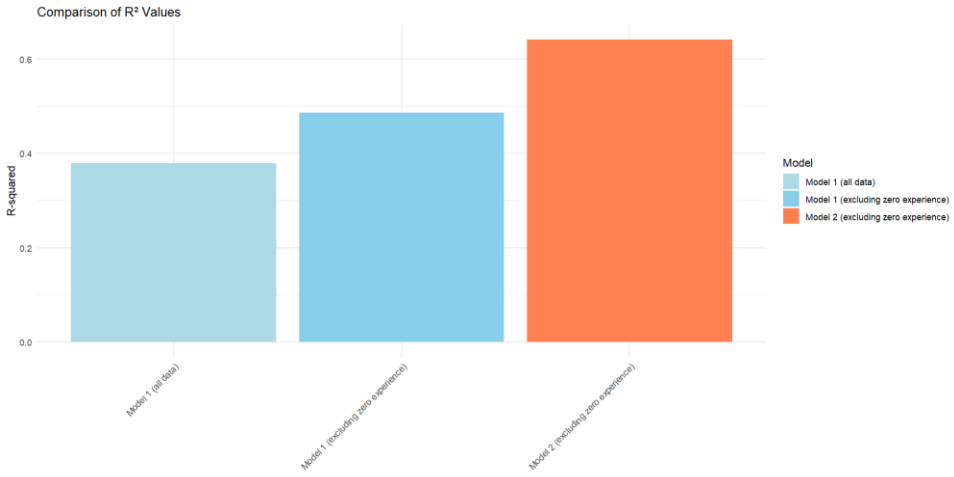
In Model 2, the marginal effect decreases as experience increases, showing that each additional year provides less benefit for more experienced artists.

e.

To evaluate which model fits the data better, we compare their R^2 values:

- Model 1 (all data): $R^2 = 0.3793$
- Model 1 (only artists with experience): $R^2 = 0.4858$
- Model 2: $R^2 = 0.6414$

Model 2 has the highest R^2 value (0.6414), indicating it explains about 64% of the variation in performance ratings, compared to only about 38-49% for Model 1. This suggests Model 2 provides a better fit to the observed data.



f.

From an economic perspective, Model 2 (logarithmic) is more plausible for several reasons:

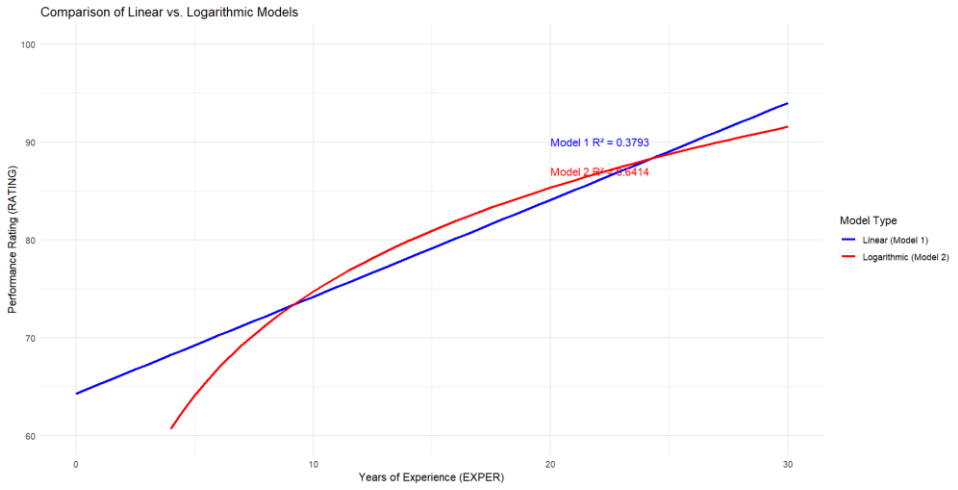
1. **Diminishing returns**: In most professions, the effect of additional experience typically diminishes over time. Model 2 captures this economic principle with its logarithmic form, while Model 1 unrealistically implies constant returns to experience.
2. **Learning curve**: Professionals generally improve rapidly early in their careers and then plateau. Model 2 correctly shows steeper improvements for less experienced artists that gradually level off.

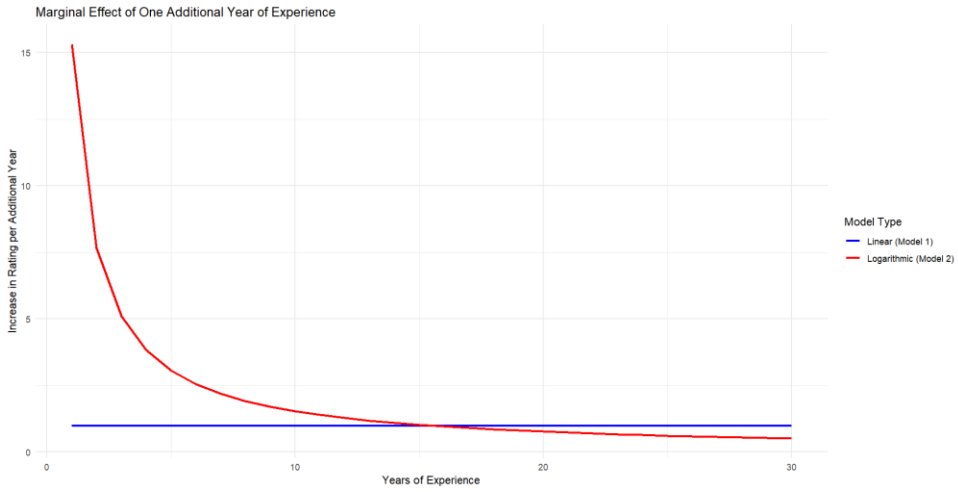
3. **Skill development**: Technical skills often follow a pattern where fundamental abilities develop quickly, but mastery takes much longer. The logarithmic curve in Model 2 better represents this pattern.

4. **Statistical evidence**: The higher R^2 for Model 2 supports the economic theory that the relationship between experience and performance is non-linear.

5. **Ceiling effect**: Performance ratings have a natural ceiling (100 points in this case). The logarithmic model approaches an upper limit as experience increases, which is more realistic than the linear model that would eventually predict ratings exceeding the maximum possible score. For example, Model 1 would predict ratings above 100 for artists with more than 36 years of experience, which is implausible.

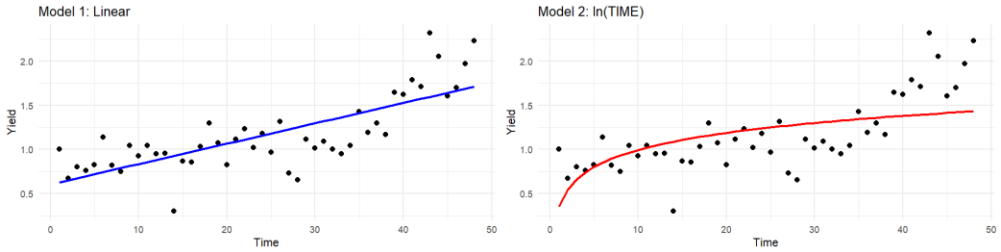
In conclusion, Model 2 not only fits the data better statistically but also aligns with economic theories of human capital development and learning curves, making it the more reasonable model.

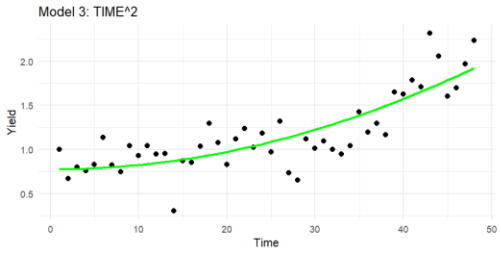


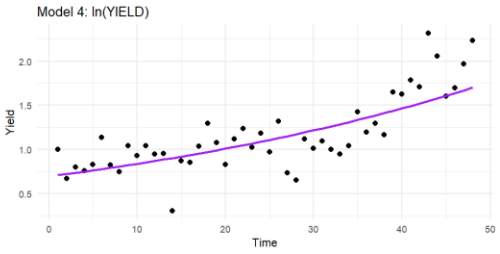


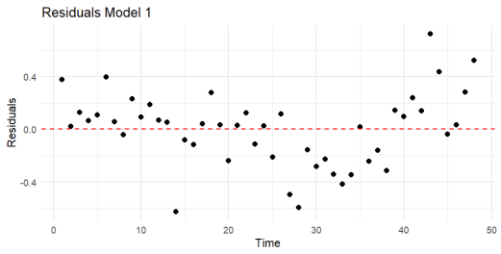
Question 28

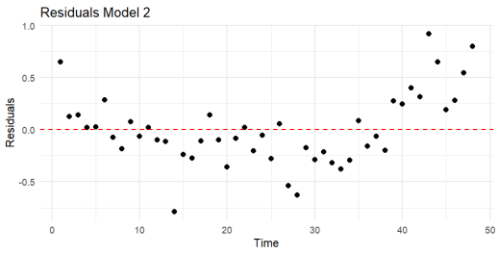
a.

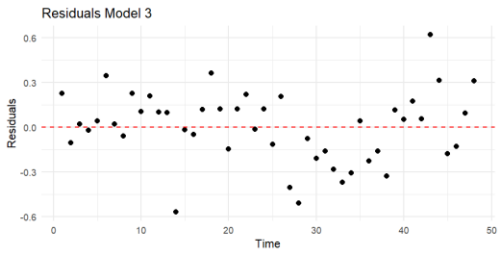


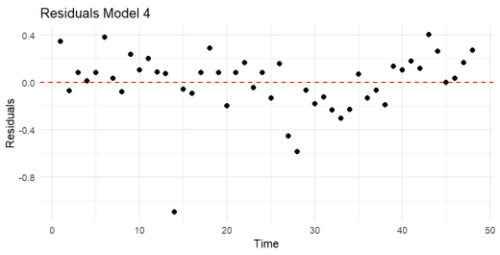












```
> # Print results
> cat("Shapiro-Wilk p-values:\n")
Shapiro-Wilk p-values:
> cat("Model 1:", shapiro1$p.value, "\n")
Model 1: 0.6792056
> cat("Model 2:", shapiro2$p.value, "\n")
Model 2: 0.1855502
> cat("Model 3:", shapiro3$p.value, "\n")
Model 3: 0.826645
> cat("Model 4:", shapiro4$p.value, "\n")
Model 4: 7.205319e-05
>
> cat("\nAdjusted R-squared:\n")

Adjusted R-squared:
> cat("Model 1:", adj_r2[1], "\n")
Model 1: 0.5686594
> cat("Model 2:", adj_r2[2], "\n")
Model 2: 0.3241945
> cat("Model 3:", adj_r2[3], "\n")
Model 3: 0.6822494
> cat("Model 4:", adj_r2[4], "\n")
Model 4: 0.4966469
```

Model 3 (TIME²) appears to be the best model:

- Best Adjusted R^2 → explains most variance.
- Best residual distribution → passes Shapiro-Wilk normality test (p = 0.8266).
- Visual fit and residuals also support this.

b.

- The derivative of $YIELD$ with respect to $TIME$ is:

$$\frac{d(YIELD)}{d(TIME)} = 2\gamma_1 TIME$$

- This means the rate of change in $YIELD$ per year increases linearly with $TIME$. For example:
  - At $TIME = 1$ (1950), the rate of change in $YIELD$ is $2 \times 4.986 \times 10^{-4} \times 1 = 0.0009972$ units per year.
  - At $TIME = 48$ (1997), the rate of change is $2 \times 4.986 \times 10^{-4} \times 48 = 0.047866$ units per year.
- This indicates that the yield increases faster as time progresses, which aligns with a quadratic trend (e.g., technological improvements in agriculture might lead to accelerating yield growth).

In practical terms:

- The coefficient $\gamma_1 = 4.986 \times 10^{-4}$ means that for each unit increase in $TIME^2$, the wheat yield increases by 0.0004986 units. However, since $TIME^2$ grows quadratically, the actual increase in $YIELD$ becomes more significant in later years.

c.

```
> # Output results
> cat("Unusual observations (indices):", outliers, "\n")
Unusual observations (indices): 6 14 28 43 44 45 46 47 48
> cat("Corresponding years:", 1950 + outliers - 1, "\n")
Corresponding years: 1955 1963 1977 1992 1993 1994 1995 1996 1997
```

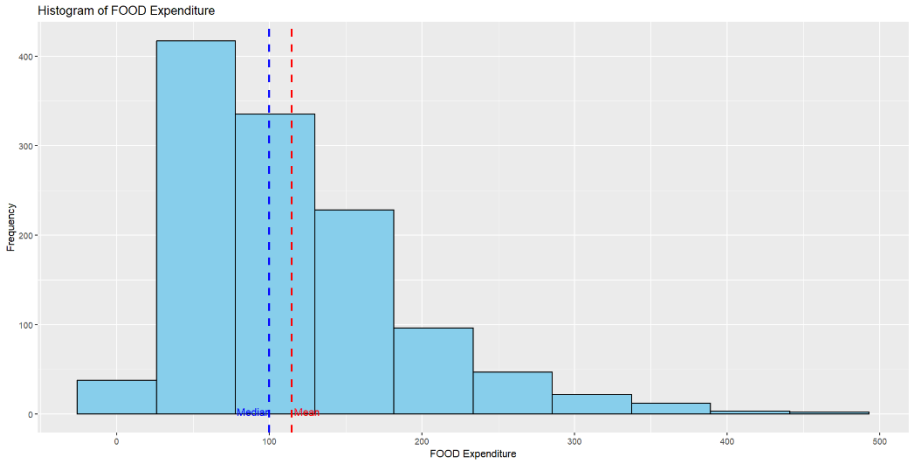
d.

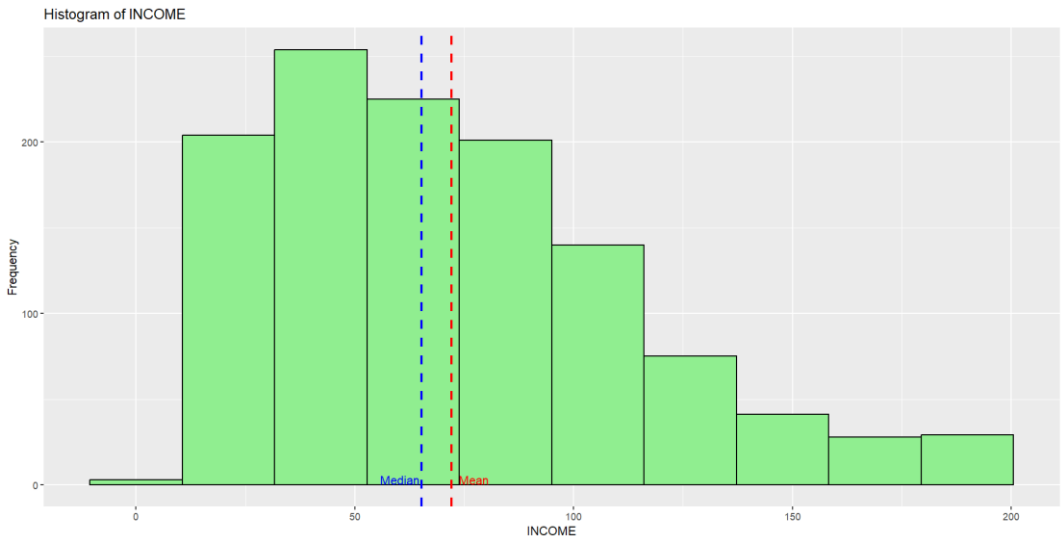
```
> # Print prediction interval and check if true value is within it
> cat("95% Prediction Interval for YIELD in 1997:\n")
95% Prediction Interval for YIELD in 1997:
> cat("  Lower bound:", pred[1, "lwr"], "\n")
  Lower bound: 1.372403
> cat("  Upper bound:", pred[1, "upr"], "\n")
  Upper bound: 2.389819
> cat("  Predicted YIELD:", pred[1, "fit"], "\n")
  Predicted YIELD: 1.881111
> cat("  True YIELD in 1997:", true_yield_1997, "\n")
  True YIELD in 1997: 2.2318
> cat("Does the interval contain the true value? ",
+     true_yield_1997 >= pred[1, "lwr"] & true_yield_1997 <= pred[1, "upr"],
+     "\n")
Does the interval contain the true value?  TRUE
```

Question 29

a. Summary statistics and histograms

```
> cat("Summary statistics for FOOD:\n")
Summary statistics for FOOD:
> print(food_summary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.63   57.78   99.80  114.44  145.00  476.67
> cat("Standard deviation:", food_sd, "\n\n")
Standard deviation: 72.6575

>
> cat("Summary statistics for INCOME:\n")
Summary statistics for INCOME:
> print(income_summary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   40.00   65.29   72.14   96.79  200.00
> cat("Standard deviation:", income_sd, "\n\n")
Standard deviation: 41.65228
```



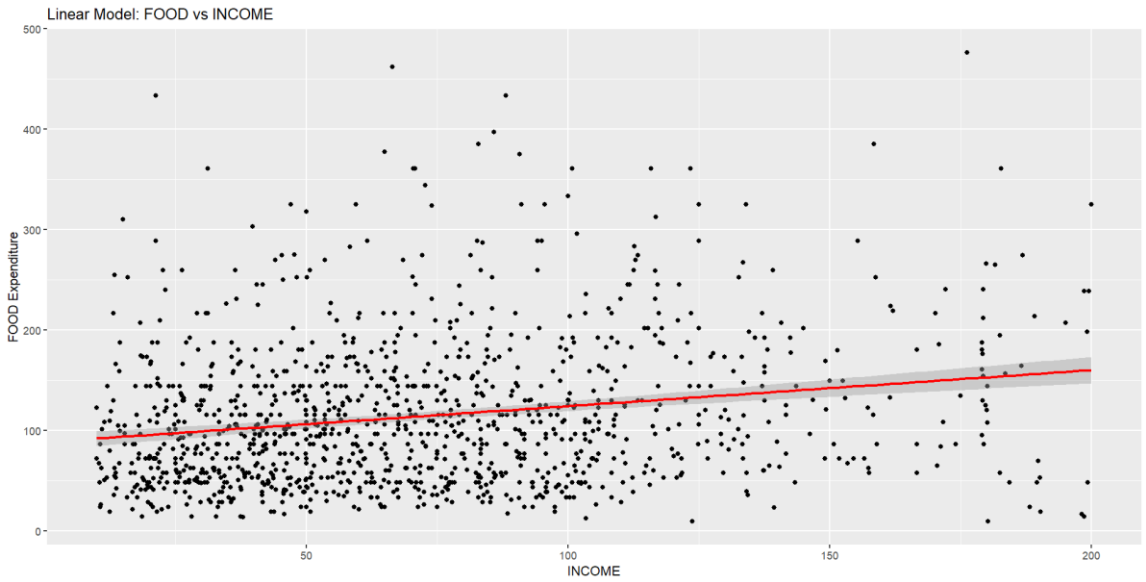



```
> cat("Jarque-Bera test for FOOD:\n")
Jarque-Bera test for FOOD:
> print(jb_food)

        Jarque Bera Test

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

> cat("\nJarque-Bera test for INCOME:\n")

Jarque-Bera test for INCOME:
> print(jb_income)

        Jarque Bera Test

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

Both distributions are positively skewed with mean's greater than medians. They are not bell shaped or symmetrical. For INCOME the Jarque-Bera statistic is 148.21 and for FOOD expenditure it is 648.65. The critical value for a test at the 5% level is 5.99. We reject the null hypothesis of normality for each variable.

b.

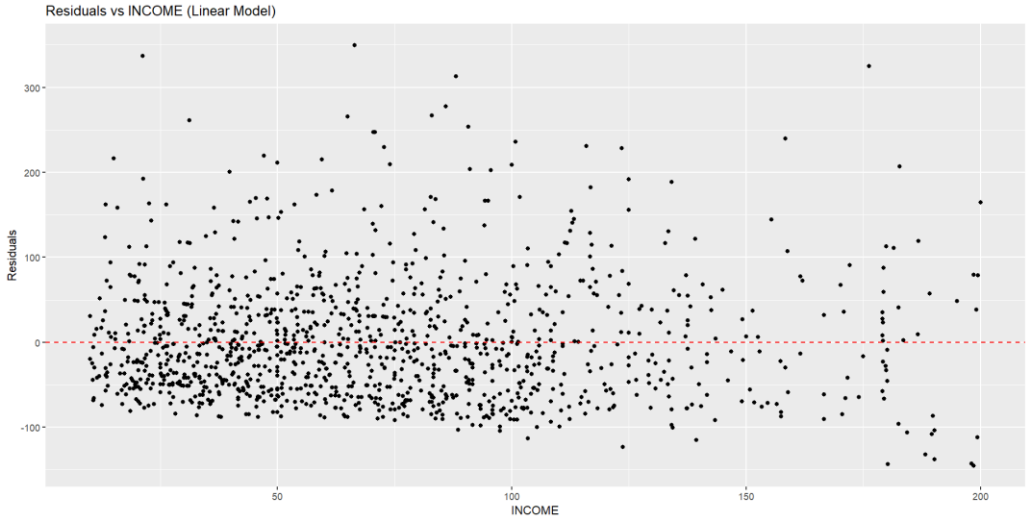
```
Linear Model: FOOD = β₁ + β₂*INCOME + e
> print(summary_linear)

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
```



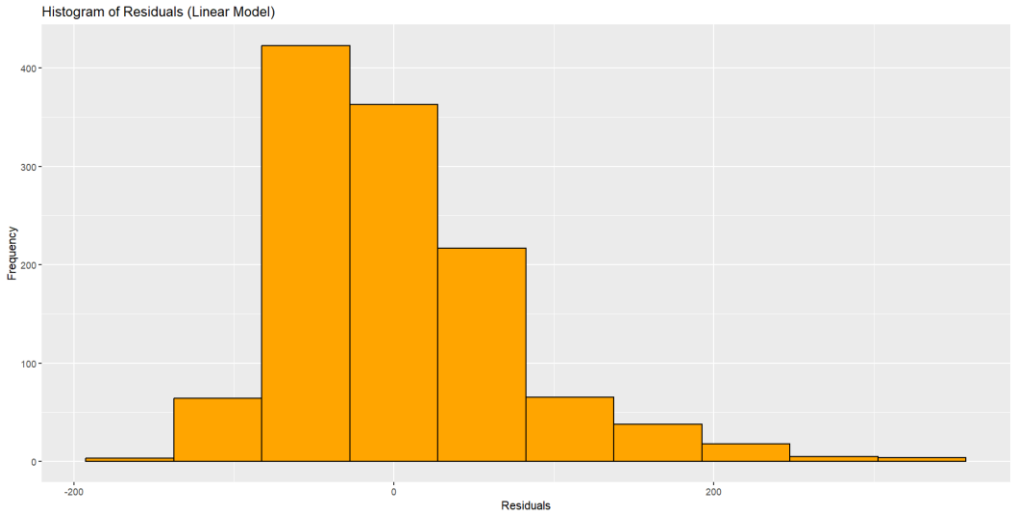
95% Confidence Interval for β_2: 0.2619215 to 0.455452

We have not estimated the effect of changing income on average food consumption very precisely: the R^2 is low, and the confidence interval is wide.

c.



The least squares residuals are plotted above. The positive skew at each income is clear. There is not a clear "spray" pattern except at high incomes. The residual histogram shows the skewness. The Jarque-Bera statistic is 624.186, which is far greater than the 5% critical value 5.99.




```
Jarque-Bera test for residuals (Linear Model):
> print(jb_residuals_linear)

        Jarque Bera Test

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

It is more important for the random error e to be normally distributed than for the variables FOOD and INCOME to be normally distributed. This is because the statistical inference (hypothesis tests, confidence intervals) in regression analysis relies on the assumption that the error term follows a normal distribution, not the variables themselves.

d.

```
> cat("\nElasticity estimates for Linear Model:\n")

Elasticity estimates for Linear Model:
> print(elasticity_results_linear)
  INCOME FOOD_Predicted Elasticity   CI_Lower   CI_Upper
1     19       95.38155 0.07145038 0.05217475 0.09072601
2     65      111.88114 0.20838756 0.15216951 0.26460562
3    160      145.95638 0.39319883 0.28712305 0.49927462
```

The estimated elasticities are **dissimilar** — they **increase noticeably** as income increases.

The interval estimates **do not overlap**, supporting the idea that the elasticities are **meaningfully different** across income levels.

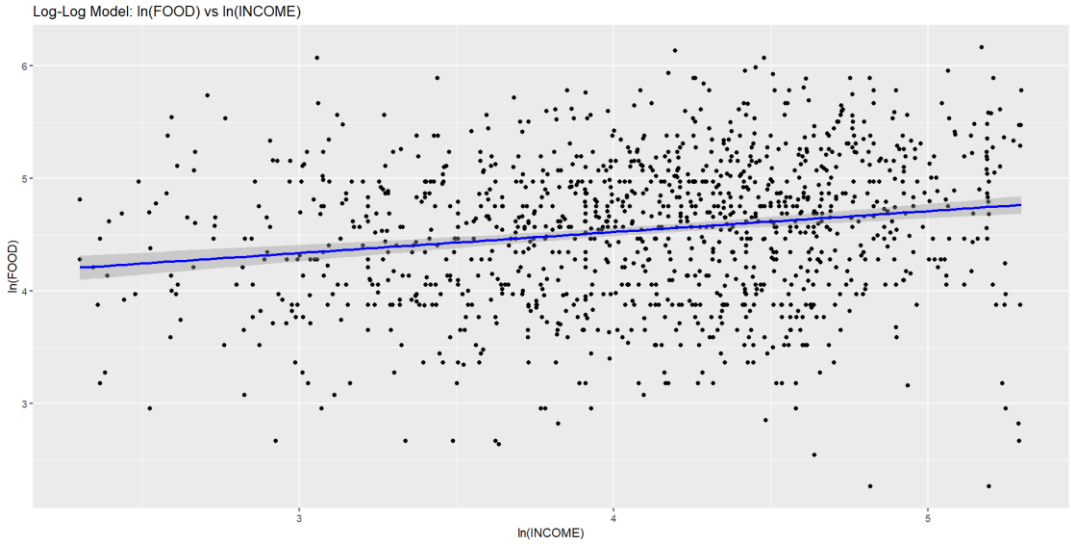
Based on economic principles, the income elasticity for food should decrease as income increases, as food is a necessity. Higher income households spend a smaller proportion of additional income on food compared to lower income households.

But in this data, it's increasing, possibly reflecting low-to-middle income levels, where people still expand their food consumption as income rises.

e.

```
Log-Log Model: ln(FOOD) = γ₁ + γ₂*ln(INCOME) + e
> print(summary_log_log)

Call:
lm(formula = ln_food ~ ln_income, data = cex5_log)

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

> generalized_r2_log_log <- cor(actual_food, predictions_original_scale)^2
>
> cat("Generalized R² for log-log model:", generalized_r2_log_log, "\n")
Generalized R² for log-log model: 0.03965161
```

The generalized R2 is 0.03965 which is slightly smaller than the R2 from the linear model (0.04228). The log-log model provides a valid functional form but does not improve model fit over the linear specification in this case. Based on both the R^2 comparison and visual inspection of the scatter plots, the linear model is slightly preferred.



f.

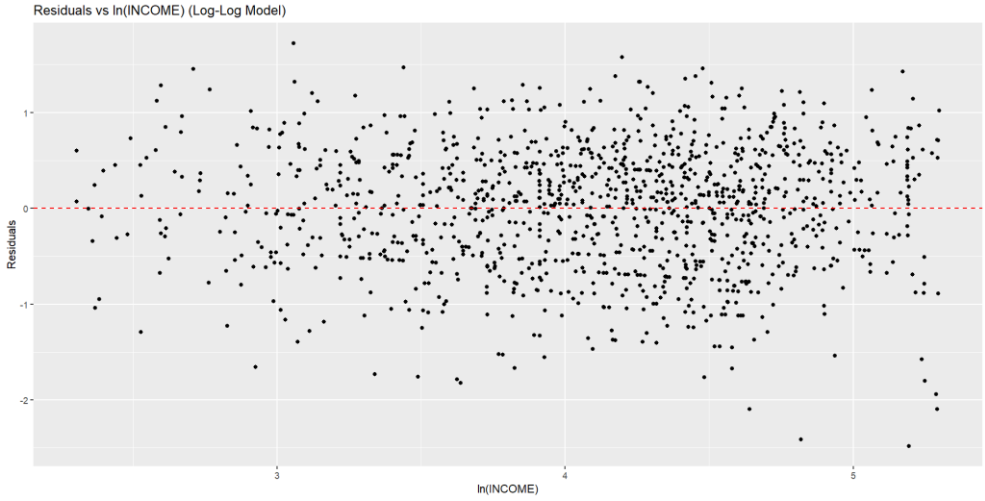
Elasticity estimate for Log-Log Model: 0.1863054

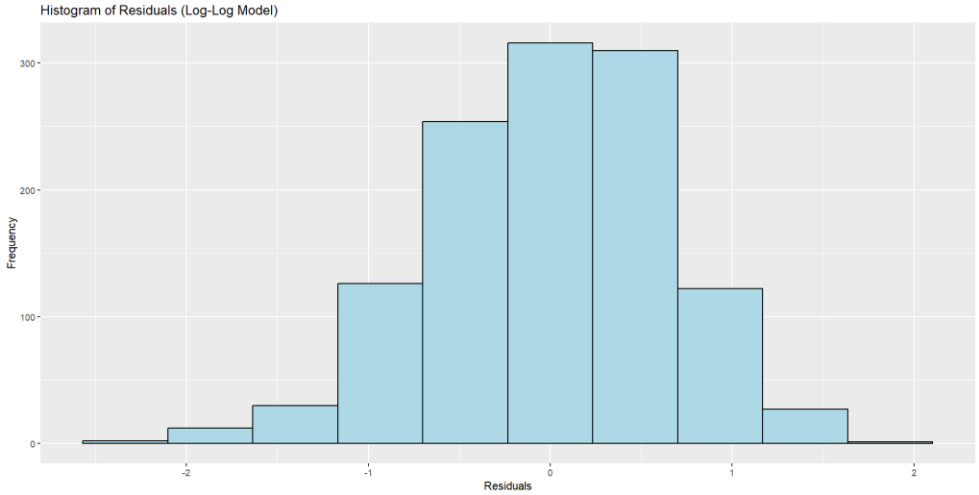
95% Confidence Interval for elasticity: 0.1293432 to 0.2432675

The elasticity estimate from the log-log model is **statistically similar** to that of the linear model at moderate income levels (e.g. INCOME = 65). This is supported by the **overlapping confidence intervals**, indicating no significant difference in estimated elasticity at that point.

```
> # Compare with elasticities from linear model
> cat("\nComparing elasticities from Linear and Log-Log models:\n")

Comparing elasticities from Linear and Log-Log models:
> cat("Log-Log model elasticity (constant):", elasticity_log_log, "\n")
Log-Log model elasticity (constant): 0.1863054
> cat("Linear model elasticity at INCOME = 19:", elasticity_linear[1], "\n")
Linear model elasticity at INCOME = 19: 0.07145038
> cat("Linear model elasticity at INCOME = 65:", elasticity_linear[2], "\n")
Linear model elasticity at INCOME = 65: 0.2083876
> cat("Linear model elasticity at INCOME = 160:", elasticity_linear[3], "\n")
Linear model elasticity at INCOME = 160: 0.3931988
>
> # Statistical evidence
> cat("\nStatistical evidence for comparison:\n")

Statistical evidence for comparison:
> cat("The elasticity from the log-log model is",
+     ifelse(elasticity_log_log >= min(elasticity_linear) & elasticity_log_log <= max(elasticity_linear),
+            "within the range of",
+            "outside the range of"),
+     "elasticities calculated from the linear model.\n")
The elasticity from the log-log model is within the range of elasticities calculated from the linear model.
```

g.





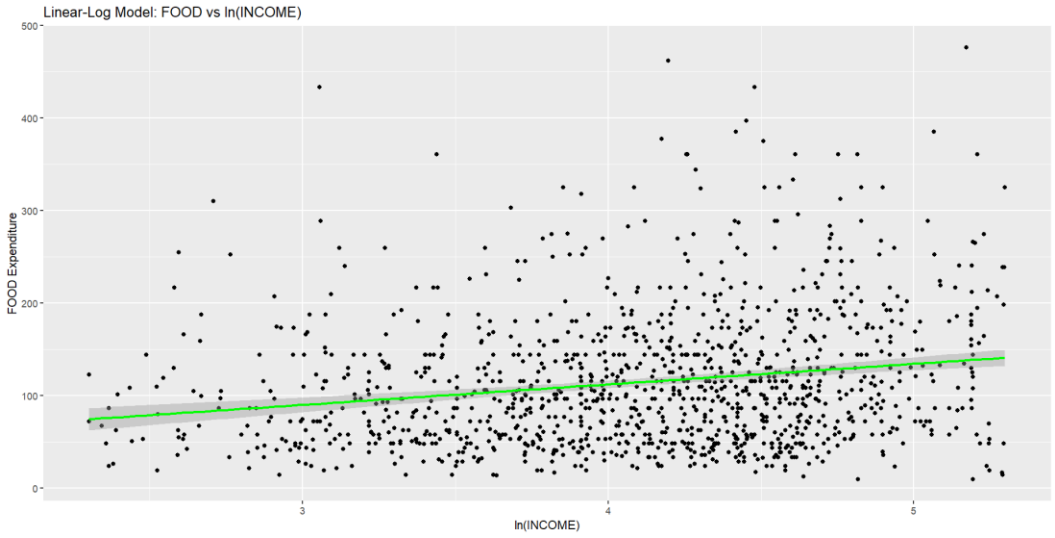

```
> # Jarque-Bera test for residuals of log-log model
> jb_residuals_log_log <- jarque.bera.test(residuals_log_log)
> cat("\nJarque-Bera test for residuals (Log-Log Model):\n")

Jarque-Bera test for residuals (Log-Log Model):
> print(jb_residuals_log_log)

        Jarque Bera Test

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

The residuals from the log-log model show no major violations of regression assumptions. The scatter appears random and symmetric, with no obvious patterns and only a slight suggestion of heteroskedasticity. This supports the choice of the **log-log model** as a statistically well-behaved specification relative to the alternatives.

There is a slight negative skew (Skewness = -0.3577) and Kurtosis is 3.0719. The Jarque Bera statistic is 25.85 which is greater than the 5% critical value 5.99. So, we reject the null hypothesis that the log-log regression errors are normal.

h.

```
Linear-Log Model: FOOD = α₁ + α₂*ln(INCOME) + e
> print(summary_linear_log)

Call:
lm(formula = food ~ ln_income, data = cex5_log)

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
```




```
R² comparison:
> cat("Linear Model R²:", summary_linear$r.squared, "\n")
Linear Model R²: 0.0422812
> cat("Log-Log Model Generalized R²:", generalized_r2_log_log, "\n")
Log-Log Model Generalized R²: 0.03965161
> cat("Linear-Log Model R²:", summary_linear_log$r.squared, "\n")
Linear-Log Model R²: 0.03799984
```

The figure is similar to that of the linear model and not as well-defined as the plot for the log-log model. Based on this visual criterion, the log-log model appears to fit the data better.

The R^2 value for the linear-log model is 0.038, which is lower than that of the linear model and also lower than the generalized R^2 from the log-log model. Using this statistical criterion, the linear model seems to provide a better fit.

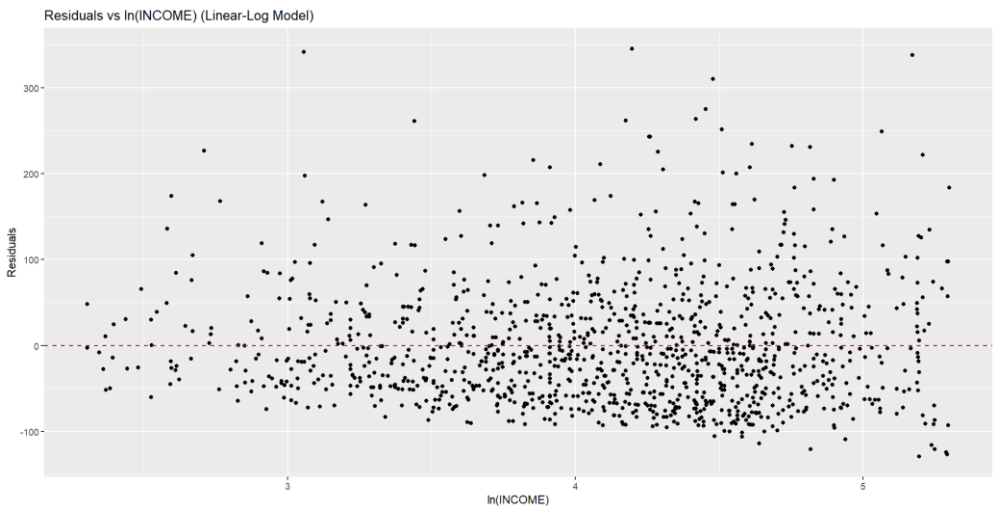
i.

```
> cat("\nElasticity estimates for Linear-Log Model:\n")

Elasticity estimates for Linear-Log Model:
> print(elasticity_results_linear_log)
  INCOME FOOD_Predicted 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
```

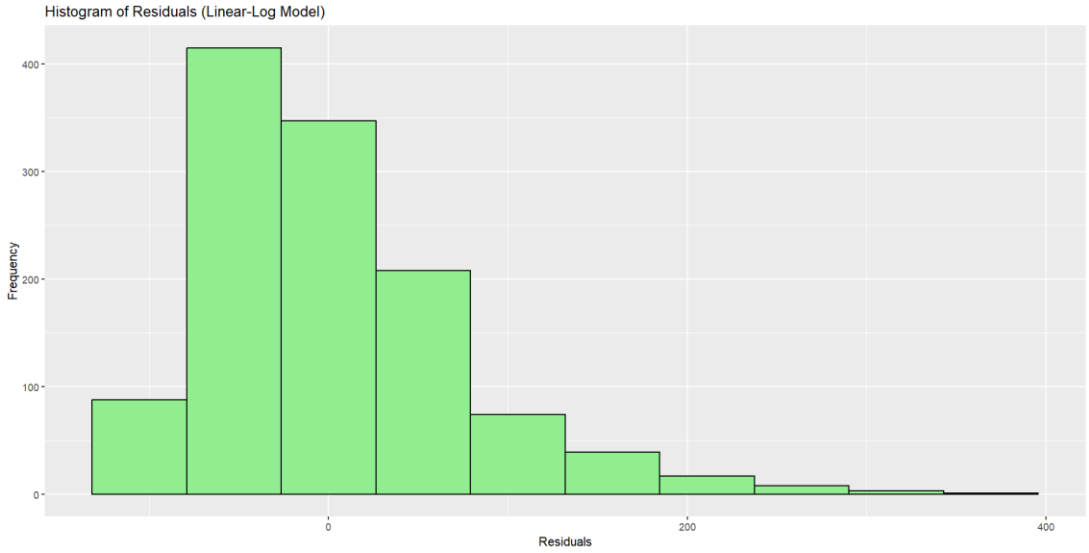
```
Comparing elasticities across all models:
> cat("Log-Log model elasticity (constant):", elasticity_log_log, "\n")
Log-Log model elasticity (constant): 0.1863054
> cat("Linear model elasticity at INCOME = 19:", elasticity_linear[1], "\n")
Linear model elasticity at INCOME = 19: 0.07145038
> cat("Linear-Log model elasticity at INCOME = 19:", elasticity_linear_log[1], "\n\n")
Linear-Log model elasticity at INCOME = 19: 0.2495828

> cat("Linear model elasticity at INCOME = 65:", elasticity_linear[2], "\n")
Linear model elasticity at INCOME = 65: 0.2083876
> cat("Linear-Log model elasticity at INCOME = 65:", elasticity_linear_log[2], "\n\n")
Linear-Log model elasticity at INCOME = 65: 0.1909624

> cat("Linear model elasticity at INCOME = 160:", elasticity_linear[3], "\n")
Linear model elasticity at INCOME = 160: 0.3931988
> cat("Linear-Log model elasticity at INCOME = 160:", elasticity_linear_log[3], "\n")
Linear-Log model elasticity at INCOME = 160: 0.1629349
```

The elasticity estimate from the **log-log model** is **statistically similar** to the elasticity from the linear and linear-log models at moderate income levels (e.g. INCOME = 65), and lies within the range of linear model elasticities. However, the **linear-log model** shows **dissimilar elasticities** at low and high income levels when compared to the linear model, suggesting that the choice of model significantly affects elasticity estimates across different income ranges.

j.



```
Jarque-Bera test for residuals (Linear-Log Model):
> print(jb_residuals_linear_log)

        Jarque Bera Test

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



How to interpret residual skewness visually:

- **Positive skew**: More residuals **below the zero line** (many small negatives, few large positives).
- **Negative skew**: More residuals **above the zero line** (many small positives, few large negatives).
- **Symmetry**: Residuals are roughly balanced above and below zero.

The residual scatter shows positive skewness at each income level and overall. The Jarque-Bera statistic is 628.07 which is far greater than the 5.99 critical value. We reject the normality of the model errors. The data scatter suggests a slight "spray" pattern.

k.

The linear model is counter-intuitive with increasing income elasticity. The linear-log model certainly satisfies economic reasoning, but the residual pattern is not an ideal random scatter. The log-log model implies that the income elasticity is constant for all income levels, which is not impossible to imagine, and the residual scatter is the most random, and the residuals are the least non-normal, based on skewness and kurtosis. On these grounds the **log-log model** seems like a good choice.