**4.4**

**a. Sketch the fitted values from Model 1 and Model 2**

**b. Explain why the four artists with no experience are not used in the estimation of Model 2.**


When EXPER=0, ln(0) is infinite small (undefined) cannot be calculated.

**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.**

$$\frac{d(\widehat{RATING})}{d(EXPER)} = 0.99$$

The marginal effect for the artist with 10 years experience or 20 years experience is the same as 0.99.

**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.**

$$\frac{d(\widehat{RATING})}{d(EXPER)} = 15.312 \times \frac{1}{EXPER}$$

The marginal effect for the artist with 10 years experience is 15.312/10=1.531 . The marginal effect for the artist with 20 years experience is 15.312/20=0.766 .

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

Model 2 fits the data better since the $\mathcal{R}^2$ of model 2 is closer to 1

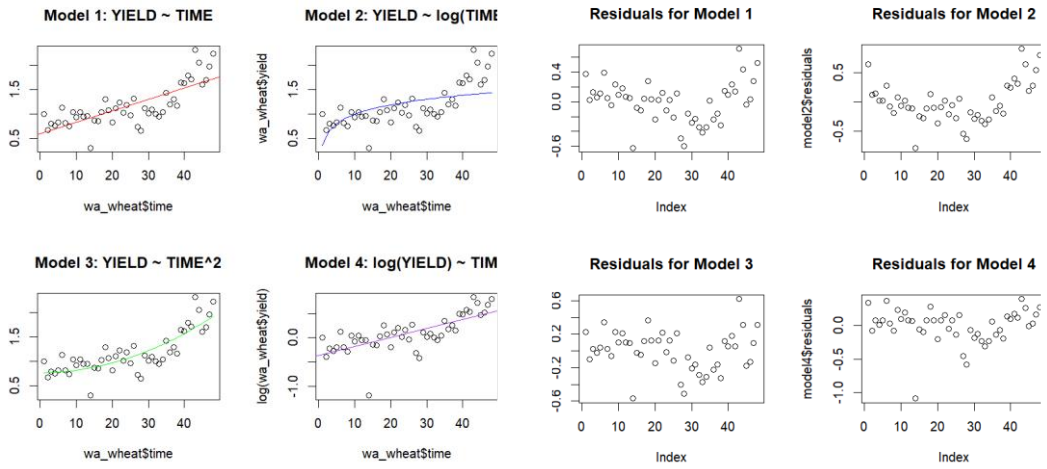
**f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning ? Explain.**

Model 2 is more reasonable based on economic reasoning since the margin effect from experience is usually decreasing in reality.

4.28.

**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 (iv) values for , which equation do you think is preferable? Explain**

(i) fitted equations (ii)residuals




| (iii) error normality tests | p-value | | (iv) $\mathcal{R}^2$ |
|---|---|---|---|
| $YIELD_t = \beta_0 + \beta_1 TIME + e_t$ | 0.6792 | Obey normal distribution | 0.5778 |
| $YIELD_t = \alpha_0 + \alpha_1 \ln(TIME) + e_t$ | 0.1856 | Obey normal distribution | 0.3386 |
| $YIELD_t = \gamma_0 + \gamma_1 TIME^2 + e_t$ | 0.8266 | Obey normal distribution | 0.6890 |
| $\ln(YIELD_t) = \phi_0 + \phi_1 TIME + e_t$ | $7.205 \times 10^{-5}$ | since p-value<0.05, model 4 disobey normal distribution. | 0.5074 |

In my opinion,I consider model 3 is preferable since $\mathcal{R}^2$ its is closer to 1.

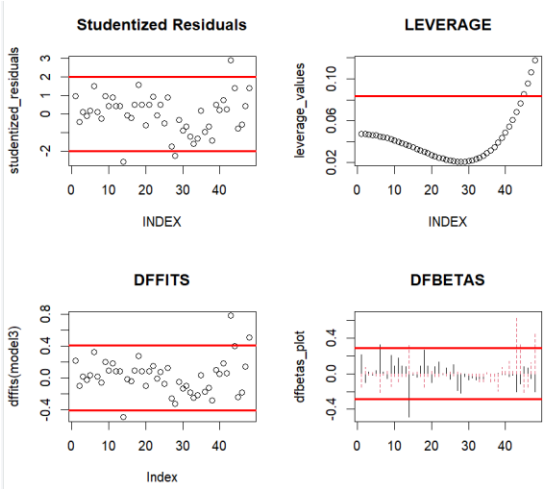
**b. Interpret the coefficient of the time-related variable in your chosen specification.**

```
> #4.28.b
> summary(model3)

Call:
lm(formula = yield ~ time2, data = wa_wheat)

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 ***
time2       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
```

Intercept=0.7737 When TIME=0(1950 year),the predicted yield is 0.7737. The coefficient of $\mathrm{TIME}^2$ is 0.0004986. The effect of the time variable on yield exhibits an accelerating growth trend.

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

unusual observations 14 28 43 45 46 47 48




**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?**

confidence interval = [1.4126 , 2.4324]

YIELD = 2.2318∈[1.4126 , 2.4324]

Yes, the interval contains the true value.

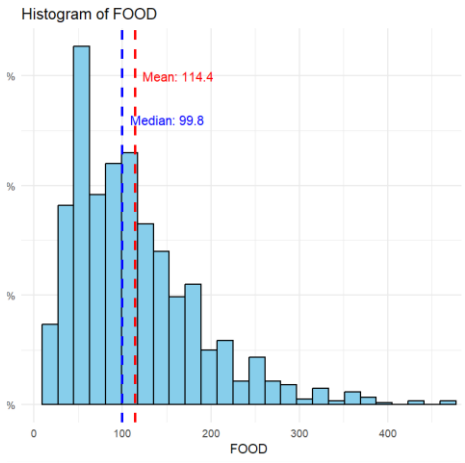
4.29

**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.**

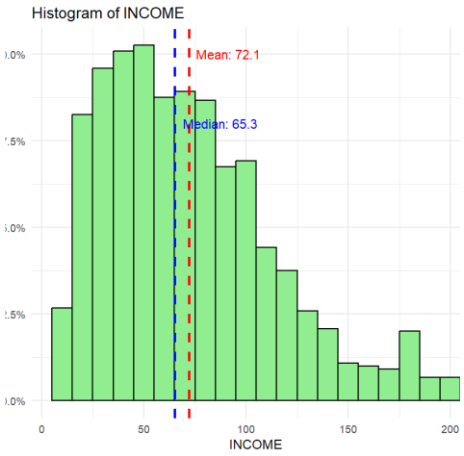
summary statistics for the variables FOOD and INCOME

| | min | max | mean | Q1 | median | Q3 | s.d |
|---|---|---|---|---|---|---|---|
| food | 9.63 | 476.67 | 114.44 | 57.78 | 99.80 | 145 | 72.6575 |
| INCOME | 10 | 200 | 72.14 | 40 | 65.29 | 96.79 | 41.6523 |

**FOOD**




**INCOME**




Both two histograms are not symmetrical and bell-shaped curves. Since their sample mean are larger than the medians there are right-skewed. Both Jarque-Bera statistic for FOOD and INCOME are larger than the critical value for a test at the 5% level(5.99),so we reject the null hypothesis of normality for each variable.

```
$jb_food

        Jarque Bera Test

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

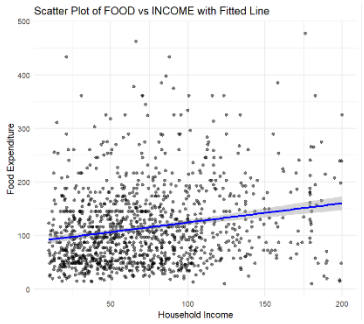
Jarque-Bera statistic is 648.65

```
$jb_income

        Jarque Bera Test

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

Jarque-Bera statistic is 148.21

**b. Estimate the linear relationship FOOD = β1 +β2INCOME +e. Create a scatter plot FOOD versus INCOME and include the fitted least squares line. Construct a 95% interval estimate for β2. Have we estimated the effect of changing income on average FOOD relatively precisely, or not?**

```
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
```

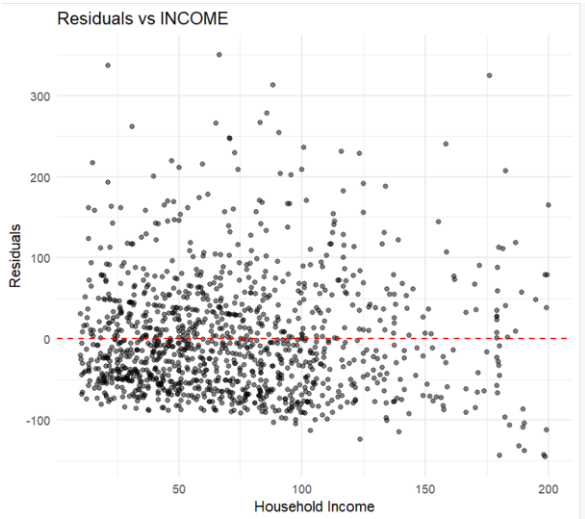


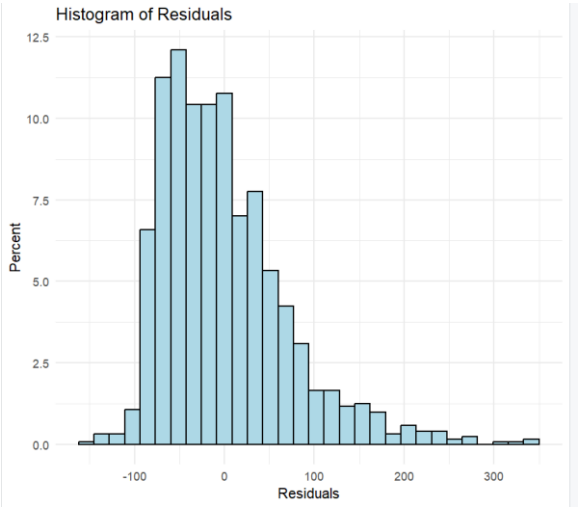

```
> confint(model, level = 0.95)
                  2.5 %    97.5 %
(Intercept) 80.5064570 96.626543
income       0.2619215  0.455452
```

$\widehat{\beta_1} = 88.5665, \widehat{\beta_2} = 0.35869$ Confidence Interval for $\widehat{\beta_2}$= [0.2613,0.4561]

The confidence interval does not include 0, indicating that income has a statistically significant effect on food expenditure.However, the interval is relatively wide, meaning there is some uncertainty about the precise impact.

**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.**



```
        Jarque Bera Test

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

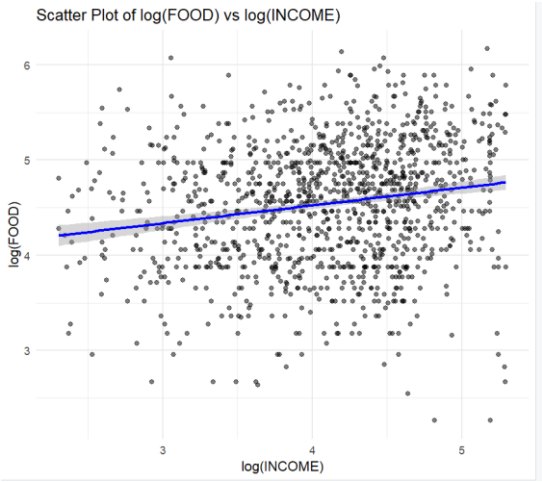
Residuals are NOT normally distributed since the residuals histogram is right-skewed. The Residuals against INCOME plot shows heteroskedasticity (variance increases as INCOME increases). It is more important for the random error term (e) to be normally distributed rather than the variables FOOD and INCOME themselves because the key assumption in OLS is:$e_t \sim N(0, \sigma^2)$

**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?**

| INCOME | point estimate | elasticity | lower bound | upper bound |
|---|---|---|---|---|
| 19 | 95.38155 | 0.07145 | 0.05217 | 0.09073 |
| **65** | 111.88114 | 0.20839 | 0.15217 | 0.26461 |
| 160 | 145.95638 | 0.39320 | 0.28712 | 0.49927 |

The estimated elasticities dissimilar. The elasticities increase as income increases, meaning higher-income households are more responsive to income changes in terms of food expenditure. The confidence intervals do not overlap, indicating statistically significant differences between the elasticities. Food is a necessity, so its income elasticity is usually low

**e. For expenditures on food, estimate the log-log relationship ln(FOOD) = γ1 +γ2ln(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 for the loglog model and compare it to the from the linear model. Which of the models seems to fit the data better?**




```
> summary(log_model)$r.squared
[1] 0.03322915
> summary(model)$r.squared
[1] 0.0422812
```

Since the linear model has a slightly higher than the log-log model, the linear model is a better fit.

**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**

A point estimate of elasticity is 0.18631.

The confidence interval is [0.1293, 0.2433].

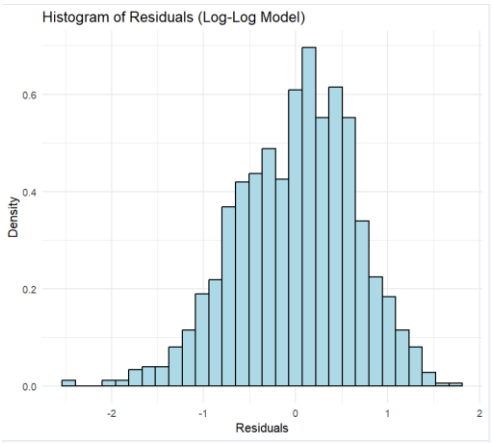
At Income = 19: [0.0522, 0.0907] (No Overlap)

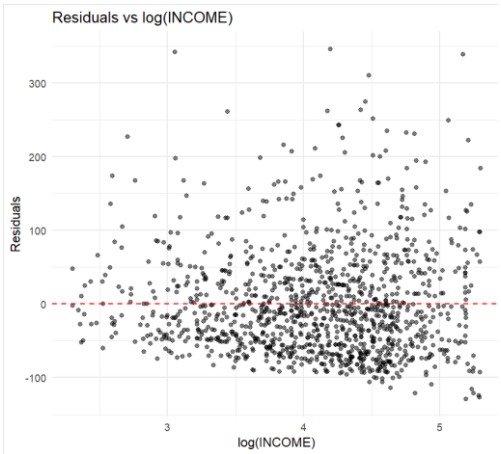
At Income = 65: [0.1522, 0.2646] (Overlap)→fail to reject H0.

At Income = 160: [0.2871, 0.4993] (No Overlap)

The confidence intervals overlap at INCOME = 65, suggesting the models provide similar estimates around the middle-income range. However, the models are statistically different at low (INCOME = 19) and high (INCOME = 160) levels.

**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?**



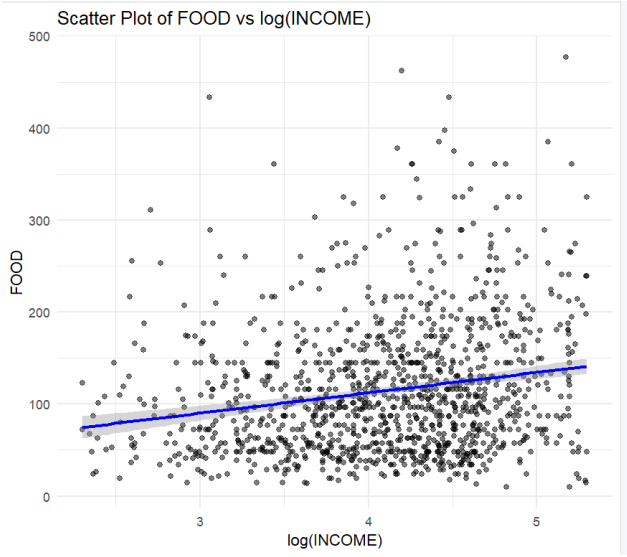

```
        Jarque Bera Test

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

The points appear randomly scattered around zero.

The residuals appear roughly normal, but the left tail is slightly extended, suggesting some left skew.

p-value < 0.05 → Reject normality → Residuals are not normal

**h. For expenditures on food, estimate the linear-log relationship FOOD = α1 +α2ln(INCOME) +e.Create a scatter plot for FOOD versus ln(INCOME) and include the fitted least squares line. Com-pare this to the plots in (b) and (e). Is this relationship more well-defined compared to the others?Compare the R2 values. Which of the models seems to fit the data better?**




For linear-log model, $\mathcal{R}^2$=0.038

By $\mathcal{R}^2$ ,linear-linear model 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.**

Log-log Model vs Linear-log Model: For all income levels, the elasticity estimates from the loglog and linear-log models are statistically similar.

Linear-log Model vs Linear-linear Model: At income = 19 and 160, the intervals do not overlap, indicating that the elasticity estimates from the linear-log and linear-linear models are statistically different.

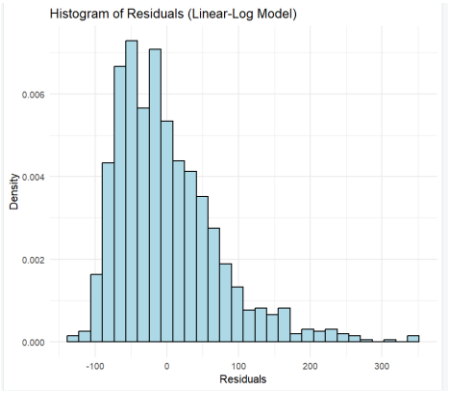
| INCOME | point estimate | elasticity | lower bound | upper bound | |
|---|---|---|---|---|---|
| **19** | 88.89788 | 0.24958 | 0.17840 | 0.32076 | No Overlap |
| **65** | 116.18722 | 0.19096 | 0.13650 | 0.24543 | Overlap |
| **160** | 136.17332 | 0.16293 | 0.11647 | 0.20940 | No Overlap |

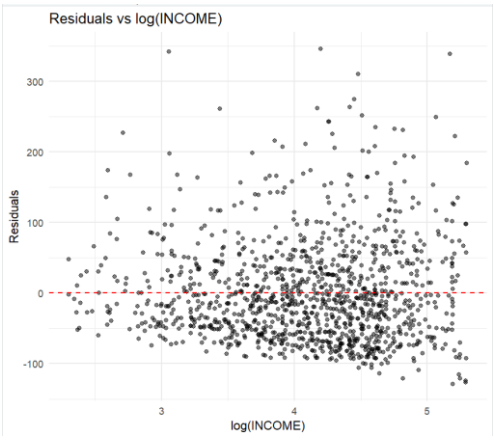
**j.Obtain the least squares residuals from the linear-log model and plot them against ln(INCOME). Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?**

Residuals are centered around zero, but the spread increases as log(INCOME) increases.

The peak is slightly shifted left, showing an excess of negative residuals.

p-value < 0.05 → Reject normality → Residuals are not normal



```
        Jarque Bera Test

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

**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.**

I prefer to Log-Log Model because it has better residual normality than other models.