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:

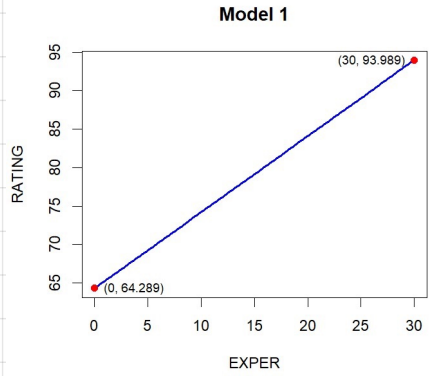
Model 1:

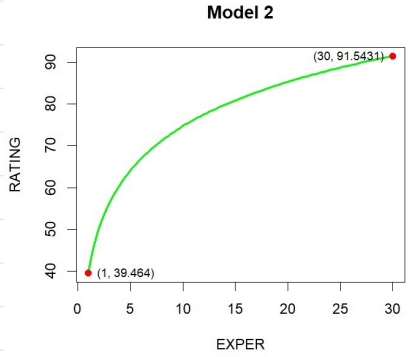
$$\widehat{RATING} = 64.289 + 0.990EXPER \quad N = 50 \quad R^2 = 0.3793$$
$$(\text{se}) \qquad (2.422) \quad (0.183)$$

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.

because $\ln(0)$ is undefined (∞).

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.

i) $\frac{\partial Rating}{\partial Exper} = 0.99 \Big|_{Exper=10}$ ii) $\frac{\partial Rating}{\partial Exper} = 0.99 \Big|_{Exper=20}$

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.

i) $\frac{\partial Rating}{\partial Exper} = 15.312 \times \frac{1}{Exper} \Big|_{Exper=10} = 1.5312$

ii) $\frac{\partial Rating}{\partial Exper} = 15.312 \times \frac{1}{Exper} \Big|_{Exper=20} = 0.7656$

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

$0.3793 < 0.4858 < 0.6414$ ∴ Model 2 fits the data better.

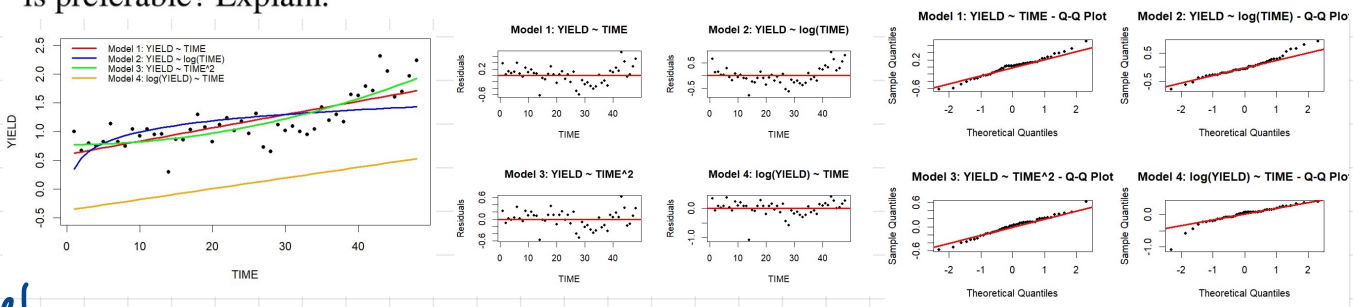
f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning? Explain.

Model 2 is more reasonable. For work experience, there is usually a phenomenon of diminishing marginal utilities.

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.



model

1. `Multiple R-squared: 0.5778,`
2. `Multiple R-squared: 0.3386,`
3. `Multiple R-squared: 0.689,`
4. `Multiple R-squared: 0.5074,`

Model 3: $YIELD = r_0 + r_1 TIME^2 + e$ is preferable ∵ R^2 is highest. Residuals are very close to a normal distribution.

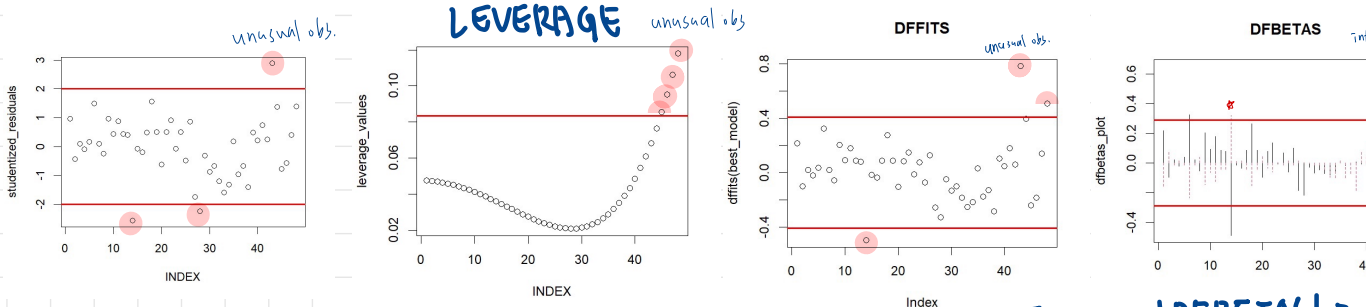
b. Interpret the coefficient of the time-related variable in your chosen specification.

```
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 ***
```

r_0: r_1:

$r_1 > 0$ suggests that yield increases at an accelerating rate over time. r_1 is highly statistically significant.

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



threshold: $|\hat{e}_i^{stu}| > 2$

$h_i > 2 \times \frac{2}{48} = 0.0833$

$|DFFITS| > 2 \times \sqrt{\frac{2}{48}} \therefore 0.408$

$|DFBETAS| > \frac{2}{\sqrt{48}} \therefore 0.2887$

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?

```
> (pred <- predict(model_train, newdata, interval = "prediction",
+                  level = 0.95))
       fit      lwr      upr
1 1.922482 1.412563 2.432401
> (origin <- wa_wheat[48, 1])
[1] 2.2318
```

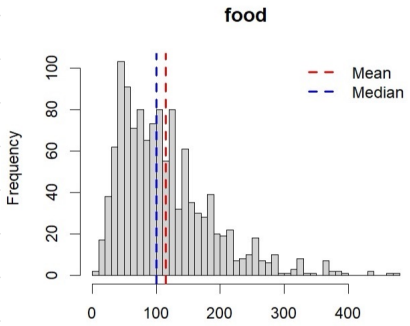
95% P.I = [1.4126, 2.4324]

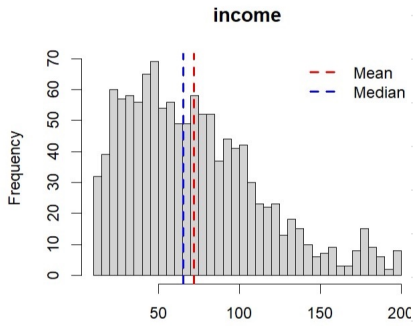
true value = 2.2318 ∈ 95% P.I.

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.

```
      food             income
 Min.   :  9.63   Min.   : 10.00
 1st Qu.: 57.78   1st Qu.: 40.00
 Median : 99.80   Median : 65.29
 Mean   :114.44   Mean   : 72.14
 3rd Qu.:145.00   3rd Qu.: 96.79
 Max.   :476.67   Max.   :200.00
```





standard deviation

```
    food   income
72.65750 41.65228
```

```
Jarque Bera Test

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

food

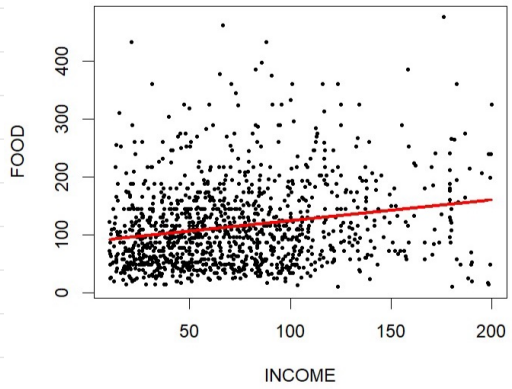
```
Jarque Bera Test

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

income → reject

These two variables are not symmetrical and bell-shape curves. Ho : X ~ Normal
Both sample means are greater than median (right skewed dist)

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?

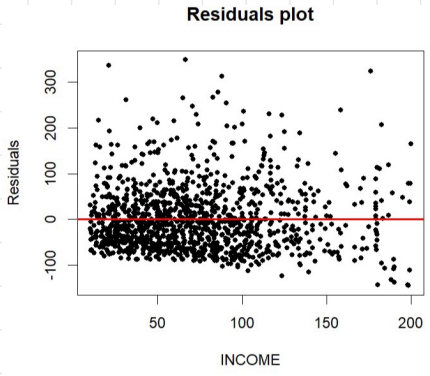


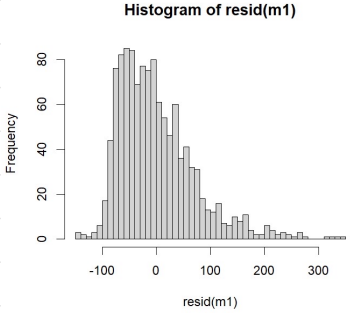
95% C.I for β_2 = [0.2619, 0.4555]

```
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 ***
```

The coefficient of slope (INCOME) is highly statistically significant.

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:  resid(m1)
X-squared = 624.19, df = 2, p-value < 2.2e-16
```

↳ reject residuals are normality

Residuals are not evenly and symmetrically distributed around 0.

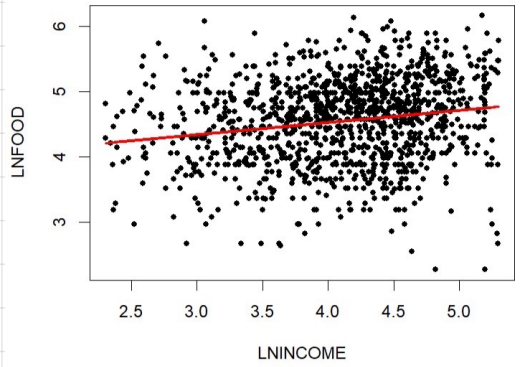
d. Calculate both a point estimate and a 95% interval estimate of the elasticity of food expenditure with respect to income at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Are the estimated elasticities similar or dissimilar? Do the interval estimates overlap or not? As *INCOME* increases should the income elasticity for food increase or decrease, based on Economics principles?

(FOOD)

INCOME	Fitted_FOOD	Elasticity	Elasticity_Lower	Elasticity_Upper
19	95.38155	0.07145038	0.05217475	0.09072601
65	111.88114	0.20838756	0.15216951	0.26460562
160	145.95638	0.39319883	0.28712305	0.49927462

① The elasticities increase as income increases.
② The confidence intervals do not overlap
③ Engel's law suggest elasticity should typically decrease, but in this case, it increases.

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?



linear - linear

```
Multiple R-squared:  0.04228
```

log - log

```
Multiple R-squared:  0.03323
```

linear - linear model is better than log-log model.

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.

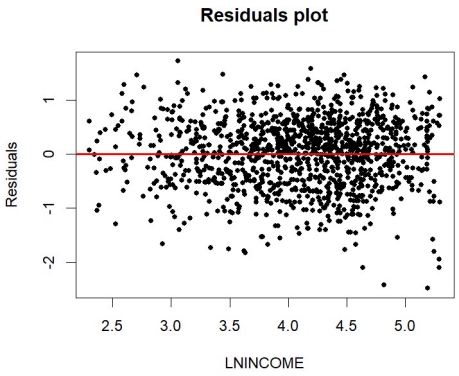
```
> cat("Point Estimate of Elasticity:", beta1_hat2, "\n")
Point Estimate of Elasticity: 0.1863054
> cat("95% Confidence Interval for Elasticity: (", beta1_CI_loglog
[1], ",", beta1_CI_loglog[2], ")\n")
95% Confidence Interval for Elasticity: ( 0.1293432 , 0.2432675 )
```

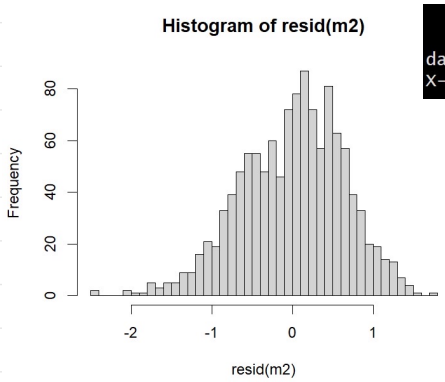
```
Income: 19  v.s. log - log.
Z-value: 3.743498
P-value: 0.0001814758
The two elasticities are significantly different (reject H0).

Income: 65  v.s. log - log
Z-value: -0.5407951
P-value: 0.5886488
No significant difference between the two elasticities (fail to reject H0).

Income: 160  v.s. log - log.
Z-value: -3.367963
P-value: 0.0007572575
The two elasticities are significantly different (reject H0).
```

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?





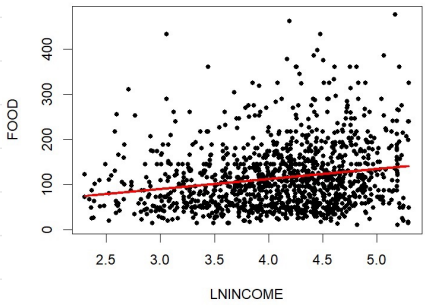

```
        Jarque Bera Test

data:  resid(m2)
X-squared = 25.85, df = 2, p-value = 2.436e-06
```

reject the regression errors are normal.

residuals slight skewed left

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?




```
Multiple R-squared:  0.038,
```

R^2 is greater than log-log model, is less than linear-linear model

By R^2, linear-linear 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.

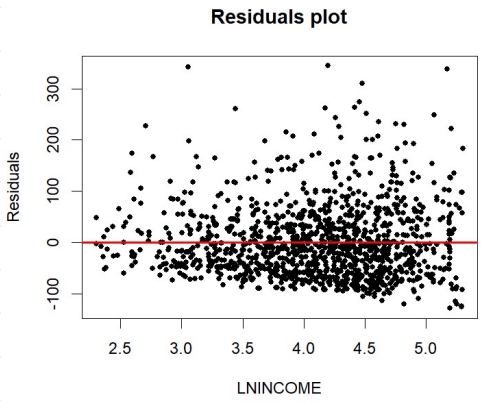
INCOME	Fitted_FOOD	Elasticity	Elasticity_Lower	Elasticity_Upper
19	88.89788	0.2495828	0.1784009	0.3207648
65	116.18722	0.1909624	0.1364992	0.2454256
160	136.17332	0.1629349	0.1164652	0.2094046

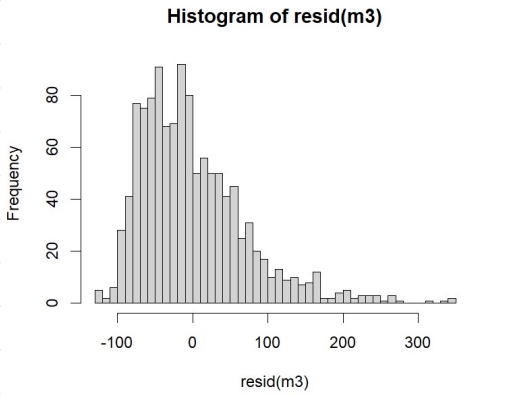
```
Comparing Elasticities: INCOME = 19 vs 65
P-value: 0.1998681
No significant difference between the two elasticities (fail to
reject H0).

Comparing Elasticities: INCOME = 19 vs 160
P-value: 0.04573626
The two elasticities are significantly different (reject H0).

Comparing Elasticities: INCOME = 65 vs 160
P-value: 0.4429038
No significant difference between the two elasticities (fail to
reject H0).
```

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?






```
        Jarque Bera Test

data:  resid(m3)
X-squared = 628.07, df = 2, p-value < 2.2e-16
```

reject the regression errors are normal.

Residuals are not evenly and symmetrically distributed around 0.
Residuals skewed right

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 log-log model. Because the residual plot is the most random and histogram of residuals is slight skewed left.