HW4

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:

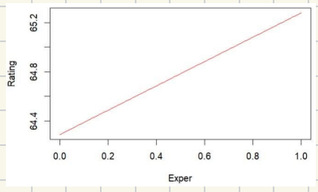
$$\widehat{RATING} = 64.289 + 0.990 EXPER \quad N = 50 \quad R^2 = 0.3793$$
$$(se) \qquad (2.422) \quad (0.183)$$

Model 2:

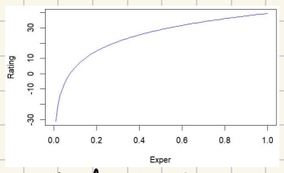
$$\widehat{RATING} = 39.464 + 15.312 \ln(EXPER) \quad N = 46 \quad R^2 = 0.6414$$
$$(se) \qquad (4.198) \quad (1.727)$$

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.

a.



b.



Since the domain of nature logarithm is $(0,\infty)$, the four artists with 0 experience can not be used in the regression model.

c. the margin effect on Rating is

$$\frac{\partial \text{ Rating}}{\partial \text{ Exper}} = 0.99$$

(i) the artists with 10 years Exper: 0.99

(ii) the artists with 20 years Exper: 0.99

d. the margin effect on Rating is

$$\frac{\partial \text{ Rating}}{\partial \text{ Exper}} = \frac{15.312}{\text{Exper}}$$

(i) the artists with 10 years Exper: $\frac{15.312}{10} = 1.5312$

(ii) the artists with 20 years Exper: $\frac{15.312}{20} = 0.7656$

e. since R^2 of model 1 $= 0.3793 < 0.6414$ (the R^2 of model 2)

$\Rightarrow$ Model 2 fits the data better

f. Model 2 is reasonable since it has larger R^2, and the rating will rapidly increase with less experience.

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:

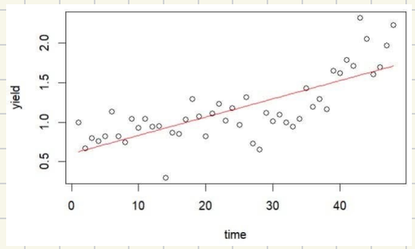
$$YIELD_t = \beta_0 + \beta_1 TIME + e_t$$
$$YIELD_t = \alpha_0 + \alpha_1 \ln(TIME) + e_t$$
$$YIELD_t = \gamma_0 + \gamma_1 TIME^2 + e_t$$
$$\ln(YIELD_t) = \phi_0 + \phi_1 TIME + e_t$$

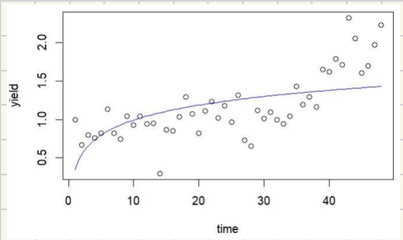
a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.
b. Interpret the coefficient of the time-related variable in your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

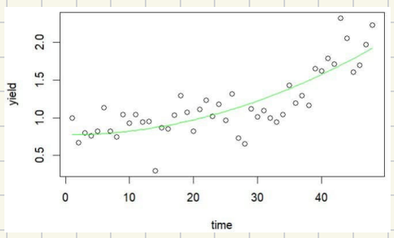
a. the estimation of four equations:

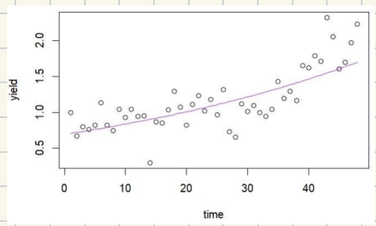
```
> theCoefficientB1
[1]  0.6032451  0.3509620  0.7736655 -0.3639376
> theCoefficientB2
[1] 0.0230779201 0.2790085294 0.0004986181 0.0186323476
```

Fitted equations:

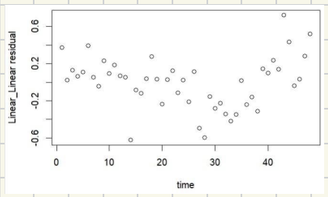


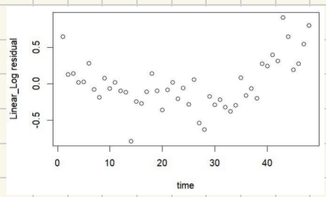


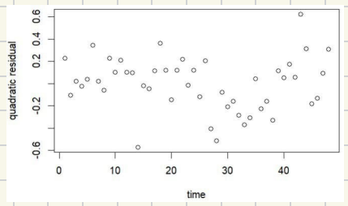


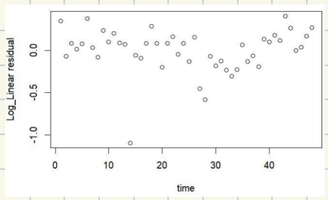


Residuals:









error normality tests:

```
> jbTest1

        Jarque Bera Test

data:  Bresidual
X-squared = 0.13257, df = 2, p-value = 0.9359

> jbTest2

        Jarque Bera Test

data:  Aresidual
X-squared = 2.7629, df = 2, p-value = 0.2512

> jbTest3

        Jarque Bera Test

data:  Cresidual
X-squared = 0.32406, df = 2, p-value = 0.8504

> jbTest4

        Jarque Bera Test

data:  Dresidual
X-squared = 83.874, df = 2, p-value < 2.2e-16
```

R^2:

```
> rsq1
[1] 0.5778369
> rsq2
[1] 0.3385733
> rsq3
[1] 0.6890101
> rsq4
[1] 0.5073566
```

According to the above data, I prefer to use model 3. Since it fits the data better, its residual does not have unusual pattern and its R^2 is the nearest to 1 with error normality tests, since the p-value $= 0.8504 > 0.05$, we will not reject that its residual has normal distribution.

b. the estimation:

```
> theCoefficientB1
[1]  0.6032451  0.3509620  0.7736655 -0.3639376
> theCoefficientB2
[1] 0.0230779201 0.2790085294 0.0004986181 0.0186323476
```

In model 3: when time $= 0$, the yield is approximately 0.7736. since $c_2 = 0.00049$, when time increase 1 unit, the yield will increase $2 \times c_2 = 0.00098$.

d. the true value: 2.2318

the Interval:

```
> lower
[1] 1.382799
> upper
[1] 2.379424
```

Yes, the true value is contained in the 95% prediction interval.

```r
yield<-c(wa_wheat$northampton)
time<-c(wa_wheat$time)
data<-data.frame(time, yield)

#a
modlinear_linear<-lm(yield~time, data=data)
modlinear_log<-lm(yield~log(time), data=data)
modquaratic<-lm(yield~I(time^2), data=data)
modlog_linear<-lm(log(yield)~time, data=data)
b1<-coef(modlinear_linear)[[1]]
b2<-coef(modlinear_linear)[[2]]
Bresidual<-resid(modlinear_linear)
a1<-coef(modlinear_log)[[1]]
a2<-coef(modlinear_log)[[2]]
Aresidual<-resid(modlinear_log)
c1<-coef(modquaratic)[[1]]
c2<-coef(modquaratic)[[2]]
Cresidual<-resid(modquaratic)
d1<-coef(modlog_linear)[[1]]
d2<-coef(modlog_linear)[[2]]
Dresidual<-resid(modlog_linear)
theCoefficientB1<-c(b1, a1, c1, d1)
theCoefficientB2<-c(b2, a2, c2, d2)
theCoefficientB1
theCoefficientB2
#plot
plot(time, yield, xlab="time", ylab="yield", col="black")
curve(b1+b2*x, add=TRUE, col="red")
curve(a1+a2*log(x), add=TRUE, col="blue")
curve(c1+c2*x^2, add=TRUE, col="green")
curve(exp(d1+d2*x), add=TRUE, col="purple")

plot(time, Bresidual, xlab="time", ylab="Linear_Linear residual", col="black")
plot(time, Aresidual, xlab="time", ylab="Linear_Log residual", col="black")
plot(time, Cresidual, xlab="time", ylab="quadratic residual", col="black")
plot(time, Dresidual, xlab="time", ylab="Log_Linear residual", col="black")
```

```r
library(tseries)
jbTest1<-jarque.bera.test(Bresidual)
jbTest2<-jarque.bera.test(Aresidual)
jbTest3<-jarque.bera.test(Cresidual)
jbTest4<-jarque.bera.test(Dresidual)
jbTest1
jbTest2
jbTest3
jbTest4

rsq1<-summary(modlinear_linear)$r.squared
rsq2<-summary(modlinear_log)$r.squared
rsq3<-summary(modquaratic)$r.squared
rsq4<-summary(modlog_linear)$r.squared
rsq1
rsq2
rsq3
rsq4
```

```r
#d
newyield<-c(wa_wheat$northampton[1:47])
newtime<-c(wa_wheat$time[1:47])
newdata1<-data.frame(newtime, newyield)
newmod<-lm(newyield~I(newtime^2), data=newdata1)
newb1<-coef(newmod)[[1]]
newb2<-coef(newmod)[[2]]
newresidual<-resid(newmod)
yield_1997<-newb1+(newb2)*(48^2)
yield_1997
t<-qt(0.975, 45)
se<-summary(newmod)$sigma
#leverage
f<-1+(1/47)+(((48-mean(newtime))^2)/sum((newtime-mean(newtime))^2))
lower<-yield_1997-t*se*(sqrt(f))
upper<-yield_1997+t*se*(sqrt(f))
lower
upper
```

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?

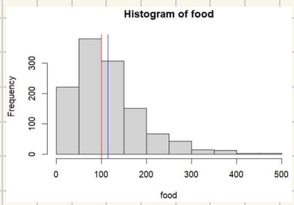
h. For expenditures on food, estimate the linear-log relationship $FOOD = \alpha_1 + \alpha_2 \ln(INCOME) + e$. Create a scatter plot for *FOOD* versus $\ln(INCOME)$ and include the fitted least squares line. Compare this to the plots in (b) and (e). Is this relationship more well-defined compared to 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.

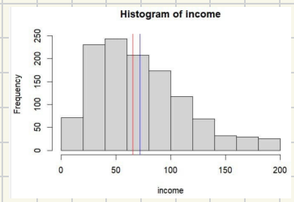
j. Obtain the least squares residuals from the linear-log model and plot them against $\ln(INCOME)$. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

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.

a.




```
> summary(food)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.63   57.78   99.80  114.44  145.00  476.67
> sd_food
[1] 72.6575
```




```
> summary(income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   40.00   65.29   72.14   96.79  200.00
> sd_income
[1] 41.65228
```

the blue line is Mean

the red line is Median

Since the two histograms are positive skewness, neither of them are symmetrical nor bell-shape curve.

The sample mean is larger than the median.

```
> jbTest_food

        Jarque Bera Test

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

> jbTest_income

        Jarque Bera Test

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

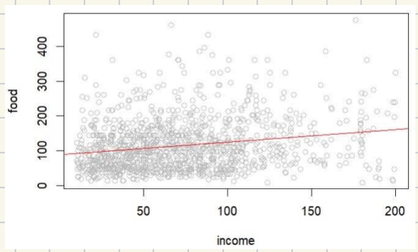
If p-value > 0.05, we will not reject H_0 = normal distribution. Neither of them are normal distribution.

b. Estimation:

```
> b1
[1] 88.5665
> b2
[1] 0.3586867
```

```
> se_b2
[1] 0.04932104
```

Plot:

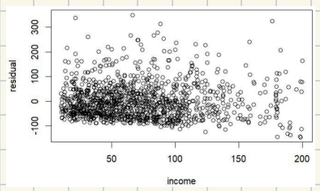


Interval:

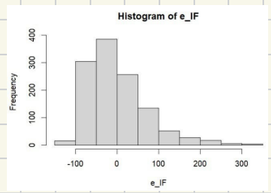
```
> lower
[1] 0.2619215
> upper
[1] 0.455452
```

Since the interval is narrow and standard error of β_2 is small ⇒ precise.

c.



It does not have any unusul pattern.




```
> e_jbTest

        Jarque Bera Test

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

Since its p-value < 0.05, it is not normal distribution.

It is more important that random error is normal distribution.

If e is normal distribution, the statistic inference (ex: hypothesis test) are valid.

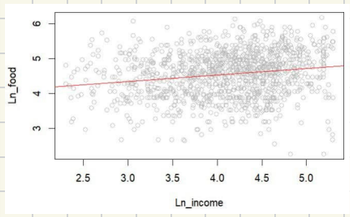
d.

```
> theIncome_p
[1]  19  65 160
> elasticity_p
[1] 0.07145038 0.20838756 0.39319883
> lower_elastic
[1] 0.05217475 0.15216951 0.28712305
> upper_elastic
[1] 0.09072601 0.26460562 0.49927462
```

The elasticities are dissimilar. the interval is not overlap.

As the income increase, the porpotion of income spend on food decrease → elasticity decrease (based on economic principle)

e.




```
> G_rsq_linear
[1] 0.04235027
> G_rsq_loglog
[1] 0.03328369
```

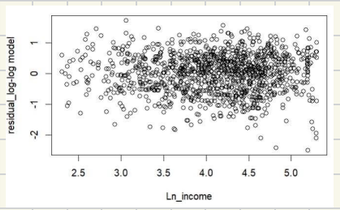
since R^2 of linear model is larger, linear model fits better.

f.

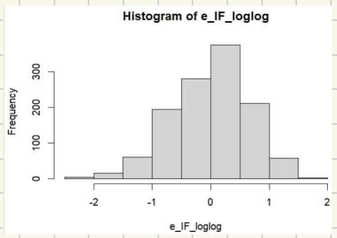
```
> b2_loglog
[1] 0.1863054
> lower_elastic_loglog
[1] 0.1293432
> upper_elastic_loglog
[1] 0.2432675
```

It is dissimilar to part(d). Since the elasticity of log-log model equal to β_2, which is a constant.

g.



It does not have unusual pattern.



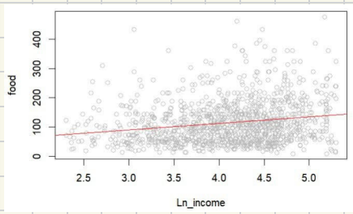

```
> e_jbTest_loglog

        Jarque Bera Test

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

Since its p-value < 0.05, it is not normal distribution.

h.




```
> rsq_linearlog
[1] 0.03799984
```

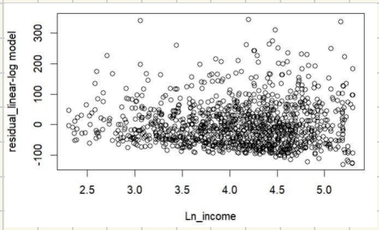
Compare with the R^2 in (e), the R^2 of linear model is greatest. Thus, linear model fits the data better.

i.

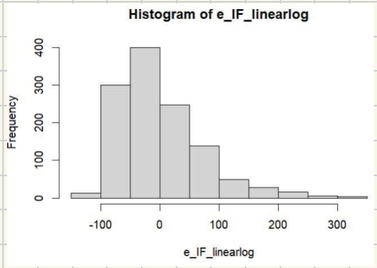
```
> elasticity_p_linearlog
[1] 0.2495828 0.1909624 0.1629349
> lower_elastic_linearlog
[1] 0.1784009 0.1364992 0.1164652
> upper_elastic_linearlog
[1] 0.3207648 0.2454256 0.2094046
```

As the income increase, the elasticity decrease.
Hence, it is dissimilar to others.

j.



It does not have unusual pattern.




```
> e_jbTest_linearlog

        Jarque Bera Test

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

Its p-value < 0.05, so it is not normal distribution.

h. I prefer log log model.
Based on economic principal, as the income increase, the elasticity decrease. However, in Linear model, its elasticity increase as the income increase. Thus, I will not choose linear model. About the other two model, log log model has higher R^2. Therefore, I prefer log log model.

```
food<-c(cex5_small$food)
income<-c(cex5_small$income)
sd_food<-sd(food)
sd_income<-sd(income)

#a
summary(food)
sd_food
summary(income)
sd_income
#histogram
hist(food)
abline(v=mean(food), col="blue")
abline(v=median(food), col="red")
hist(income)
abline(v=mean(income), col="blue")
abline(v=median(income), col="red")
#jbTest
jbTest_food<-jarque.bera.test(food)
jbTest_income<-jarque.bera.test(income)
jbTest_food
jbTest_income

#b
data=data.frame(income, food)
modIF<-lm(food~income, data=data)
b1<-coef(modIF)[[1]]
b2<-coef(modIF)[[2]]
b1
b2
#plot
plot(income, food, xlab="income", ylab="food", col="grey")
abline(b1, b2, col="red")
```

```
#95% interval
summaryModIF<-summary(modIF)
se_b2<-summaryModIF$coefficients[2, 2]
n<-length(food)
t<-qt(0.975, n-2)
lower<-b2-t*(se_b2)
upper<-b2+t*(se_b2)
lower
upper

#c
e_IF<-resid(modIF)
plot(income, e_IF, xlab="income", ylab="residual")
#histogram of residual
hist(e_IF)
#jbtest
e_jbTest<-jarque.bera.test(e_IF)
e_jbTest

#d
#point
theIncome_p<-c(19, 65, 160)
theFood_p<-b1+b2*theIncome_p
margin_effect<-b2
elasticity_p<-margin_effect*(theIncome_p/theFood_p)
elasticity<-margin_effect*(income/food)
n_elastic<-length(income)
t_elastic<-qt(0.975, n_elastic-2)
lower_elastic<-elasticity_p-t_elastic*(theIncome_p/theFood_p)*se_b2
upper_elastic<-elasticity_p+t_elastic*(theIncome_p/theFood_p)*se_b2
theIncome_p
elasticity_p
lower_elastic
upper_elastic
```

```r
#e
mod_loglog<-lm(log(food)~log(income), data=data)
plot(log(income), log(food), xlab="Ln_income", ylab="Ln_food", col="grey")
b1_loglog<-coef(mod_loglog)[[1]]
b2_loglog<-coef(mod_loglog)[[2]]
se_b2_loglog<-summary(mod_loglog)$coefficients[2, 2]
abline(b1_loglog, b2_loglog, col="red")
#r square
rsq_linear<-summary(modIF)$r.squared#linear
rsq_loglog<-summary(mod_loglog)$r.squared#log-log
N<-nrow(data)
p1<-length(coef(modIF))-1
p2<-length(coef(mod_loglog))-1
G_rsq_linear<-1-exp((log(1-rsq_linear)*N)/(N-p1-1))
G_rsq_loglog<-1-exp((log(1-rsq_loglog)*N)/(N-p2-1))
G_rsq_linear
G_rsq_loglog

#f
#95%interval
n_elastic_loglog<-length(elasticity_loglog)
t_elastic_loglog<-qt(0.975, n_elastic_loglog-2)
lower_elastic_loglog<-b2_loglog-t_elastic_loglog*se_b2_loglog
upper_elastic_loglog<-b2_loglog+t_elastic_loglog*se_b2_loglog
b2_loglog
lower_elastic_loglog
upper_elastic_loglog

#g
e_IF_loglog<-resid(mod_loglog)
plot(log(income), e_IF_loglog, xlab="Ln_income", ylab="residual_log-log model")
hist(e_IF_loglog)
e_jbTest_loglog<-jarque.bera.test(residuals(mod_loglog))
e_jbTest_loglog
```

```r
#h
mod_linearlog<-lm(food~log(income), data=data)
plot(log(income), food, xlab="Ln_income", ylab="food", col="grey")
b1_linearlog<-coef(mod_linearlog)[[1]]
b2_linearlog<-coef(mod_linearlog)[[2]]
se_b2_linearlog<-summary(mod_linearlog)$coefficient[2, 2]
abline(b1_linearlog, b2_linearlog, col="red")
#r square
rsq_linearlog<-summary(mod_linearlog)$r.squared
rsq_linearlog

#i
#95%interval
theIncome_p<-c(19, 65, 160)
theFood_p_linearlog<-b1_linearlog+b2_linearlog*log(theIncome_p)
margin_effect_linearlog<-b2_linearlog/theIncome_p
elasticity_p_linearlog<-margin_effect_linearlog*(theIncome_p/theFood_p_linearlog)
n_elastic_linearlog<-length(income)
t_elastic_linearlog<-qt(0.975, n_elastic_linearlog-2)
lower_elastic_linearlog<-elasticity_p_linearlog-t_elastic_linearlog*(se_b2_linearlog)/theFood_p_linearlog
upper_elastic_linearlog<-elasticity_p_linearlog+t_elastic_linearlog*(se_b2_linearlog)/theFood_p_linearlog
elasticity_p_linearlog
lower_elastic_linearlog
upper_elastic_linearlog

#j
e_IF_linearlog<-resid(mod_linearlog)
plot(log(income), e_IF_linearlog, xlab="Ln_income", ylab="residual_linear-log model")
hist(e_IF_linearlog)
e_jbTest_linearlog<-jarque.bera.test(residuals(mod_linearlog))
e_jbTest_linearlog
```