2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Plot house price against house size in a scatter diagram.

b. Estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

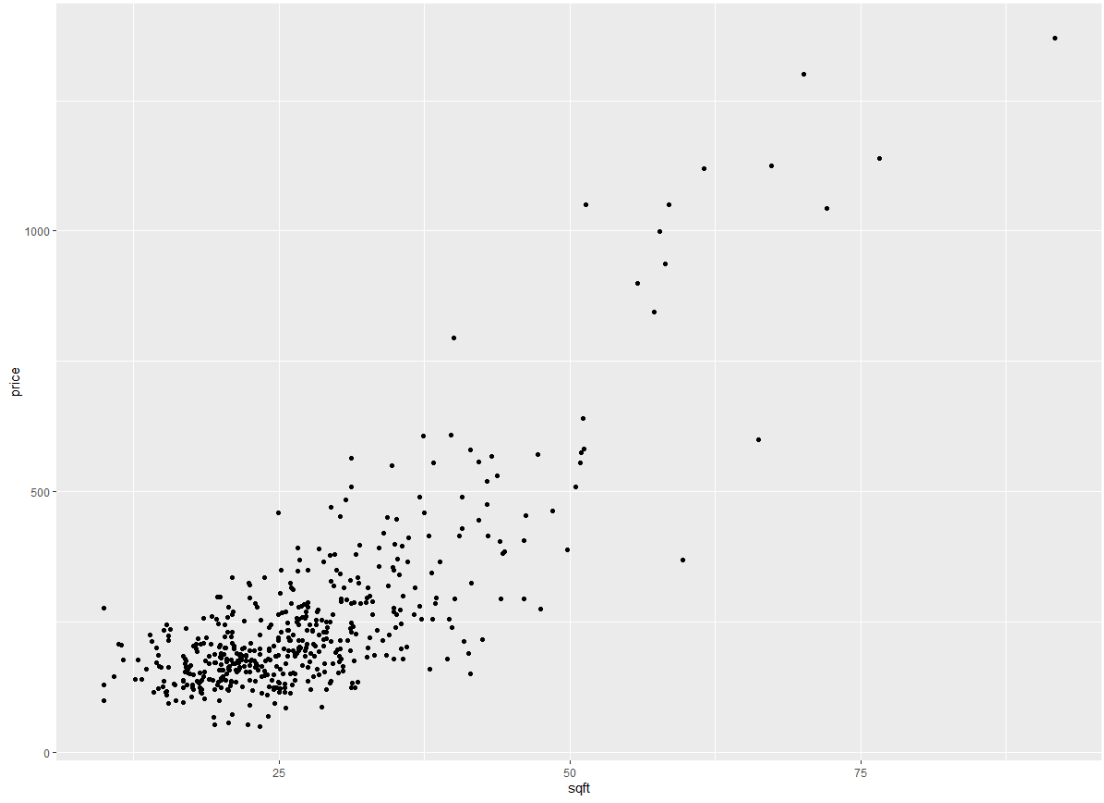
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

f. For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?

g. One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?

a.



b.

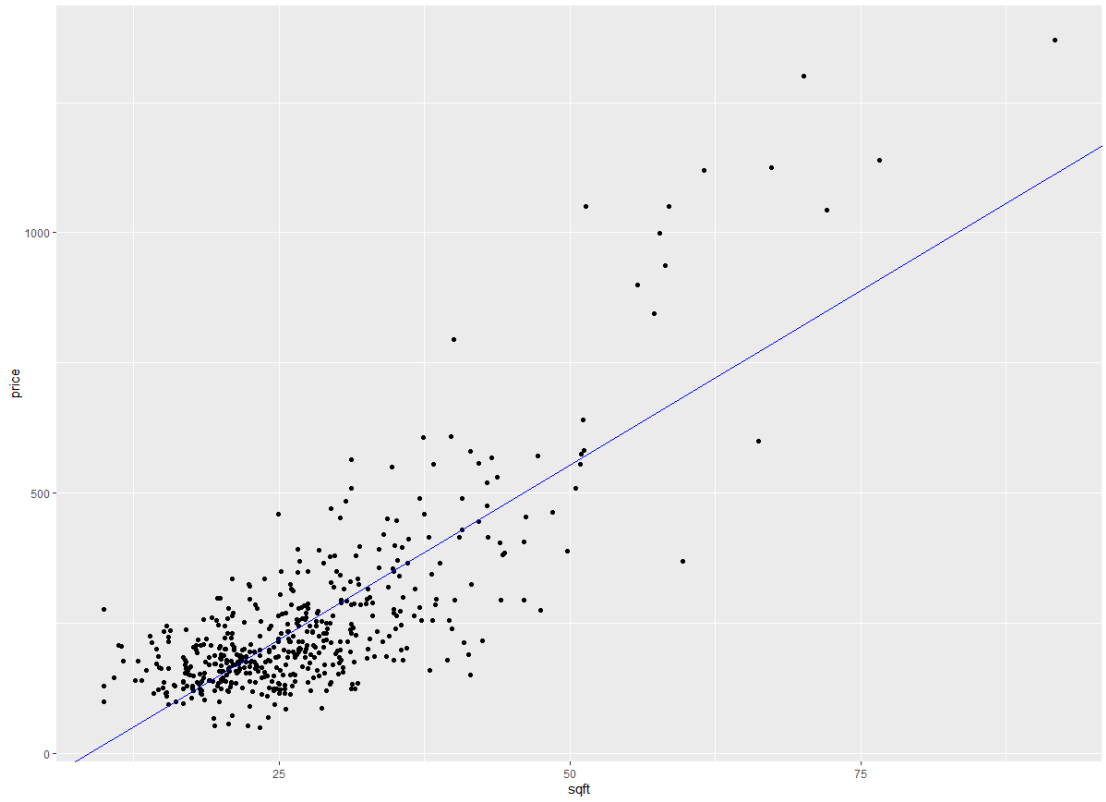
b1=-115.4236

b2=13.4029

$\widehat{PRICE} = -115.4236 + 13.4029 \times SQFT$

當 SQFT=0，PRICE 的估計值為-115.4236，若 SQFT 增加一單位(100 平方英尺)

則 PRICE 估計值增加 13.4029 單位($1000)



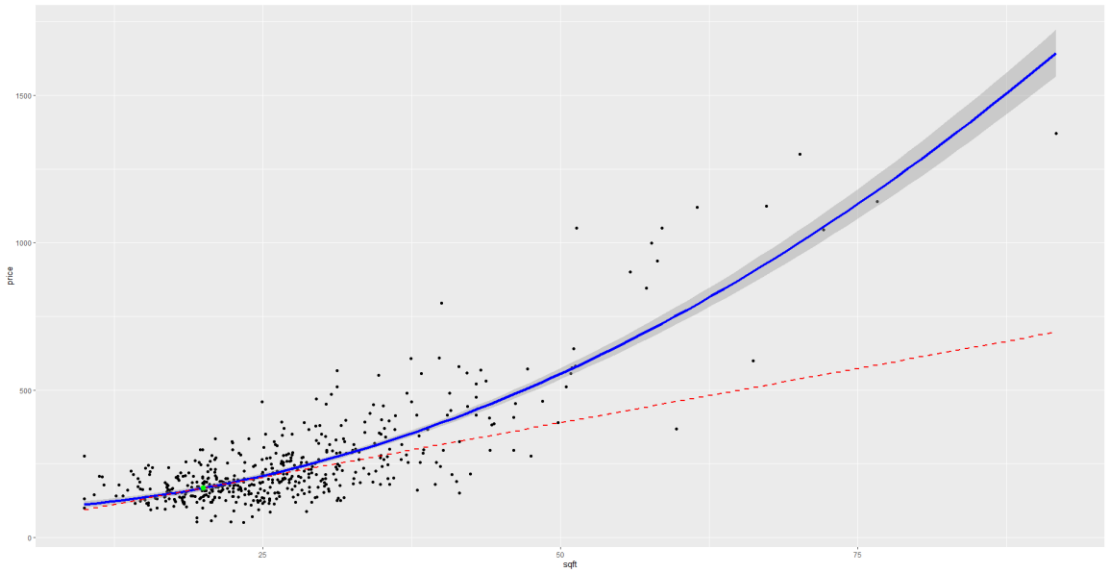
c.

在房屋面積為 2000 平方英尺(sqft=20)的情況下，增加 100 平方英尺的邊際效果為

$\widehat{PRICE} = 93.56585 + 0.184519 \times SOFT^2$

$\frac{dPRICE}{dSOFT}$ = 2*a2*SOFT = 2*0.184519*20 = 7.38076

d.

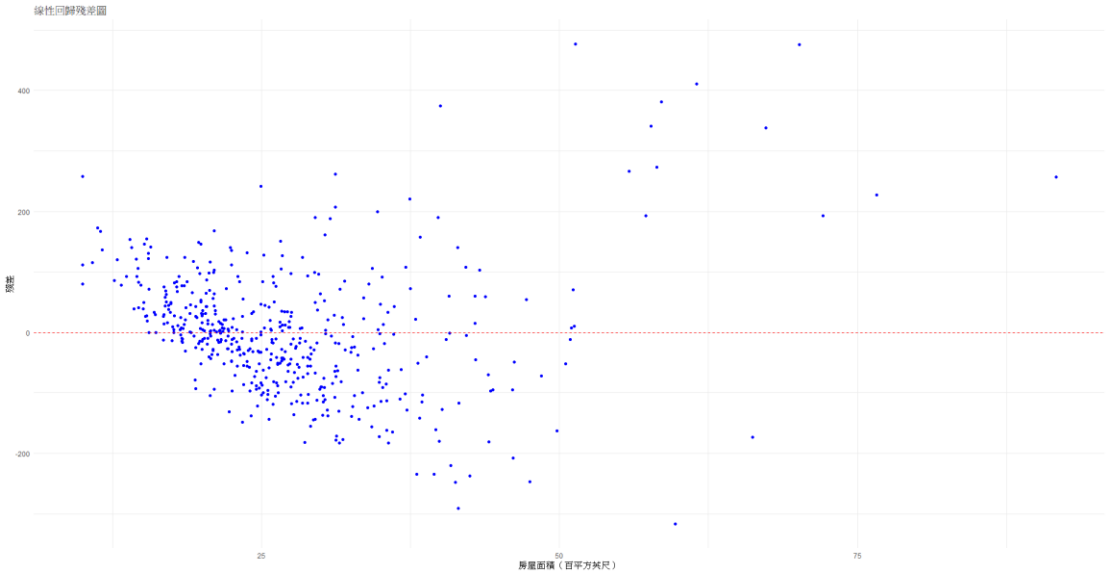


e.

elasticity = $\frac{dPRICE/PRICE}{dSOFT/SOFT}$ =(2*b2*20^2)/(b1+b2)=1.574556

f.

linear



```
        Shapiro-Wilk normality test

data:  collegetown$residuals_linear
W = 0.95602, p-value = 4.687e-11
```

常態檢定顯著

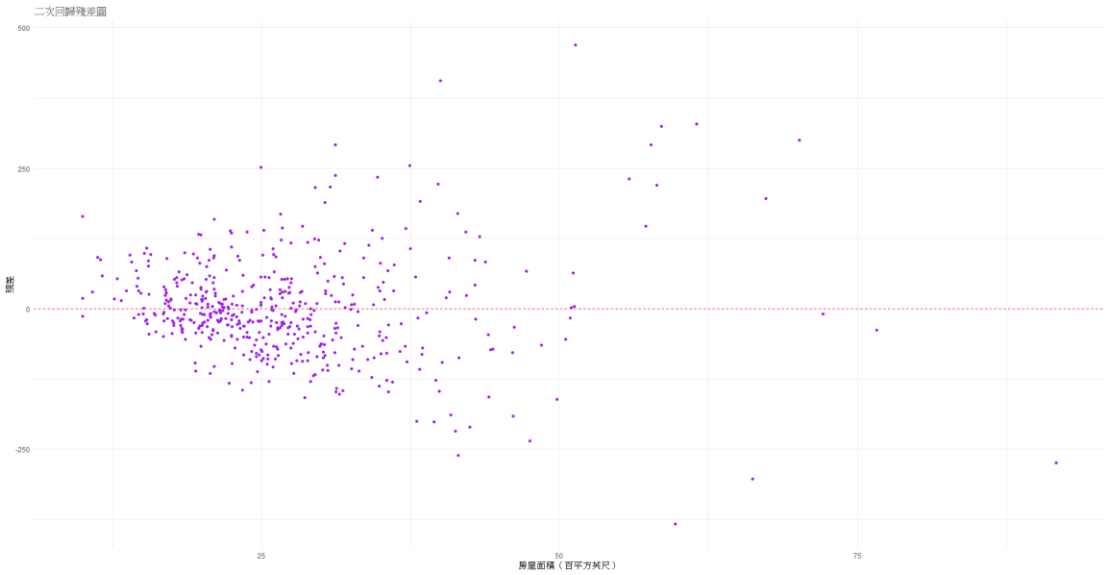
```
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 330.5479, Df = 1, p = < 2.22e-16
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 136.81, df = 1, p-value < 2.2e-16
```

P<0.05　SR3:Conditional Homoskedasticity 這個假設可能有問題

```
        Runs Test

data:  collegetown$residuals_linear
statistic = -7.6103, runs = 166, n1 = 250, n2 = 250, n = 500, p-value =
2.735e-14
alternative hypothesis: nonrandomness
```

p-value<0.05 殘差可能隨機 SR 4:Conditional Uncorrelated Error 這個假設有問題

quadratic




```
        Shapiro-Wilk normality test

data:  collegetown$residuals_quadratic
W = 0.94462, p-value = 1.027e-12
```

常態檢定顯著

```
> ncvTest(m2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 358.9102, Df = 1, p = < 2.22e-16
> bptest(m2)

        studentized Breusch-Pagan test

data:  m2
BP = 130.79, df = 1, p-value < 2.2e-16
```

P<0.05　SR3:Conditional Homoskedasticity 這個假設可能有問題

```
        Runs Test

data:  collegetown$residuals_quadratic
statistic = -11.012, runs = 128, n1 = 250, n2 = 250, n = 500, p-value < 2.2e-16
alternative hypothesis: nonrandomness
```

p-value<0.05 殘差可能隨機 SR 4:Conditional Uncorrelated Error 這個假設有問題

g.

SSE(b)= 5262847

SSE(c)= 4222356

(c)$\widehat{PRICE} = 93.56585 + 0.184519 \times SOFT^2$

(c)的 SSE 較小，所以(c)的模型較好，$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y})^2$，表示真實值與預測值的平方總和，SSE 越小代表誤差越小

2.25 Consumer expenditure data from 2013 are contained in the file *cex5_small*. [Note: *cex5* is a larger version with more observations and variables.] Data are on three-person households consisting of a husband and wife, plus one other member, with incomes between $1000 per month to $20,000 per month. *FOODAWAY* is past quarter's food away from home expenditure per month per person, in dollars, and *INCOME* is household monthly income during past year, in $100 units.

a. Construct a histogram of *FOODAWAY* and its summary statistics. What are the mean and median values? What are the 25th and 75th percentiles?

b. What are the mean and median values of *FOODAWAY* for households including a member with an advanced degree? With a college degree member? With no advanced or college degree member?

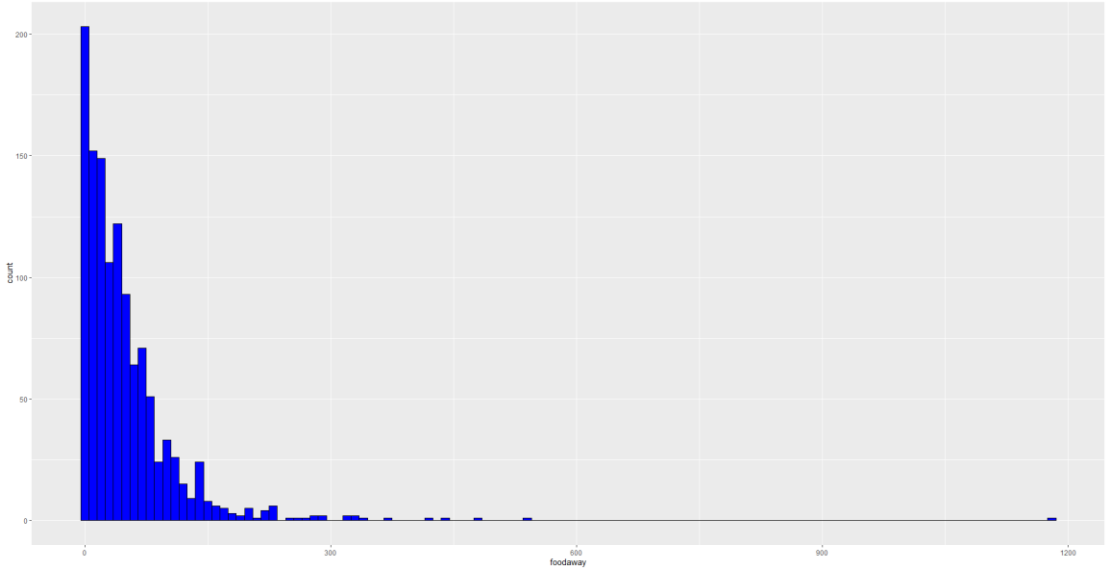
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

d. Estimate the linear regression $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

e. Plot ln(*FOODAWAY*) against *INCOME*, and include the fitted line from part (d).

f. Calculate the least squares residuals from the estimation in part (d). Plot them vs. *INCOME*. Do you find any unusual patterns, or do they seem completely random?

(a)



mean = 49.27085

median = 32.555

quantiles

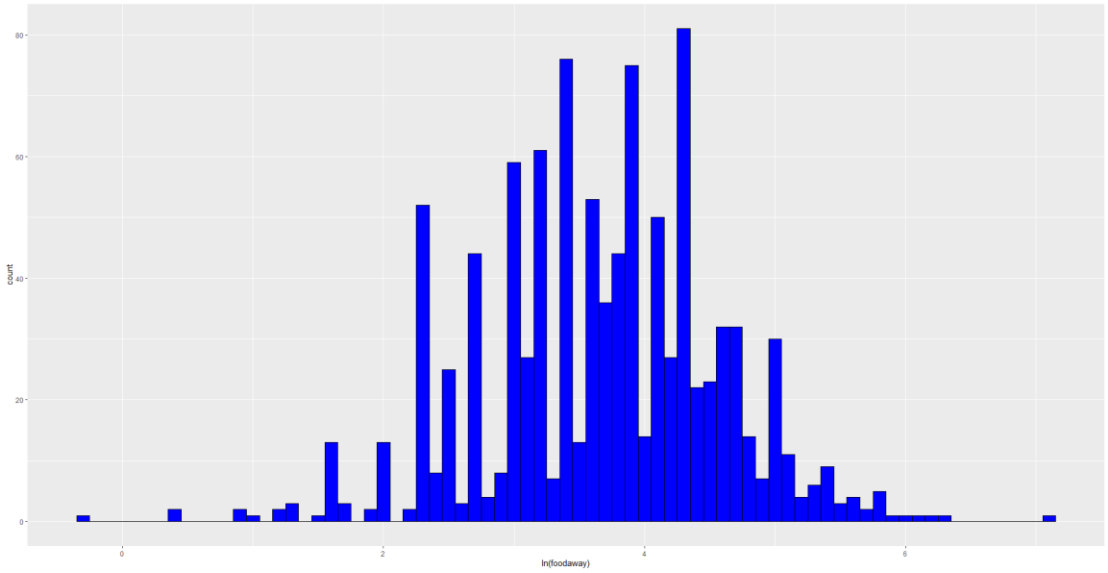
25% 75%

12.0400 67.5025

(b)

	Advance_degree	College_degree	No
Mean	73.15494	48.59718	39.01017
Median	48.15	36.11	26.02

(c)



ln(foodaway)

Min.　:-0.3011

1st Qu.: 3.0759

Median : 3.6865

Mean　: 3.6508

3rd Qu.: 4.2797

Max.　: 7.0724

NA's　:178

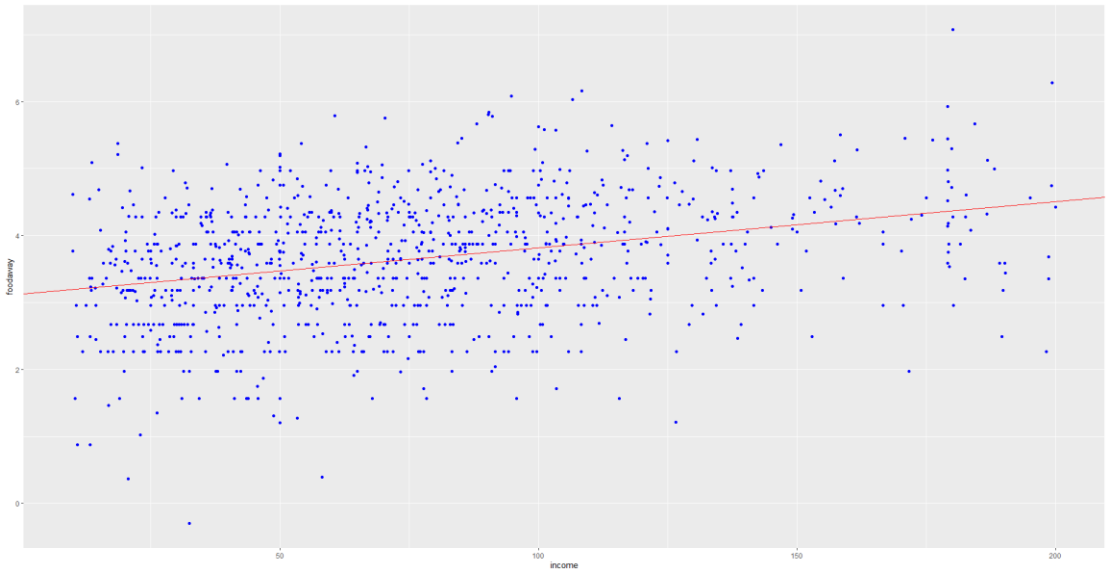
若原本 foodaway 的值為 0 取 ln 的值會趨近負無限大，這是造成缺失值有 178 個的原因

(d)

$\ln(\widehat{FOODAWAY}) = 3.1293004 + 0.0069017 \times INCOME$

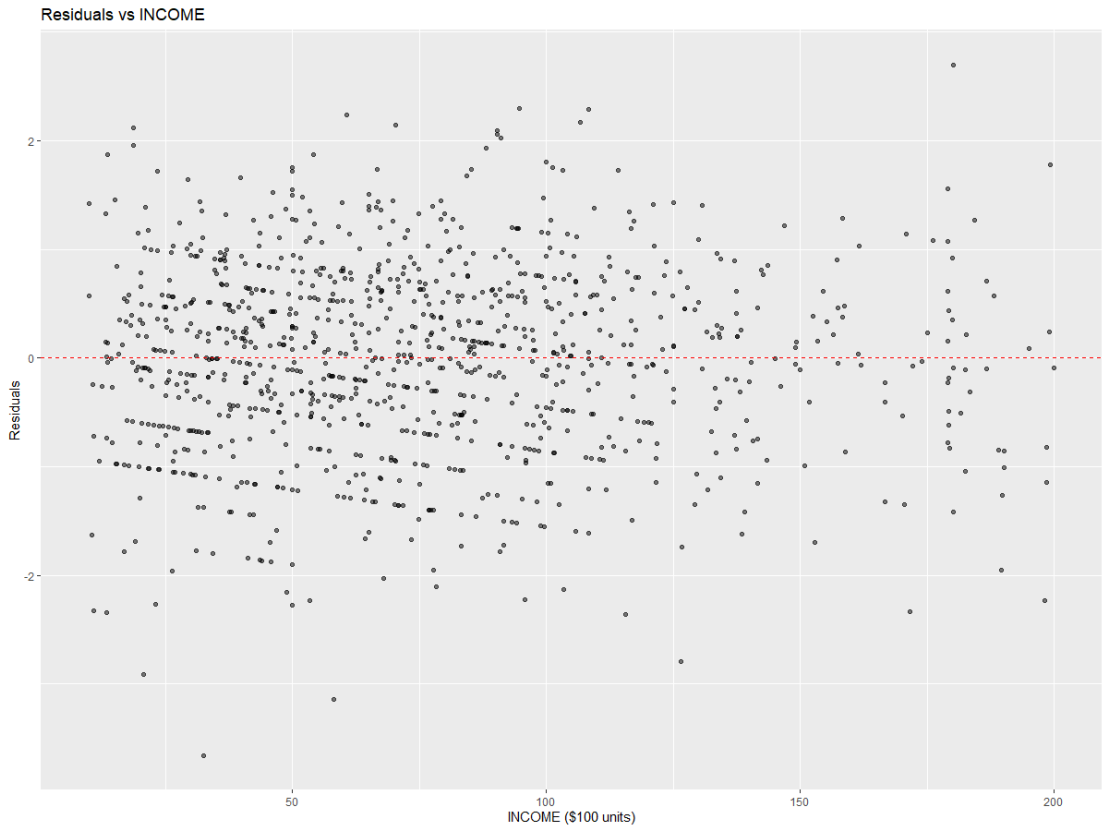
Slope =0.0069017 代表當 INCOME 增加一單位($100 units)，FOODAWAY 增加 0.6%

(e)



(f)

The least squares residuals=782.9716



有一些超過 2 或-2 的離群值 INCOME<100 的點較>100 的點密集

```
        Runs Test

data:  m1$residuals
statistic = -0.50073, runs = 504, n1 = 511, n2 = 511, n = 1022, p-value = 0.6166
alternative hypothesis: nonrandomness
```

p-value>0.05 所以殘差項具隨機性

2.28 How much does education affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version.]

- **a.** Obtain the summary statistics and histograms for the variables *WAGE* and *EDUC*. Discuss the data characteristics.
- **b.** Estimate the linear regression $WAGE = \beta_1 + \beta_2 EDUC + e$ and discuss the results.
- **c.** Calculate the least squares residuals and plot them against *EDUC*. Are any patterns evident? If assumptions SR1–SR5 hold, should any patterns be evident in the least squares residuals?
- **d.** Estimate separate regressions for males, females, blacks, and whites. Compare the results.
- **e.** Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 16 years of education. Compare these values to the estimated marginal effect of education from the linear regression in part (b).
- **f.** Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *WAGE* and *EDUC*. Which model appears to fit the data better?

a.

EDUC

Min. : 0.0

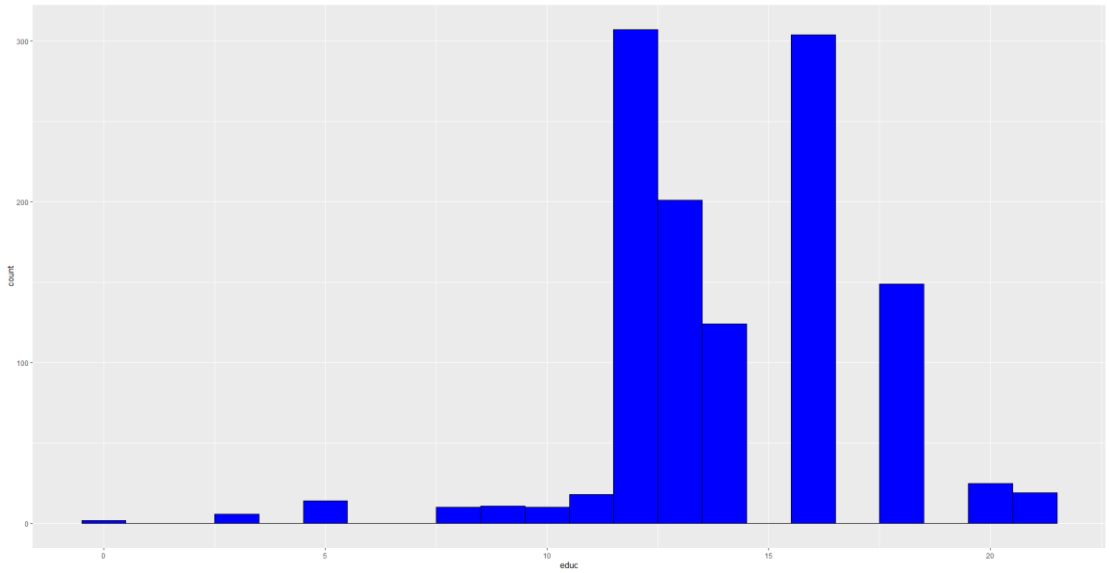
1st Qu.:12.0

Median :14.0

Mean :14.2

3rd Qu.:16.0

Max. :21.0



WAGE

Min. : 3.94

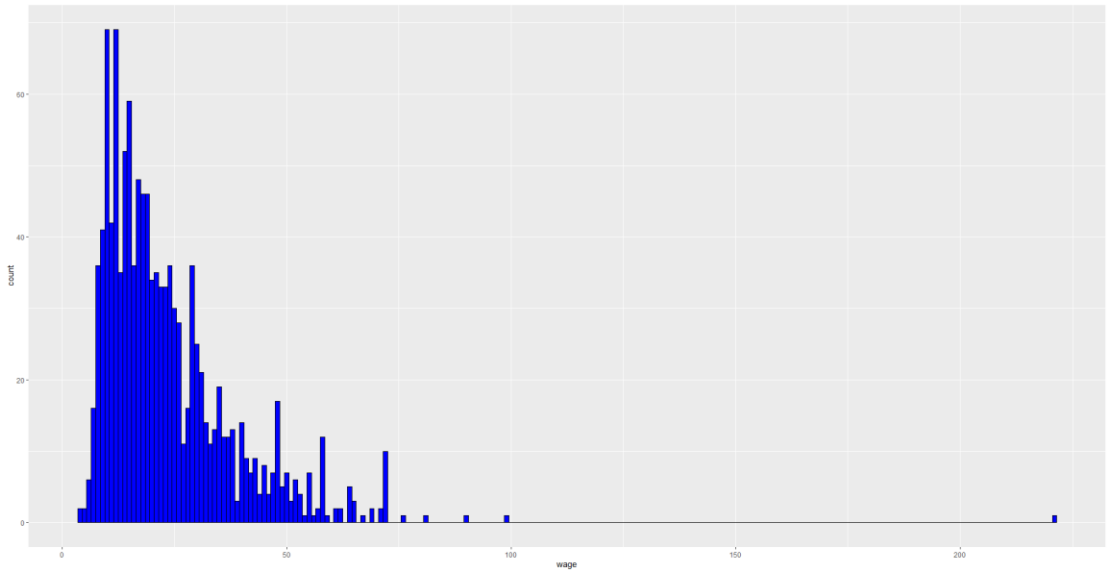
1st Qu.: 13.00

Median : 19.30

Mean : 23.64

3rd Qu.: 29.80

Max. :221.10



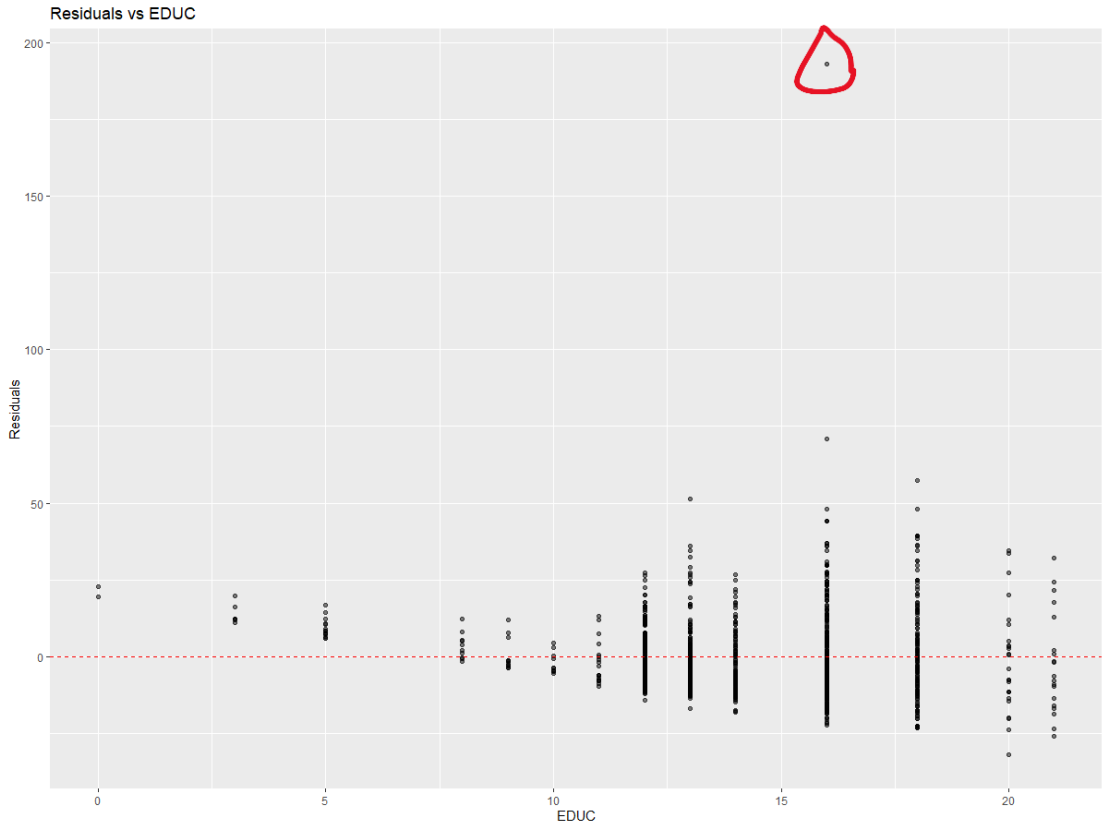
(b)

```
Call:
lm(formula = wage ~ educ, data = cps5_small)

Residuals:
    Min      1Q  Median      3Q     Max
-31.785  -8.381  -3.166   5.708 193.152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.4000     1.9624    -5.3 1.38e-07 ***
educ          2.3968     0.1354    17.7  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.55 on 1198 degrees of freedom
Multiple R-squared:  0.2073,    Adjusted R-squared:  0.2067
F-statistic: 313.3 on 1 and 1198 DF,  p-value: < 2.2e-16
```

$\widehat{WAGE} = -10.4 + 2.3968 \times \mathrm{E}DUC$

C.

SSE = 220062.3



有一個離群值特別明顯

若 SR1-SR5 假設皆成立，殘差圖不應該出現差距如此大的離群值

d.

Female

```
lm(formula = wage ~ educ, data = female)

Residuals:
    Min      1Q  Median      3Q     Max
-30.837  -6.971  -2.811   5.102  49.502

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.6028     2.7837  -5.964 4.51e-09 ***
educ          2.6595     0.1876  14.174  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.5 on 526 degrees of freedom
Multiple R-squared:  0.2764,    Adjusted R-squared:  0.275
F-statistic: 200.9 on 1 and 526 DF,  p-value: < 2.2e-16
```

$\widehat{WAGE} = -16.6028 + 2.6595 \times \mathrm{E}DUC$

Male

```
lm(formula = wage ~ educ, data = male)

Residuals:
    Min      1Q  Median      3Q     Max
-27.643  -9.279  -2.957   5.663 191.329

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.2849     2.6738  -3.099  0.00203 **
educ          2.3785     0.1881  12.648  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.71 on 670 degrees of freedom
Multiple R-squared:  0.1927,    Adjusted R-squared:  0.1915
F-statistic:   160 on 1 and 670 DF,  p-value: < 2.2e-16
```

$\widehat{WAGE} = -8.2849 + 2.3785 \times \mathrm{E}DUC$

Black

```
lm(formula = wage ~ educ, data = black)

Residuals:
    Min      1Q  Median      3Q     Max
-15.673  -6.719  -2.673   4.321  40.381

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.2541     5.5539  -1.126    0.263
educ          1.9233     0.3983   4.829 4.79e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.51 on 103 degrees of freedom
Multiple R-squared:  0.1846,    Adjusted R-squared:  0.1767
F-statistic: 23.32 on 1 and 103 DF,  p-value: 4.788e-06
```

$\widehat{WAGE} = -6.2541 + 1.9233 \times \mathrm{E}DUC$

White

```
lm(formula = wage ~ educ, data = white)

Residuals:
    Min      1Q  Median      3Q     Max
-32.131  -8.539  -3.119   5.960 192.890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -10.475      2.081  -5.034  5.6e-07 ***
educ           2.418      0.143  16.902  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.79 on 1093 degrees of freedom
Multiple R-squared:  0.2072,    Adjusted R-squared:  0.2065
F-statistic: 285.7 on 1 and 1093 DF,  p-value: < 2.2e-16
```

$\widehat{WAGE} = -10.475 + 2.418 \times \mathrm{E}DUC$

e.

```
lm(formula = wage ~ I(educ^2), data = cps5_small)

Residuals:
    Min      1Q  Median      3Q     Max
-34.820  -8.117  -2.752   5.248 193.365

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.916477   1.091864   4.503 7.36e-06 ***
I(educ^2)   0.089134   0.004858  18.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.45 on 1198 degrees of freedom
Multiple R-squared:  0.2194,    Adjusted R-squared:  0.2187
F-statistic: 336.6 on 1 and 1198 DF,  p-value: < 2.2e-16
```

$\widehat{WAGE} = 4.916477 + 0.089134 \times EDUC^2$

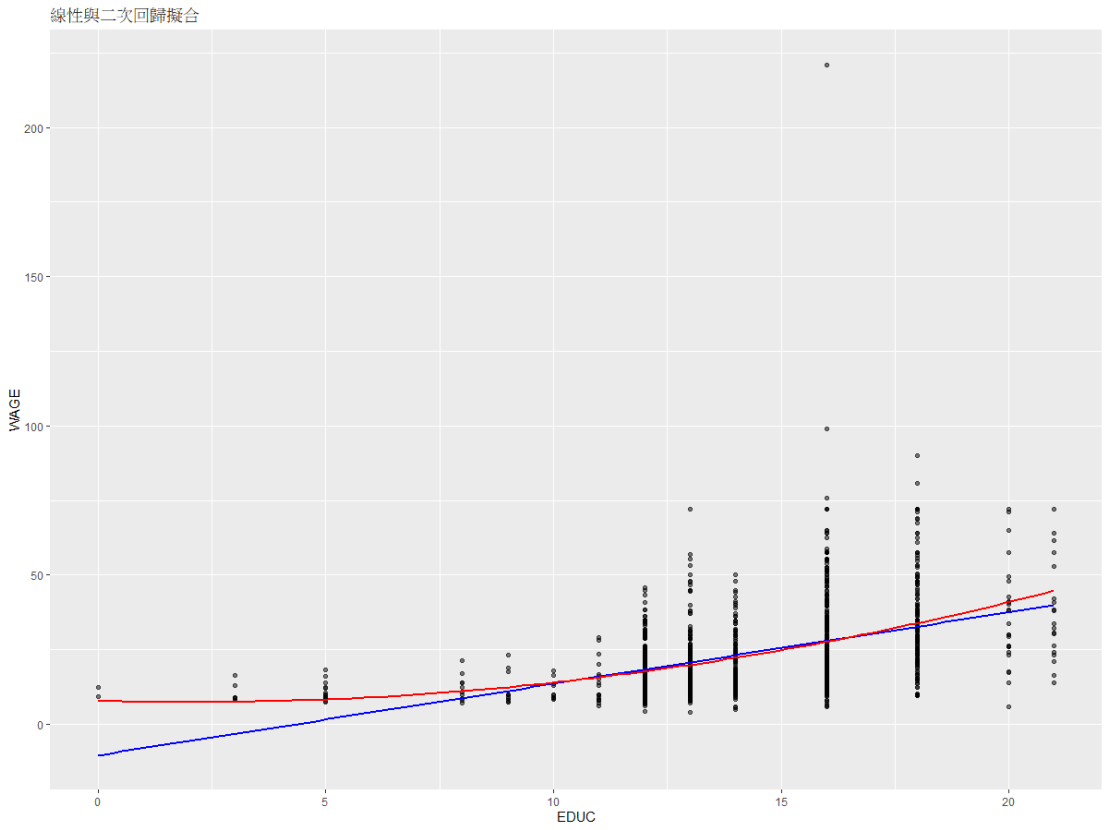
EDUC=12

$\frac{dWAGE}{dEDUC}$ = 2*a2*EDUC = 2*0.089134*12 = 2.139216

EDUC=16

$\frac{dWAGE}{dEDUC}$ = 2*a2*EDUC = 2*0.089134*16 = 2.852288

f.



The quadratic model from part (e) appears to fit the data better.