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.

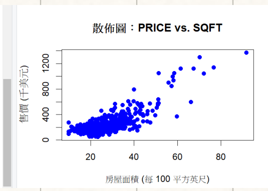
e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

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.)

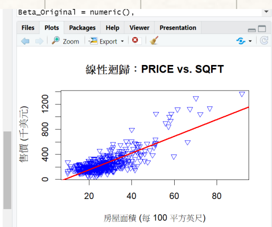
```
  price  sqft age pool fireplace close twostory occupied
1  96.5 16.77  10    0         0     0        0        0
2 106.0 17.51  10    0         0     0        0        0
3 114.0 18.37   9    0         0     0        0        1
4 119.9 29.10   9    0         0     0        0        0
5 122.7 14.68   9    0         0     0        0        1
6 125.0 18.23   8    0         0     0        0        0
> plot(collegetown$sqft, collegetown$price,
+      xlab = "房屋面積 (每 100 平方英尺)",
+      ylab = "售價 (千美元)",
+      main = "散佈圖：PRICE vs. SQFT",
+      pch  = 19,         # 散點形狀
+      col  = "blue")     # 散點顏色
錯誤: 找不到物件 'collegetown'
> plot(c_t_data$sqft, c_t_data$price,
+      xlab = "房屋面積 (每 100 平方英尺)",
+      ylab = "售價 (千美元)",
+      main = "散佈圖：PRICE vs. SQFT",
+      pch  = 19,         # 散點形狀
+      col  = "blue")     # 散點顏色
> |
```



(b.) $PRICE = -115.4236 + 13.4029\, SQFT$

房屋面積每增加1單位100平方英尺，售價會增加13.4029千美元

```
Call:
lm(formula = price ~ sqft, data = c_t_data)

Residuals:
    Min      1Q  Median      3Q     Max
-316.93  -58.90   -3.81   47.94  477.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -115.4236    13.0882  -8.819   <2e-16 ***
sqft          13.4029     0.4492  29.840   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102.8 on 498 degrees of freedom
Multiple R-squared:  0.6413,    Adjusted R-squared:  0.6406
F-statistic: 890.4 on 1 and 498 DF,  p-value: < 2.2e-16

> plot(c_t_data$sqft, c_t_data$price,
+      xlab = "房屋面積 (每 100 平方英尺)",
+      ylab = "售價 (千美元)",
+      main = "線性迴歸：PRICE vs. SQFT",
+      pch  = 15, col = "blue")
> abline(lin_model, col = "red", lwd = 2)
```



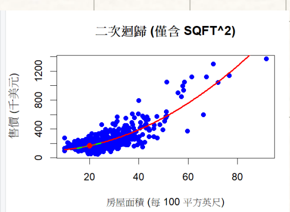
(c.) $PRICE = 93.5659 + 0.1845\, SQFT^2$

```
Call:
lm(formula = price ~ I(sqft^2), data = c_t_data)

Residuals:
    Min      1Q  Median      3Q     Max
-383.67  -48.39   -7.50   38.75  469.70

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.565854   6.072226   15.41   <2e-16 ***
I(sqft^2)    0.184519   0.005256   35.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 92.08 on 498 degrees of freedom
Multiple R-squared:  0.7122,    Adjusted R-squared:  0.7117
F-statistic:  1233 on 1 and 498 DF,  p-value: < 2.2e-16

> # 產生 SQFT 範圍內的數據點
```

$\frac{\partial PRICE}{\partial SQFT} = 2 \times 0.1845 \times SQFT$，若 SQFT=20，Marginal Effect = 7.38

```
> alpha2 <- coef(quad_only_model)["I(sqft^2)"]
> ME_2000 <- 2 * alpha2 * 20   # 2000 sq ft = 20 * (100 sq ft)
> ME_2000
I(sqft^2)
  7.38076
```

(d.) 當SQFT=20時，切線斜率=7.38

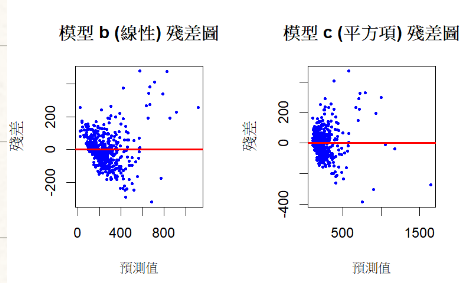


(e.) $\hat{\varepsilon} = 2 \times 0.1845 \times \frac{20^2}{93.5659 + 0.1845 \times 20^2} = 0.8819$

The elasticity is

$$\varepsilon = \frac{\Delta y / y}{\Delta x / x} = \frac{\Delta y}{\Delta x}\frac{x}{y} = m\frac{SQFT}{PRICE} = (2\beta_2 SQFT)\frac{SQFT}{PRICE} = 2\beta_2\frac{SQFT^2}{PRICE} = 2\beta_2\frac{SQFT^2}{\beta_1 + \beta_2 SQFT^2}$$

(f.)



當SQFT愈大時，殘差愈大，違反同質變異數假設。

(g.)

```
> # 取得殘差平方和 (RSS = SSE)
> sse_lin  <- sum(resid(lin_model)^2)
> sse_quad <- sum(resid(quad_only_model)^2)
>
> sse_lin
[1] 5262847
> sse_quad
[1] 4222356
```

SSE(Model B) > SSE(Model C)

⇒ Model C is better than Model B.

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?

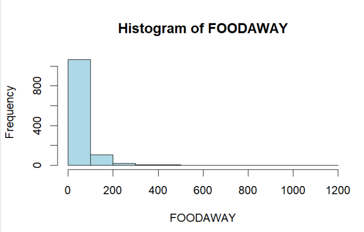
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

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)

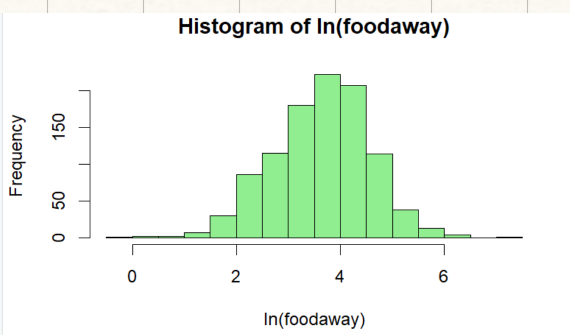



```
> # 直方圖
> hist(data$foodaway, main = "Histogram of FOODAWAY", 
ue")
>
> # 摘要統計
> summary(data$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   12.04   32.55   49.27   67.50 1179.00 
>
> # 25th 和 75th 百分位數
> quantile(data$foodaway, probs = c(0.25, 0.75))
    25%     75% 
12.0400 67.5025 
```

(b)

```
> # 平均值和中位數（大學學歷）
> mean_college <- mean(data$foodaway[data$college == 1], na.rm = TRUE)
> median_college <- median(data$foodaway[data$college == 1], na.rm = TRUE)
>
> # 平均值和中位數（無學位）
> mean_no_degree <- mean(data$foodaway[data$advanced == 0 & data$college == (
TRUE)
> median_no_degree <- median(data$foodaway[data$advanced == 0 & data$college
rm = TRUE)
>
> # 顯示結果
> cat("進階學位 - 平均值:", mean_advanced, "中位數:", median_advanced, "\n")
進階學位 - 平均值: 73.15494 中位數: 48.15
> cat("大學學位 - 平均值:", mean_college, "中位數:", median_college, "\n")
大學學位 - 平均值: 48.59718 中位數: 36.11
> cat("無學位 - 平均值:", mean_no_degree, "中位數:", median_no_degree, "\n")
無學位 - 平均值: 39.01017 中位數: 26.02
```

(c)

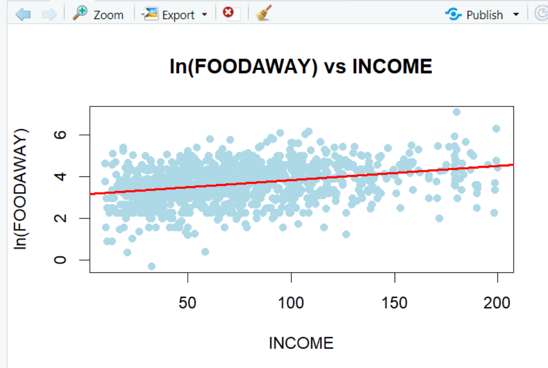



```
 # 列出刪除的資料數量
 cat("刪除的資料數量：", deleted_count, "\n")
刂除的資料數量： 178
 hist(valid_data$lnfoodaway, main = "Histogram of
)", col = "lightgreen")
 # 計算對數後的摘要統計
 summary(valid_data$lnfoodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3011  3.0759  3.6865  3.6508  4.2797  7.0724 
```

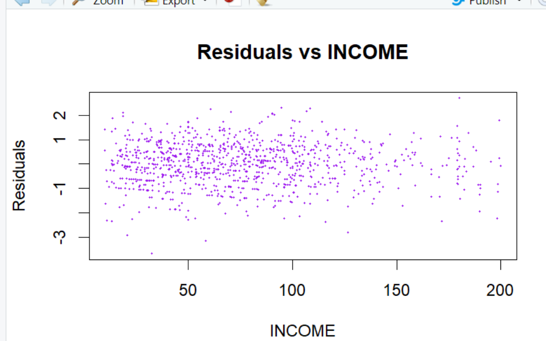
因為負值和零無法取對數,故刪除178筆data

(d.) (e) $\ln FOODAWAY = 3.1293 + 0.0069\ Income$

```
> summary(model)

Call:
lm(formula = lnfoodaway ~ income, data = valid_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6547 -0.5777  0.0530  0.5937  2.7000 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.1293004  0.0565503   55.34   <2e-16 ***
income      0.0069017  0.0006546   10.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8761 on 1020 degrees of freedom
Multiple R-squared:  0.09826,	Adjusted R-squared:  0.09738 
F-statistic: 111.1 on 1 and 1020 DF,  p-value: < 2.2e-16

>
> # 解釋斜率
> cat("斜率解釋：當收入增加 100 美元時，預期 ln(FOODAWAY) 變化量為", coef(model)[2], "\n")
斜率解釋：當收入增加 100 美元時，預期 ln(FOODAWAY) 變化量為 0.006901748
> # 繪圖
> plot(data$income, data$lnfoodaway, main = "ln(FOODAWAY) vs INCOME",
+      xlab = "INCOME", ylab = "ln(FOODAWAY)", pch = 16, col = "lightblue")
> abline(model, col = "red", lwd = 2)
```



(f)




```
> residuals <- residuals(model)
>
> # 計算最小殘差平方和 (RSS)
> RSS <- sum(residuals^2)
> cat("最小殘差平方和 (RSS)：", RSS, "\n")
最小殘差平方和 (RSS)： 782.9716
```

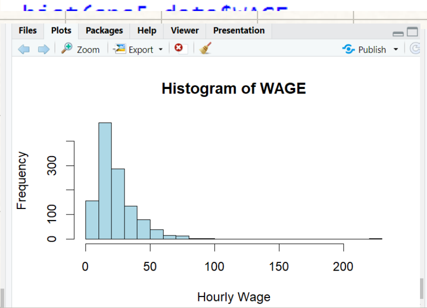
→殘差項 seem completely random

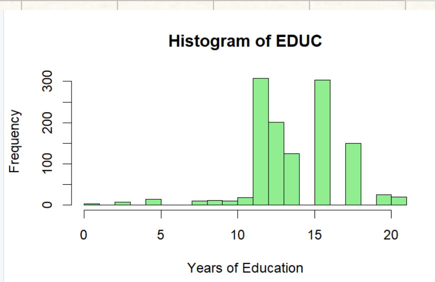
2.28 How much does education affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version.]

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)

```
t , south , WAGE , WEST )
> # 摘要統計
> summary(cps5_data$WAGE)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.94   13.00   19.30   23.64   29.80  221.10
> summary(cps5_data$EDUC)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.0    12.0    14.0    14.2    16.0    21.0
>
> # WAGE 直方圖
```





右偏 左偏

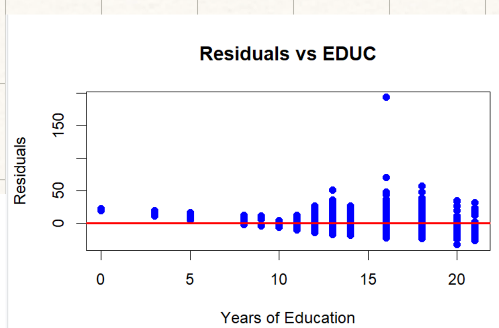
(b)

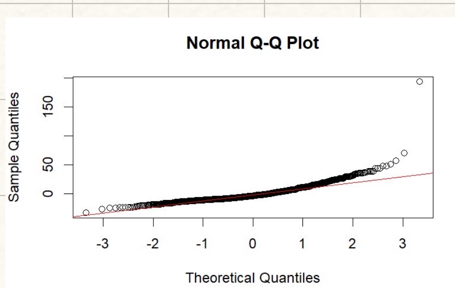
WAGE= -10.4+2.3968EDUC

```
Call:
lm(formula = WAGE ~ EDUC, data = cps5_data)

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

>
> # 斜率解釋
> cat("解釋：每增加一年教育，工資增加約", coef(model_linear)[2], "美元。\n")
解釋：每增加一年教育，工資增加約 2.396761 美元。
```

(c)

```
> bptest(model_linear)

	studentized Breusch-Pagan test

data:  model_linear
BP = 7.4587, df = 1, p-value = 0.006313
```





不符合SR5

(d)

```
lm(formula = WAGE ~ EDUC + FEMALE, data = cps5_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.896  -8.054  -2.795   5.648 191.104 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -9.7714     1.9425   -5.03 5.65e-07 ***
EDUC          2.4855     0.1348   18.44  < 2e-16 ***
FEMALE       -4.2914     0.7845   -5.47 5.48e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.39 on 1197 degrees of freedom
Multiple R-squared:  0.2267,	Adjusted R-squared:  0.2254 
F-statistic: 175.4 on 2 and 1197 DF,  p-value: < 2.2e-16

>
> # 格式化顯示回歸方程式
> coef_values <- coef(model_gender)
> equation <- paste0("WAGE = ", round(coef_values[1], 4),
+                    " + ", round(coef_values[2], 4), " * EDUC",
+                    " + ", round(coef_values[3], 4), " * FEMALE")
> cat("性別回歸方程式：", equation, "\n")
性別回歸方程式： WAGE = -9.7714 + 2.4855 * EDUC + -4.2914 * FEMALE
> cat("截距 (β1)：", round(coef_values[1], 4), "\n")
截距 (β1)： -9.7714
> cat("教育斜率 (β2)：", round(coef_values[2], 4), "\n")
教育斜率 (β2)： 2.4855
> cat("性別斜率 (β3)：", round(coef_values[3], 4), "\n")
```

```
lm(formula = WAGE ~ EDUC + BLACK, data = cps5_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.933  -8.533  -3.068   5.771 192.951 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -9.9840     1.9731  -5.060 4.85e-07 ***
EDUC          2.3833     0.1355  17.595  < 2e-16 ***
BLACK        -2.5744     1.3852  -1.858   0.0633 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.54 on 1197 degrees of freedom
Multiple R-squared:  0.2096,	Adjusted R-squared:  0.2083 
F-statistic: 158.7 on 2 and 1197 DF,  p-value: < 2.2e-16

>
> # 格式化顯示回歸方程式
> coef_values <- coef(model_race)
> equation <- paste0("WAGE = ", round(coef_values[1], 4),
+                    " + ", round(coef_values[2], 4), " * EDUC",
+                    " + ", round(coef_values[3], 4), " * BLACK")
> cat("種族回歸方程式：", equation, "\n")
種族回歸方程式： WAGE = -9.984 + 2.3833 * EDUC + -2.5744 * BLACK
> cat("截距 (β1)：", round(coef_values[1], 4), "\n")
截距 (β1)： -9.984
> cat("教育斜率 (β2)：", round(coef_values[2], 4), "\n")
教育斜率 (β2)： 2.3833
> cat("種族斜率 (β3)：", round(coef_values[3], 4), "\n")
```

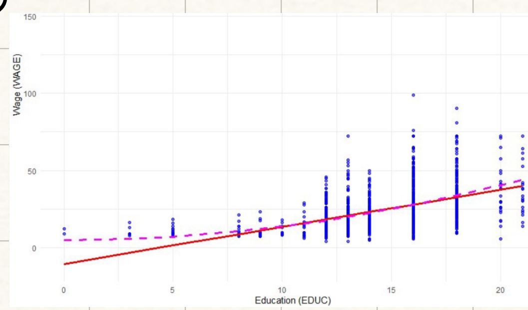
(e)

```
Call:
lm(formula = WAGE ~ I(EDUC^2), data = cps5_data)

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

>
> # 格式化顯示回歸方程式
> coef_values <- coef(model_quad_only)
> equation <- paste0("WAGE = ", round(coef_values[1], 4),
+                    " + ", round(coef_values[2], 4), " * EDUC^2")
> cat("平方項回歸方程式：", equation, "\n")
平方項回歸方程式： WAGE = 4.9165 + 0.0891 * EDUC^2
> cat("截距 (α1)：", round(coef_values[1], 4), "\n")
截距 (α1)： 4.9165
> cat("平方項係數 (α2)：", round(coef_values[2], 4), "\n")
```

12: marginal effect $= 2 \times 0.0891 \times 12$

16: $2 \times 0.0891 \times 16$

(f)



quadratic 比較 fitted.