**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.

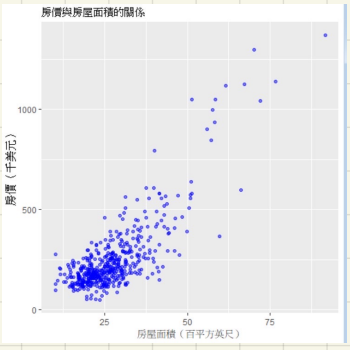
**d.** Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

**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?

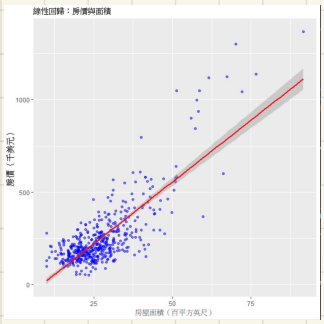
**g.** One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?

1a)




房價與房屋面積呈正相關

(b)




$Price = \beta_1 + \beta_2 SQFT + e$

$\widehat{Price} = -115.42 + 13.4 \cdot SQFT$

截距 (Intercept, $\beta_0$) $= -115.42$

↳ 當房屋面積 $= 0$，$\widehat{Price} = -115.42$ 千美元

斜率 ($\beta_1$) $= 13.4$

↳ 每增加 1 單位的 SQFT，房價平均上升 13.4 千美元

```
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 ***
---
```

(c)

```
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 ***
---
```

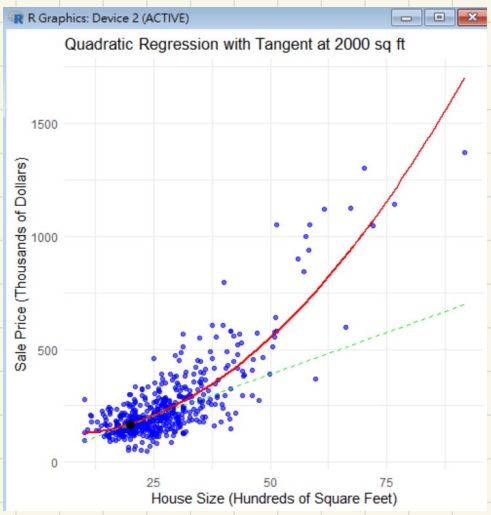
$Price = \alpha_1 + \alpha_2 SQFT^2 + e$

$\widehat{Price} = 93.5659 + 0.1845\ SQFT^2$

$\frac{d\,PRICE}{d\,SQFT} = 2 \times 0.1845 \times SQFT$

當房屋面積為2000平方英尺，每增加100平方英尺，房價約增加7380美元

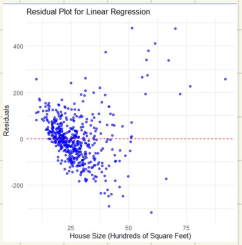
(d)



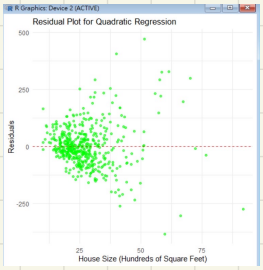
e. $\hat{\varepsilon} = 0.882$

```
> #2.17.e
> elasticity <- marginal_effect * (sqft_value / price_at_2000)
> elasticity
    sqft2
0.8819511
```

f.




當SQFT增加時，殘差的變異性變大，代表無法捕捉隨著房屋面積增長的價格變化




殘差分布仍有異方差性

↳ 同方差性(Homoskedasticity)假設可能被違反

g.

```
> #2.17.g
> sse_linear <- sum(collegetown$linear_residuals^2)
> sse_quad <- sum(collegetown$quad_residuals^2)
> sse_linear
[1] 5262847
> sse_quad
[1] 4222356
>
```

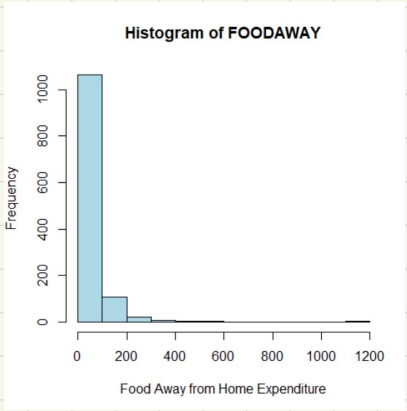
$sse\_linear = 5262847 > sse\_quadratic = 4222356$

⇒二次回歸模型擬合效果更好

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)




```
+
+   # 計算均值與中位數
+   mean_value <- round(mean(cex5_small$foodaway, na.rm = TRUE), 2)
+   median_value <- round(median(cex5_small$foodaway, na.rm = TRUE), 2)
+
+   # 計算 25th, 50th (median), 和 75th 百分位數
+   quantiles <- round(quantile(cex5_small$foodaway, probs = c(0.25, 0.50, 0.75$
+
+   # 輸出結果
+   print(paste("Mean of FOODAWAY:", mean_value))
+   print(paste("Median of FOODAWAY:", quantiles[2]))
+   print("25th, 50th, and 75th Percentiles:")
+   print(quantiles)
+ } else {
+   print("Error: FOODAWAY variable not found in dataset.")
+ }
[1] "Mean of FOODAWAY: 49.27"
[1] "Median of FOODAWAY: 32.56"
[1] "25th, 50th, and 75th Percentiles:"
  25%   50%   75%
12.04 32.56 67.50
```

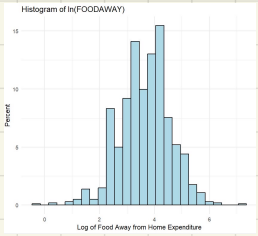
mean = 49.275 $25^{th} = 12.04$, $50^{th} = 32.56$, $75^{th} = 67.6$

(b)

```
[1] "--- FOODAWAY Statistics by Education Level ---"
[1] "Advanced Degree (N = 257 ): Mean = 73.15  Median = 48.15"
[1] "College Degree (N = 369 ): Mean = 48.6  Median = 36.11"
[1] "No College (N = 574 ): Mean = 39.01  Median = 26.02"
```

| | N | Mean | Median |
|---|---|---|---|
| advanced | 257 | 73.15 | 48.15 |
| college | 369 | 48.6 | 39.01 |
| None | 574 | 49.57 | 26.02 |

(c)




取 ln 可以修正右偏分布.

(d)

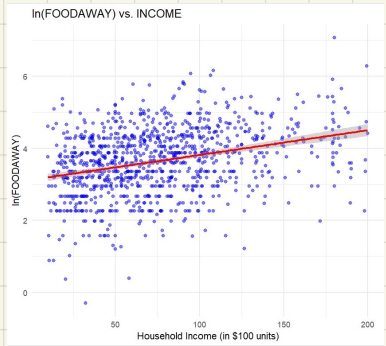
$\ln(FOODAWAY) = 3.1293 + 0.0069\ INCOME$

(se) (0.0566) (0.0007)

```
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 ***
```

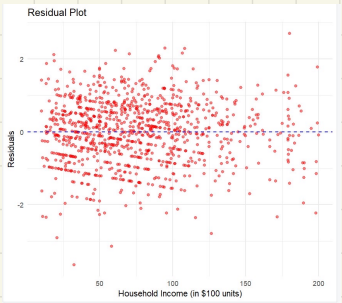
household income 增加 $100, food away expenditures per person 0.69%

(e)




ln(FOODWAY) 與收入呈正相關

(f)

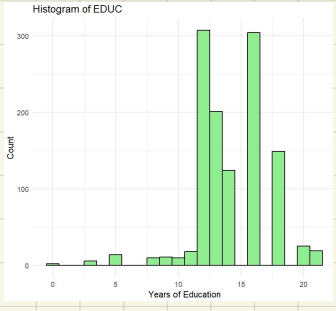


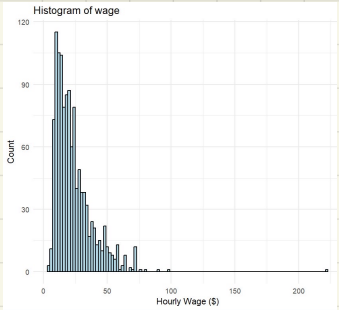

OLS 殘差沒有明顯的模式

**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?

(g)




教育年數與時薪可能存在一定關聯，高教育水平的人可能有更高的薪資

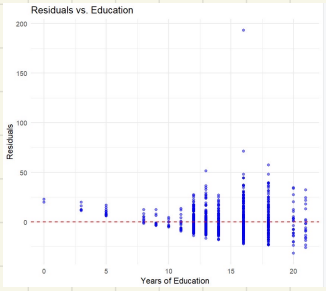
(b)

```
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
```

$\widehat{WAGE} = -10.4 + 2.3968\ EDUC$

EDUC 对 wage 有正向影響

(c)




EDUC and wage 並非完全線性，high EDUC 的 wage 波动較大

1. Heteroscedasticity 異方差問題：wage 的變異性隨教育年数增加而擴大

2. 可能存在非線性關係

(d)

1. 男性與女性的比較

| 群體 | 截距 ( Intercept ) | 教育係數 ( educ ) | $R^2$ ( 決定係數 ) | 標準誤 |
|---|---|---|---|---|
| 男性 | -8.28 | 2.38 | 0.1927 | 14.71 |
| 女性 | -16.60 | 2.66 | 0.2764 | 11.50 |

2. 黑人與白人的比較

| 群體 | 截距 ( Intercept ) | 教育係數 ( educ ) | $R^2$ ( 決定係數 ) | 標準誤 |
|---|---|---|---|---|
| 黑人 | -6.25 | 1.92 | 0.1846 | 10.51 |
| 白人 | -10.48 | 2.42 | 0.2072 | 13.79 |

EDUC 对 wage 影響力排序：女性 > 白人 > 男性 > 黑人

e.1

```
             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 ***
```

$WAGE = 4.916477 + 0.089134\ EDUC^2$

Marginal Effect : $\frac{d\ WAGE}{d\ EDUC} = 2 \times a_2 \times EDUC$

(b) 小題 $\frac{d\ WAGE}{d\ EDUC} = 2.3968$

這表示線性回歸認為每年教育對工資的影響是恆定的（約 2.40 美元），但二次回歸顯示影響是隨著教育增加而變大的：

在 EDUC = 12 時，二次回歸的影響較小（ 2.14 vs. 2.40 ）。
在 EDUC = 16 時，二次回歸的影響較大（ 2.85 vs. 2.40 ）。
這說明教育對工資的影響具有加速效應，較高教育水平的人獲得的工資增幅更大。

當 EDUC = 12

$\frac{d\ WAGE}{d\ EDUC} = 2 \times 0.0891 \times 12 = 2.138$

受教育 12 年，額外一年教育可增加 2.138 美元的 wage

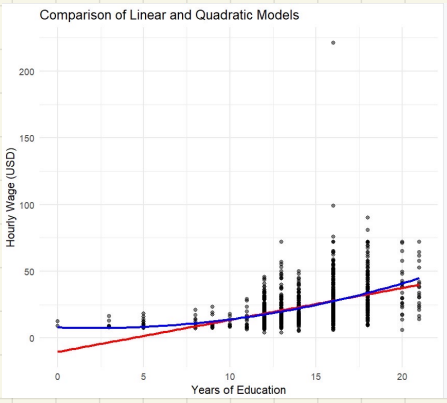
EDUC = 16

$\Rightarrow 2 \times 0.0891 \times 16 = 2.850$

" 16 年 "

增加 2.850 "

f.1




紅色曲線（線性回歸）：顯示教育年數與薪資之間的線性關係，即每多一年教育，薪資增長的幅度保持恆定。

藍色曲線（二次回歸）：顯示隨著教育年數增加，薪資增長的幅度也逐漸變大。

黑色點（觀測數據）：顯示數據的實際分布，薪資的變異性在高教育年數時變得更大。

線性模型（紅色）：假設教育的回報是固定的，無論受教育程度如何，每多一年教育，薪資增加的幅度都相同。
但從圖中可以看到，高教育水平（15 年以上）時，線性模型低估了實際薪資，說明這種假設可能不合理。

二次模型（藍色）：允許教育對薪資的影響隨著教育年數的增加而增強，即教育的邊際回報隨著時間推移變大。
從圖中可見，二次模型對於高教育群體的擬合效果比線性模型更好。

二次回歸模型（藍色）比線性模型（紅色）更適合描述教育與薪資之間的關係。