2.17

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.

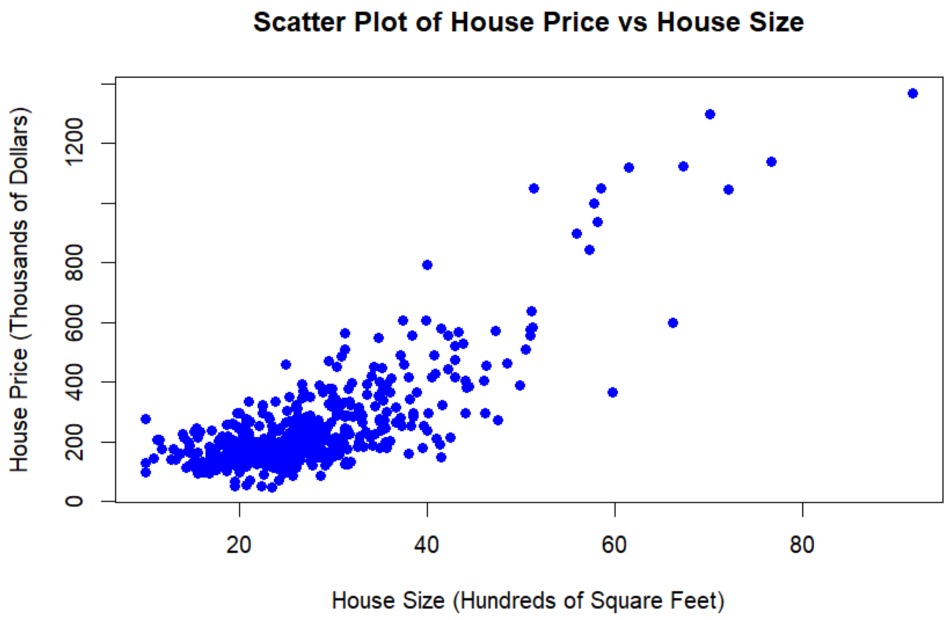
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

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)

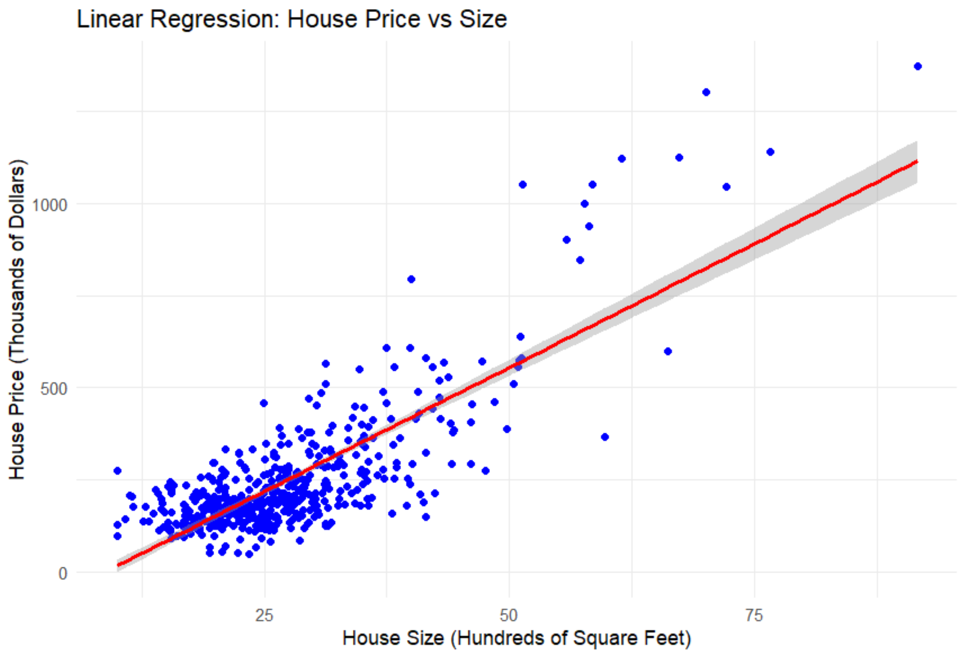
$$PRICE = \beta_1 + \beta_2 \cdot SQFT + e$$

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

$$(SE) \qquad (13.0882) \quad (0.4492)$$

在其他條件不變下，SQFT 每增加 100 平方英尺，預期房價將增加 13,402.9 美元。當 SQFT=0 時，預期房價為-115,423.6 美元。

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

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
```



(c)

$$PRICE = \alpha_1 + \alpha_2 \cdot SQFT^2 + e$$

$$\widehat{PRICE} = 93.565854 + 0.184519 \cdot SQFT$$

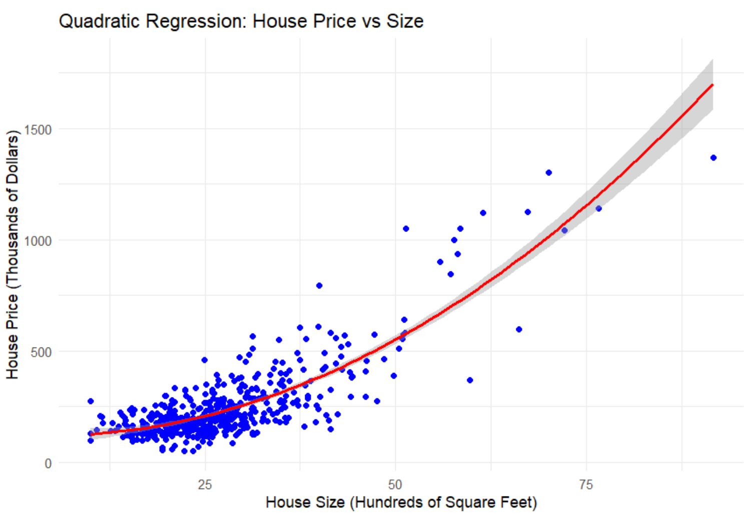
$$(SE) \qquad (6.072226) \qquad (0.005256)$$

在其他條件不變下，當 SQFT 達 2,000 平方英尺，SQFT 每增加 100 平方英尺，將會使預期房價增加 7,380.8 美元

```
Call:
lm(formula = price ~ sqft + sqft2, data = collegetown)

Residuals:
    Min      1Q  Median      3Q     Max 
-386.71  -48.66   -8.78   37.64  472.80 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 124.68914   24.49255   5.091 5.07e-07 ***
sqft         -1.87040    1.42603  -1.312     0.19    
sqft2         0.20796    0.01863  11.163  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.01 on 497 degrees of freedom
Multiple R-squared:  0.7132,	Adjusted R-squared:  0.7121 
F-statistic:   618 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
> # 計算邊際效應（額外 100 sqft 對價格的影響）
> sqft_2000 <- 20  # 2000 square feet -> 20 in hundreds
> marginal_effect_100 <- coef(lm_quad)["sqft"] + 2 * coef(lm_quad)["sqft2"] * sqft_2000
> cat("Marginal effect of additional 100 sqft at 2000 sqft:", marginal_effect_100, "\n")
Marginal effect of additional 100 sqft at 2000 sqft: 6.448092
```

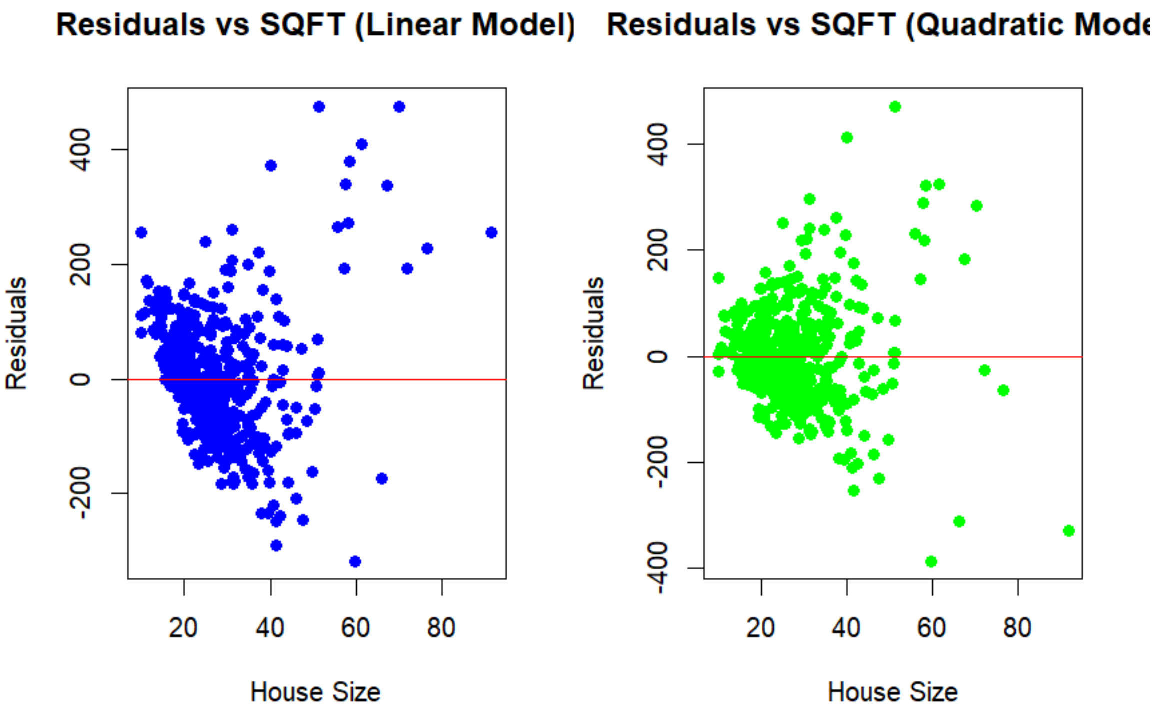
(d)



(e)

```
> elasticity_2000 <- (coef(lm_quad)["sqft"] + 2 * coef(lm_quad)["sqft2"] * sqft_2000) * (sqft_2000
/ predict(lm_quad, newdata = data.frame(sqft = sqft_2000, sqft2 = sqft_2000^2)))
> cat("Elasticity of price with respect to sqft at 2000 sqft:", elasticity_2000, "\n")
Elasticity of price with respect to sqft at 2000 sqft: 0.7565249
```

(f)



可以發現違反同質變異數假設，殘差不符合其同質變異數時的 scatter plot。

(g)

```
> cat("SSE (Linear Model):", sse_linear, "\n")
SSE (Linear Model): 5262847
> cat("SSE (Quadratic Model):", sse_quad, "\n")
SSE (Quadratic Model): 4207791
> if (sse_quad < sse_linear) {
+   cat("Quadratic model fits better as it has a lower SSE.\n")
+ } else {
+   cat("Linear model fits better as it has a lower SSE.\n")
+ }
Quadratic model fits better as it has a lower SSE.
```

2.25

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?

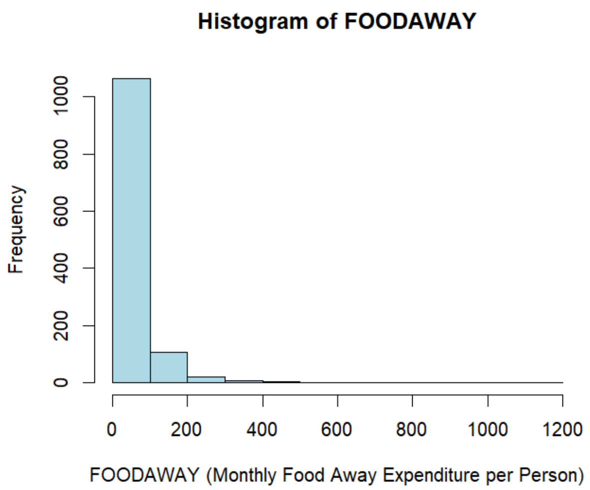
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

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)



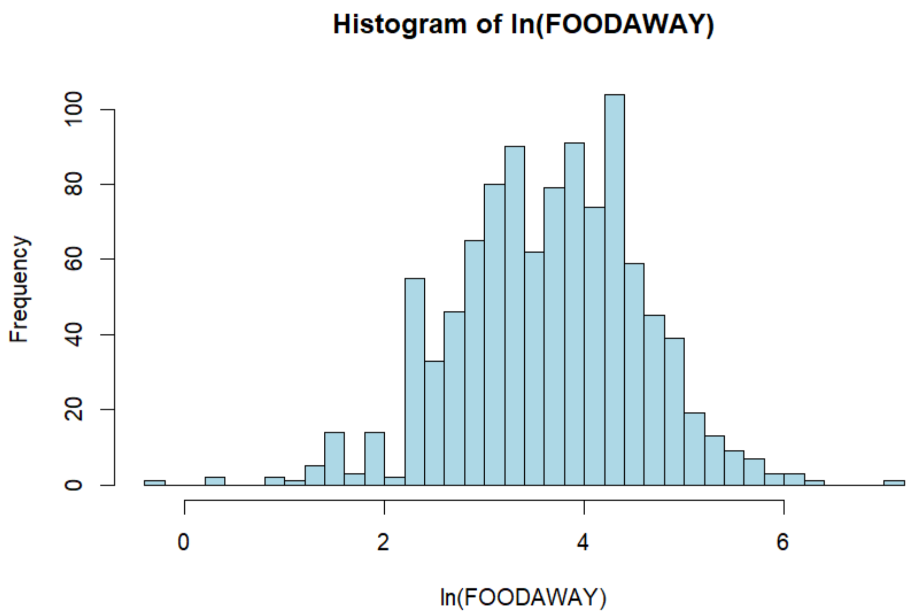
Values	
mean_foodaway	49.27085
median_foodaway	32.555
percentile_25	Named num 12
percentile_75	Named num 67.5

(b)

Values	
foodaway_advanced	num [1:257] 85 68.37 48.15 9.63 47.26 ...
foodaway_college	num [1:369] 39.8 0 108.3 66.1 19.3 ...
mean_advanced	73.1549416342412
mean_college	48.5971815718157
mean_foodaway	49.27085
median_advanced	48.15
median_college	36.11
median_foodaway	32.555
n_advanced	257L
n_college	369L
n_no_degree	574L
percentile_25	Named num 12
percentile_75	Named num 67.5
y_max	807L
y_ticks	num [1:4] 0 400 800 1200

(c)

```
> cat("Summary Statistics of ln(FOODAWAY):\n")
Summary Statistics of ln(FOODAWAY):
> cat("Mean:", mean_log_foodaway, "\n")
Mean: 3.650804
> cat("Median:", median_log_foodaway, "\n")
Median: 3.686499
> cat("Min:", min_log_foodaway, "\n")
Min: -0.3011051
> cat("Max:", max_log_foodaway, "\n")
Max: 7.072422
> cat("25th Percentile:", q1_log_foodaway, "\n")
25th Percentile: 3.075929
> cat("75th Percentile:", q3_log_foodaway, "\n")
75th Percentile: 4.279717
> cat("Number of Observations (ln(FOODAWAY)):", n_log_foodaway, "\n")
Number of Observations (ln(FOODAWAY)): 1022
> cat("Number of Observations (FOODAWAY):", n_foodaway, "\n")
Number of Observations (FOODAWAY): 1200
```



由於我們取自然對數時,若函數裡面是 0 會導致 negative infinite 出現，因此我們必須剔除 FOODAWAY 是 0 的值,故 FOODAWAY 和 ln(FOODAWAY)會有不同的觀測值數目。

(d)

Regression Model:

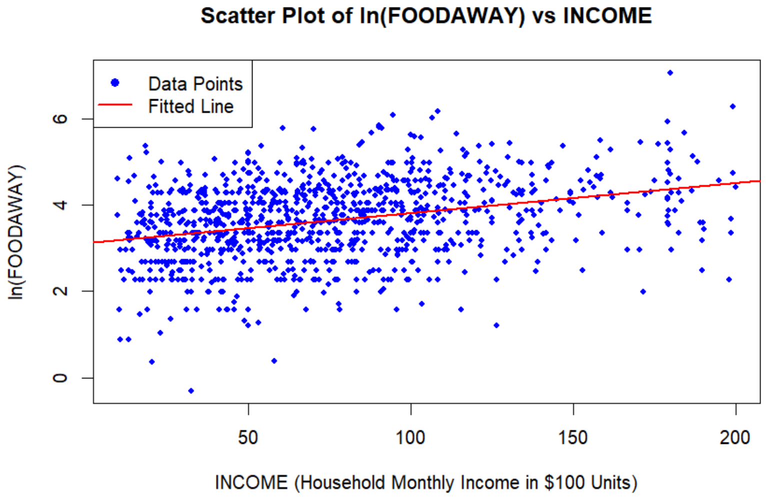
```
Call:
lm(formula = log_foodaway ~ income, data = clean_data)

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
```

interpretation:

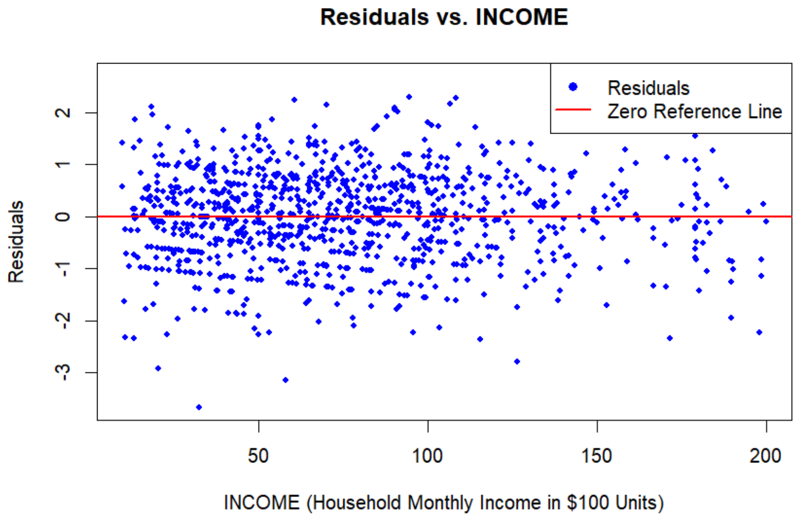
由結果可知 estimated slope = 0.0069 並且顯著,表示在其他條件不變之

下,若 income 上升$100(題幹有給單位),平均而言每人每月外出用餐的支出會變動 1%(as it is a log-linear model)。

(e)



(f)



整體而言,殘差圖看起來是完全隨機的,此回歸應當符合古典回歸假設,即殘差平均數為零,且符合同質變異數假設。

2.28

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?

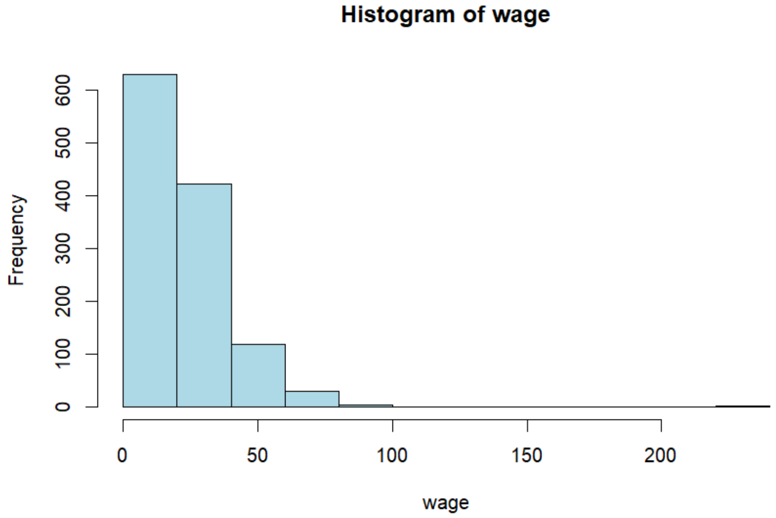
d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

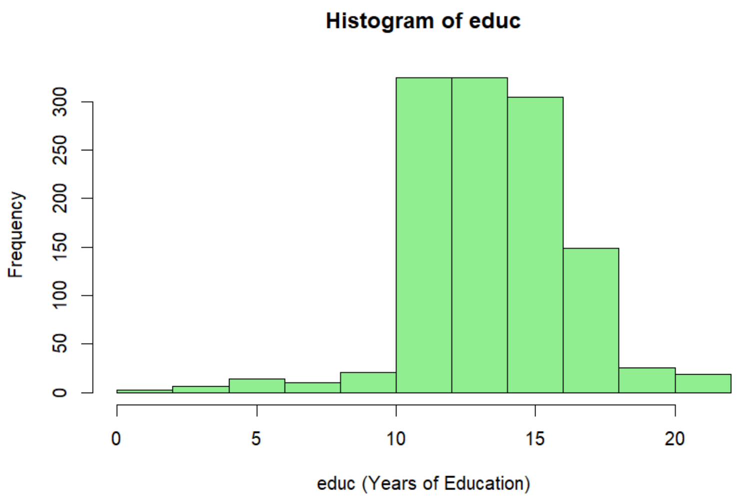
e. Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 16 years of education. Compare these values to the estimated marginal effect 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)

```
> summary(df[, c("wage", "educ")])
      wage             educ     
 Min.   :  3.94   Min.   : 0.0  
 1st Qu.: 13.00   1st Qu.:12.0  
 Median : 19.30   Median :14.0  
 Mean   : 23.64   Mean   :14.2  
 3rd Qu.: 29.80   3rd Qu.:16.0  
 Max.   :221.10   Max.   :21.0  
```





由 summary 和直方圖可以發現,薪資水平式明顯的右偏分配,上限很高,顯見貧富差距還算明顯。而教育年限相對只有些微左偏,因為大部分人都會受義務教育,僅部分人口會完全沒受教育或提早離開校園。

(b)

$$\widehat{WAGE} = -10.4000 + 2.3968 \cdot EDUC$$
$$(SE) \qquad (1.9624) \quad (0.1354)$$

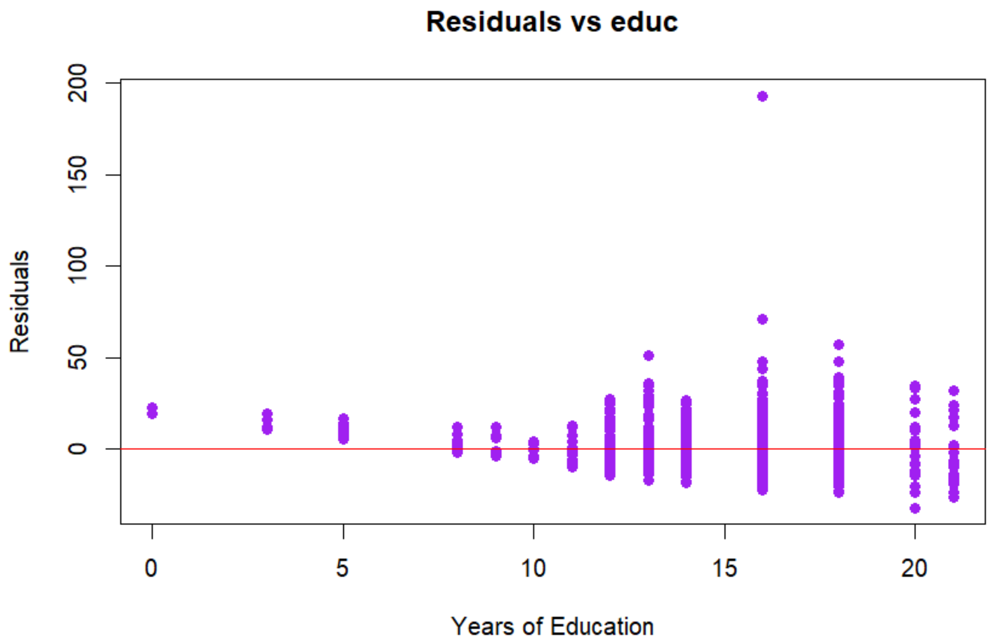
Interpretation:

由工資對受教育年限的結果我們可以看出,斜率是顯著的 2.3968,表示在其他條件不變之下,每多受教育一年,工資率平均而言會上升 2.3968 單位，也就是工資率和受教育年限應有正向關係。但由表中我們也可以看出簡單回歸的缺失,因為即便受教年限為 0 也不可能出現負數的工資率。

```
> summary(lm_linear)

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

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
```

(c)

由殘差圖可以發現當 Years of Education 增加時工資率的波動程度有變大的趨勢 。因此如果當 SR1-SR5 符合，不該出現變異數放大的殘差圖趨勢，故不符合同質變異數假說。



(d)

```
> summary(model_male)

Call:
lm(formula = wage ~ educ, data = cps5_small, subset = (female ==
    0))

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

> summary(model_female)

Call:
lm(formula = wage ~ educ, data = cps5_small, subset = (female ==
    1))

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
```

我們先從男性及女性來觀察與比較，兩個集合的工資率對受教育年限都具顯著的正斜率，表示其他條件不變之下，受教育程度越高時平均而言工資率對兩族群來說都會上升 。值得注意的是女性在這方面的受惠程度較高 。又我們看殘差的極大值會發現男性是驚人的 191.328 ,遠高於女性的 49.502，雖然只是單筆數據不足以代表，但可能的推測是工資率的發放有性別不平等的現象。

```
> summary(model_black)

Call:
lm(formula = wage ~ educ, data = cps5_small, subset = (black ==
    1))

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
```

```
> summary(model_white)

Call:
lm(formula = wage ~ educ, data = cps5_small, subset = (black ==
    0))

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
```

接下來來觀察黑人與白人的敘述統計及迴歸分析，兩者都具有顯著正斜率，顯示其他條件不變之下，受教育程度越高，平均而言工資率對兩群體來說都會上升。另外白人在這方面受惠程度較高，即便黑人多受教育能帶來的邊際效益也較低。

(e)

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

Interpretation:

計算邊際影響上，

12 years of education 下，每多一單位教育會增加薪資 2.1392 單位

16 years of education 下，每多一單位教育會增加薪資 2.852 單位

```
Call:
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

> cat("Marginal Effect at 12 years of education:", ME_12, "\n")
Marginal Effect at 12 years of education: 2.139216 
> cat("Marginal Effect at 16 years of education:", ME_16, "\n")
Marginal Effect at 16 years of education: 2.852288 
```

由上表可知 ,在其他條件不變之下,受教育年限的平方項每上升一單位,平均而言工資率會上升 0.089134 單位,而且我們使用 quadratic 的模型成功解決了截距項為負不合理的問題。原先受 12 年教育的人若多受一年教育,對工資率影響的效果平均而言是 0.089134*(13^2 – 12^2) = 2.22835 單位。原先受 16 年教育的人若多受一年教育,對工資率影響的效果平均而言是 0.089134*(17^2 – 16^2) = 2.941422 單位。相較(b)小題一般模型而言(無論當前受教育年限,每多受一年教育,平均工資率的增長都是斜率 2.3968 單位) ,使用平方項的回歸式較符合實際情況,因為所受的教育越高,帶來的邊際效益應該不相同。

(f)

紅線更貼合樣本點，顯示 Quadratic Model 可能更適合解釋樣本資料的特性