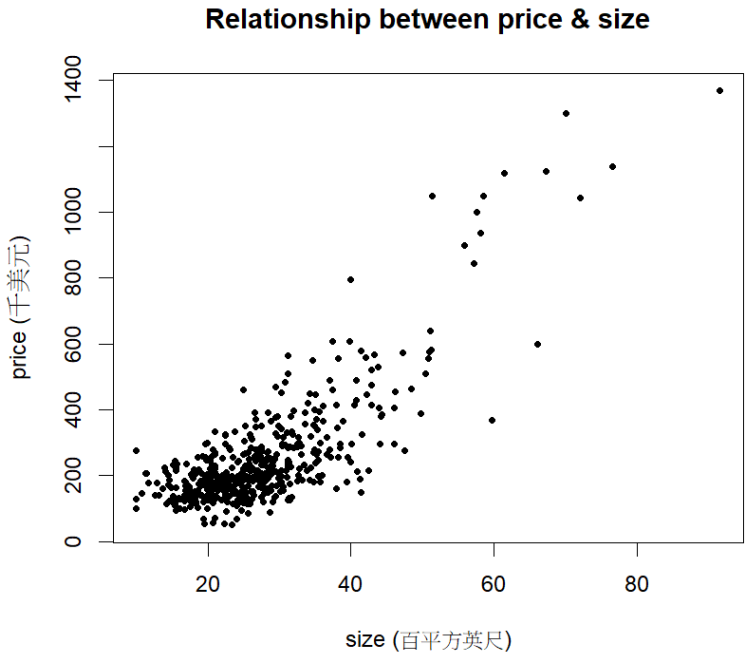
2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

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.

(b)

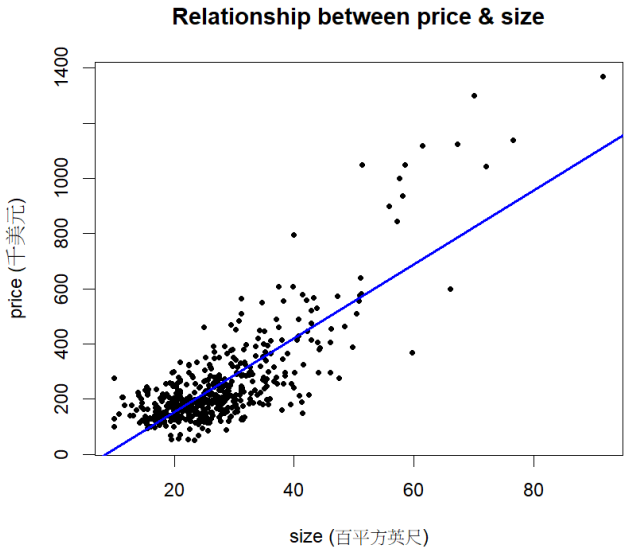
```
Call:
lm(formula = price ~ sqft, data = collegetown)

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
```

Intercept = -115.4326

表示當 sqft =0 時，一間房子的期望價值為-115.4326

sqrt's beta = 13.4026

房子的面積每增加 100 sqft，價格會增加 13.4026



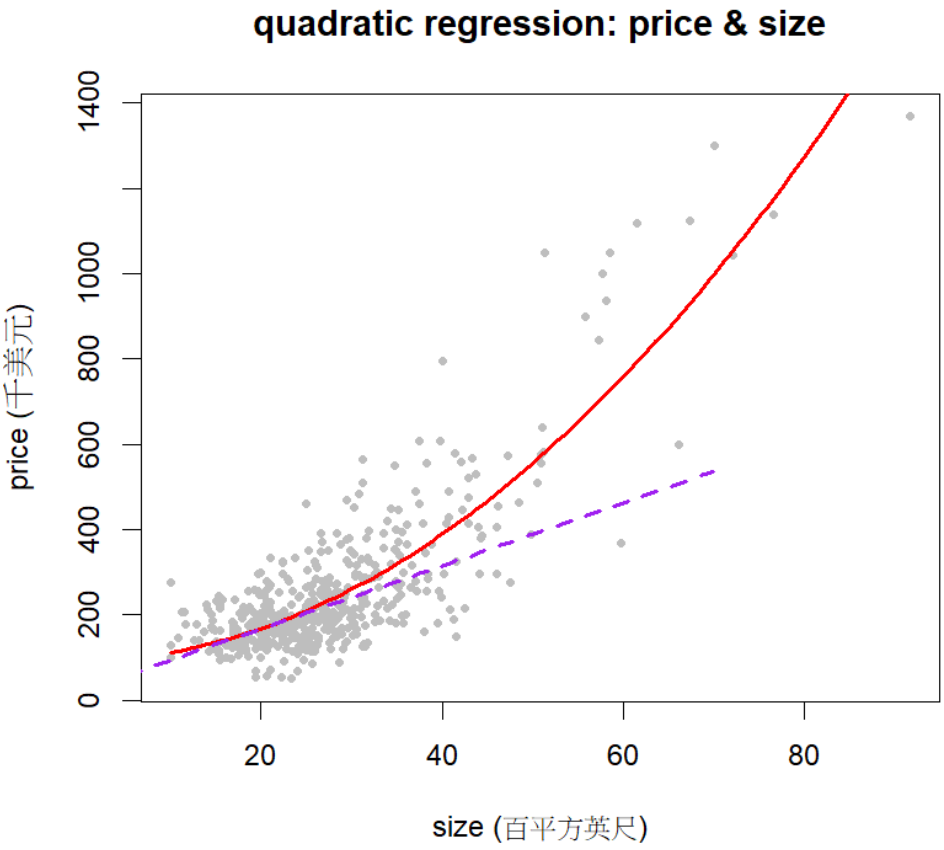
(c)

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

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

> marginal_effect_20 <- 2 * coef(quad_model)["I(sqft^2)"] * 20
> print(marginal_effect_20)
I(sqft^2)
  7.38076
```

Marginal effect = 7.38076

當房屋面積從 2000 平方英尺增加 100 平方英尺時，價格會增加 7.38076(千美元)

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.

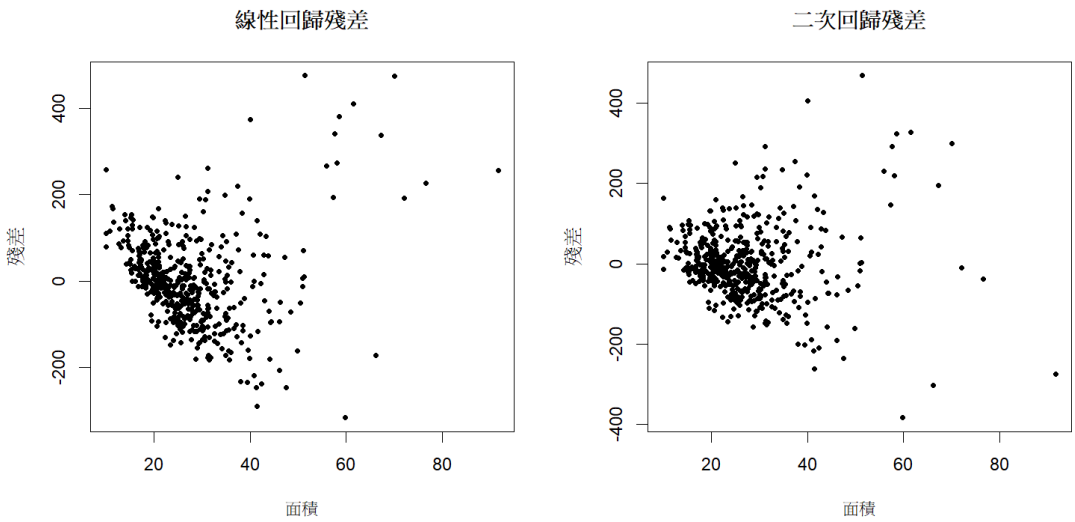


(e)

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.

```
> lines(tangent_x_vals, tangent_y_vals, col = "purple", lwd = 2, lty = 2)
> Elasticity <- (2 * alpha2 * sqft_2000) * (sqft_2000 / predicted_price_2000)
> print(Elasticity)
I(sqft^2)
0.8819511
```

(f)



殘差有擴散的跡象，隨面積(sqft)增加，殘差也跟著擴大，所以違反了 Homoskedasticity 的前提假設。

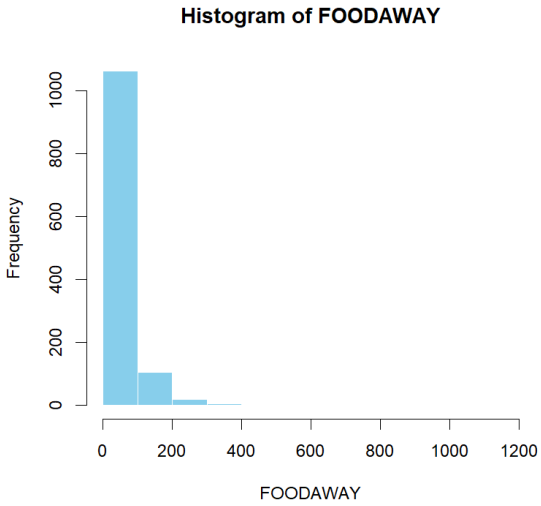
(g)

```
[1] "linear regression model SSE: 5262846.94710885"
> print(paste("quadratic regression model SSE:", SSE_quad))
[1] "quadratic regression model SSE: 4222356.34932398"
>
> if (SSE_lm < SSE_quad) {
+   print("linear regression model is better")
+ } else {
+   print("quadratic regression model is better")
+ }
[1] "quadratic regression model is better"
```

SSE 越低代表擬合越好，估計越準確。二次項回歸的表現更好。

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?



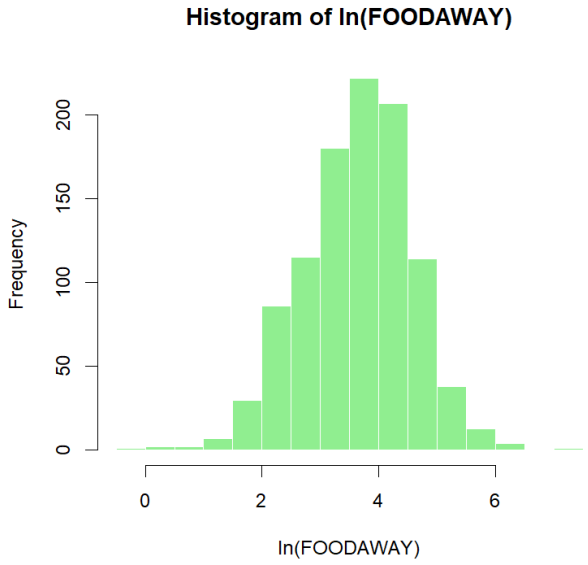

```
summary(cex5_small$foodaway)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   12.04   32.55   49.27   67.50 1179.00
```

```
> print(mean_foodaway)
[1] 49.27085
> print(median_foodaway)
[1] 32.555
> print(quantile_foodaway)
     25%      75%
12.0400 67.5025
```

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?

```
> mean_advanced
[1] 73.15494
> median_advanced
[1] 48.15
>
> mean_college
[1] 48.59718
> median_college
[1] 36.11
>
> mean_no_degree
[1] 39.01017
> median_no_degree
[1] 26.02
```

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




```
> summary(cex5_small$ln_foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.3011  3.0759  3.6865  3.6508  4.2797  7.0724

> mean_ln_foodaway
[1] 3.650804
> median_ln_foodaway
[1] 3.686499
> quantile_ln_foodaway
     25%      75%
3.075929 4.279717
```

有些數值為 0 不能取 log，所以在去除為 0 的數值後，資料筆數會減少。

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).

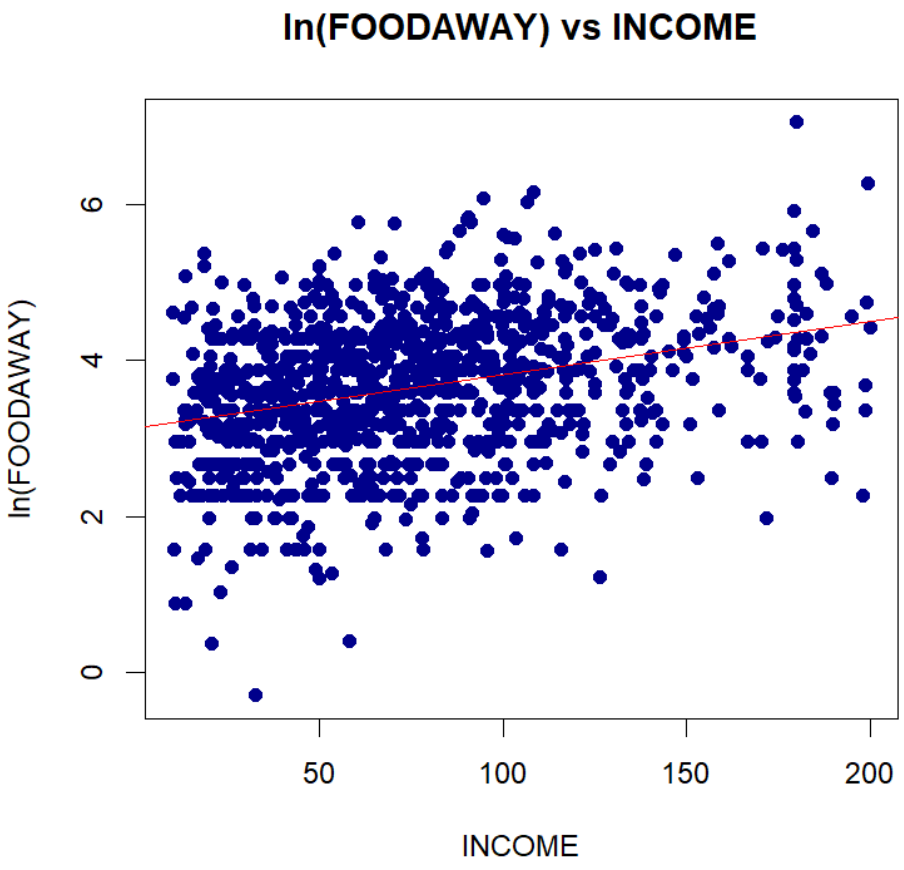
f. Calculate the least squares residuals from the estimation in part (d). Plot them vs. *INCOME*. Do you find any unusual patterns, or do they seem completely random?

(d)

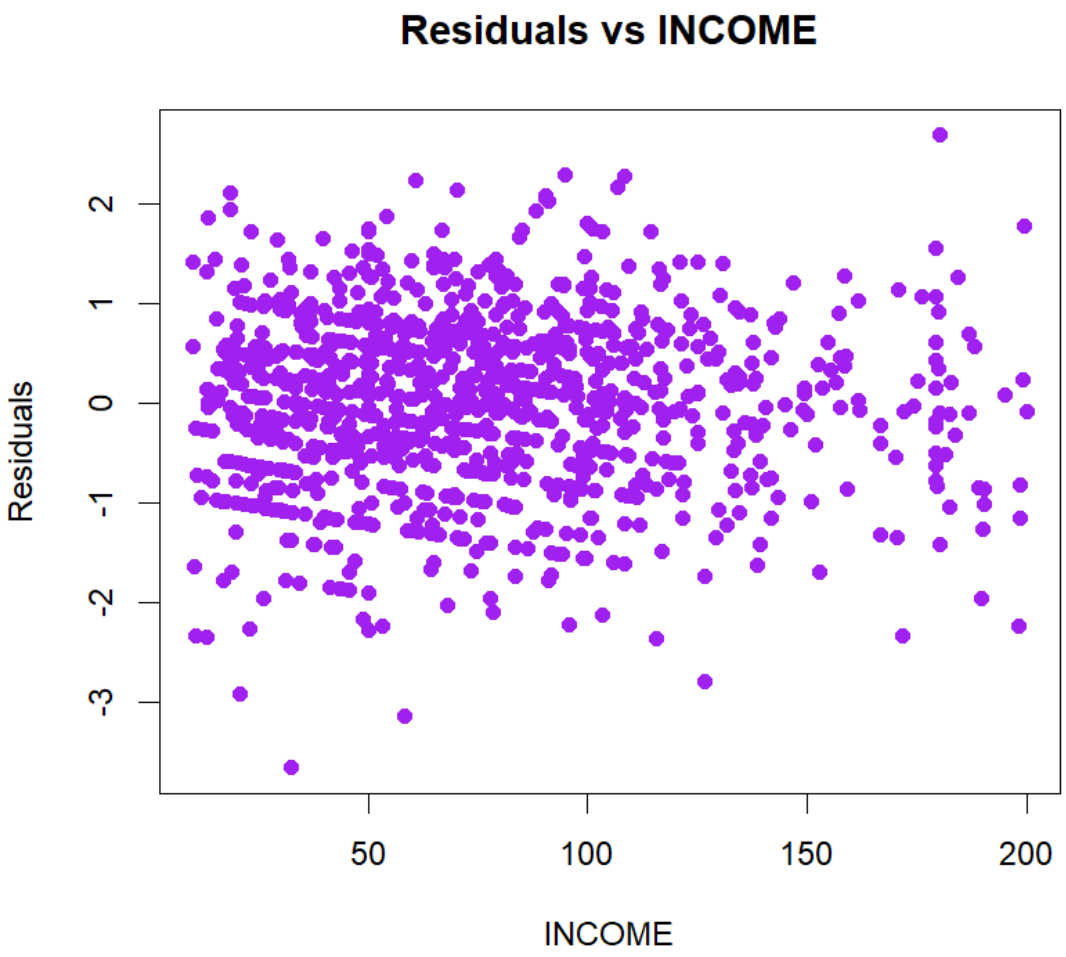
```
Call:
lm(formula = ln_foodaway ~ income, data = cex5_small)

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
```

Income 每增加 1 單位，會對 food away 造成 0.69%的影響。

(e)



(f)

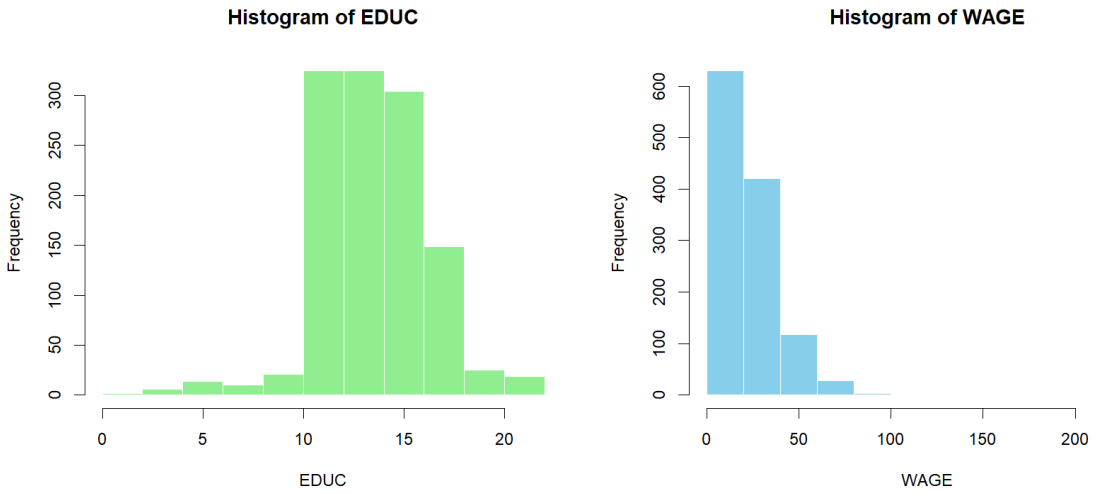


They seem completely random.

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.

```
[1] "WAGE"
> print(summary(cps5_small$wage))
   Min. 1st Qu.  Median    Mean
   3.94   13.00   19.30   23.64
3rd Qu.    Max.
  29.80  221.10
> print('EDUC')
[1] "EDUC"
> print(summary(cps5_small$educ))
   Min. 1st Qu.  Median    Mean
    0.0    12.0    14.0    14.2
3rd Qu.    Max.
   16.0    21.0
```



EDUC 表現出在受教育年分於 10~16 年間的人數最多

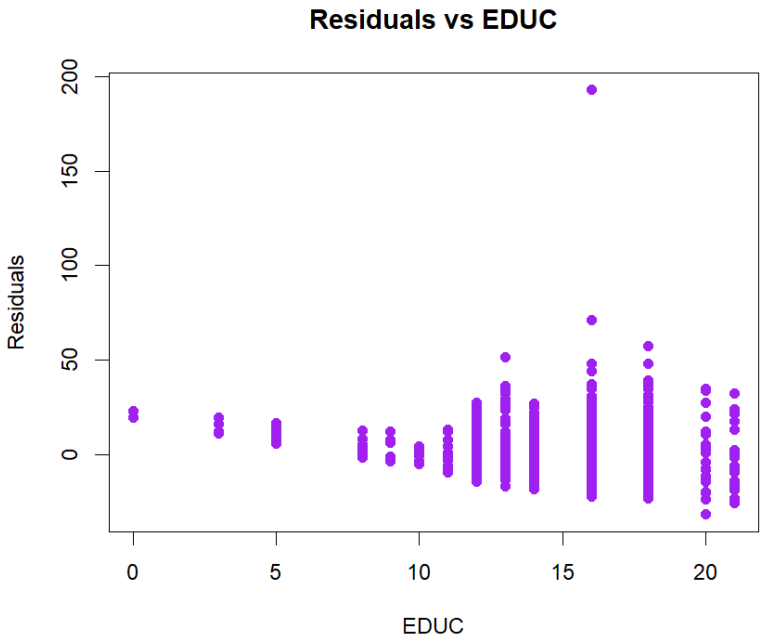
WAGE 呈現明顯右偏的分布，表示大部分的人工資偏低

b. Estimate the linear regression $WAGE = \beta_1 + \beta_2 EDUC + e$ and discuss the results.

```
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
```

Educ 的係數為 2.3968，在其他條件不變下，educ 每增加一年，時薪增加 2.3968 單位。

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?



殘差隨 educ 增加擴大。不滿足 SR5: Conditional Homoskedasticity

若 SR1~SR5 成立，殘差應該呈現隨機均勻分布的樣式。不應該有殘差隨 educ 增加而擴大的趨勢出現。

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

```
> summary(model_male)

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


> summary(model_black)

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

性別比較：女性的 educ 的係數較高，表示教育對薪資的影響較大。截距項較男性小，起薪可能較低。

種族比較：黑人的 educ 係數較低，教育對薪資的影響較小。截距項也較白人低。

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).

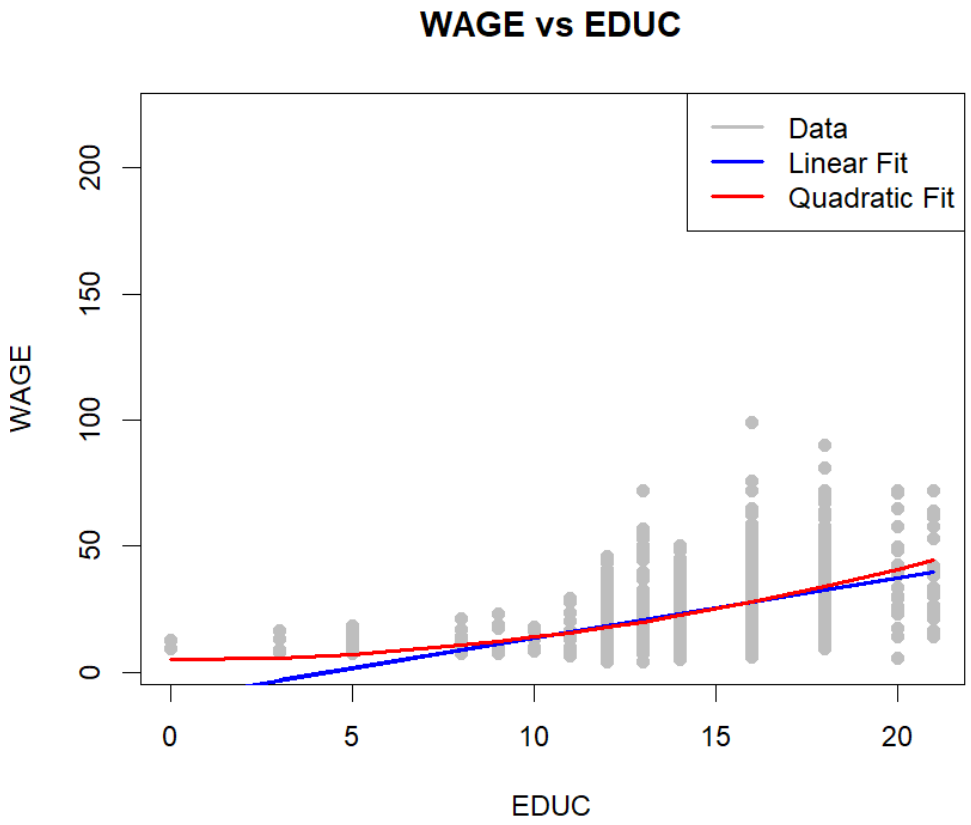
```
Call:
lm(formula = wage ~ I(educ^2), data = cps5_small)

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
```

二次項的係數>0 表示邊際效果隨 educ 增加會遞增。

```
> marginal_effect_12 <- 2 * coef(model_quadratic)["I(educ^2)"] * 12
> marginal_effect_16 <- 2 * coef(model_quadratic)["I(educ^2)"] * 16
> print('when educ = 12, ME=')
[1] "when educ = 12, ME="
> print(marginal_effect_12)
I(educ^2)
 2.139216
> print('when educ = 16, ME=')
[1] "when educ = 16, ME="
> print(marginal_effect_16)
I(educ^2)
 2.852288
```

與(b)的線性回歸模型相比，二次項回歸表示 educ 對 wage 的效果為遞增

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?



二次項回歸表現更好，除了較貼近數據的趨勢，在 educ 較低的地方，也避免 wage < 0 的結果出現。