POEHW0303—313707015 陳麗靜

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.

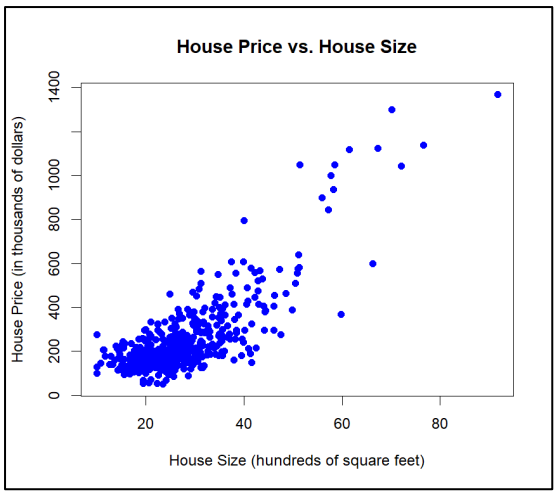
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

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)

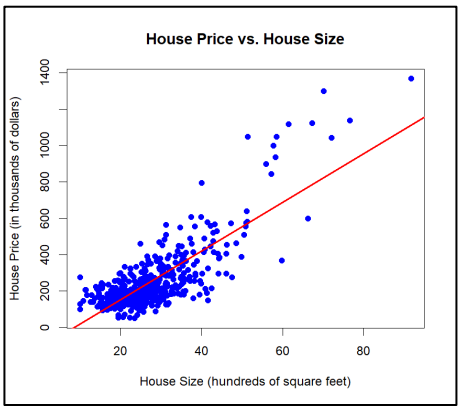
$$PRICE = \beta_1 + \beta_2 \cdot SQFT + e$$

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

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

其他條件不變下，居住面積每增加 100 平方英尺，預期房價將增加 13,402.94 美元。當面積為零時，預期價格為-115,423.60 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)

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

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

$(SE) \quad (6.072226) \quad (0.005256)$

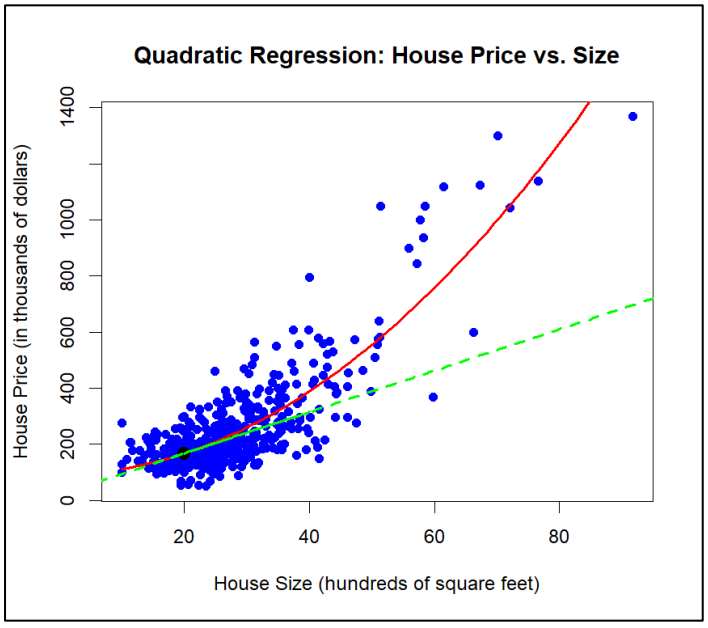
其他條件不變下，當居住面積達 2000 平方英尺時，每額外增加 100 平方英尺的居住面積將使預期房價增加 7,380.80 美元。

```
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
```

```
[1] "當房屋面積為 2000 平方英尺時，額外 100 平方英尺對房價的影響為： 7.3808"
```

(d)

紅色線為(c)小題的二次回歸曲線，綠色線為面積 2000 平方英尺處的切線。



(e)

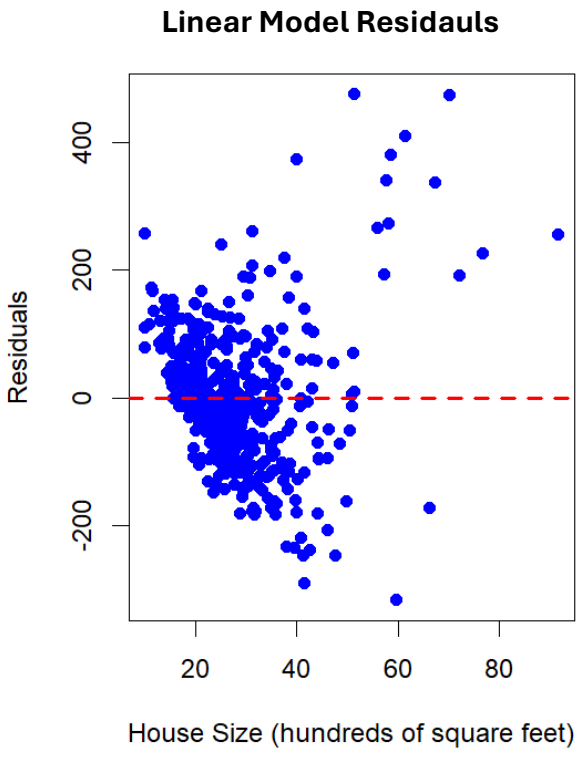
根據(c)小題的回歸模型 $PRICE = \alpha_1 + \alpha_2 \cdot SQFT^2 + e$，對 SQFT 微分得：$d(PRICE)/d(SQEF) = 2\alpha_2 \cdot SQFT$。再帶入彈性公式$E = (2\alpha_2 \cdot SQFT) \times SQFT/PRICE$。

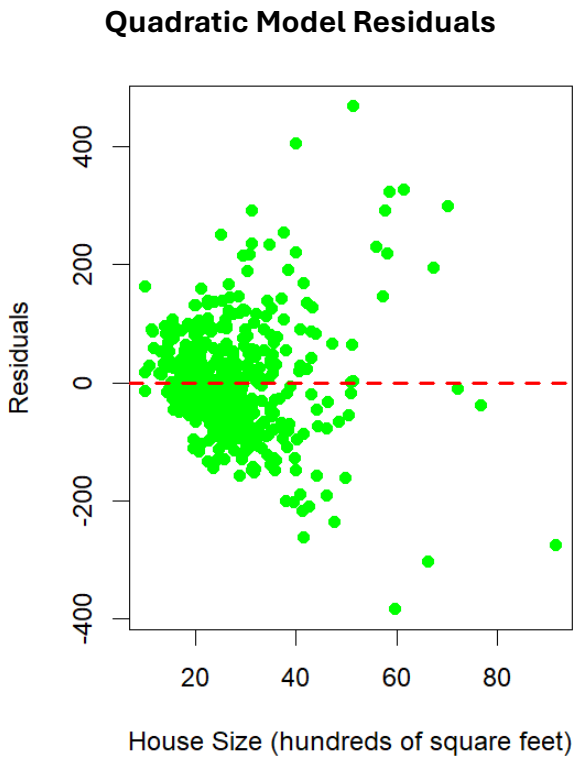
故當 SQRT=20 (2000 平方英尺，單位為 100 平方英尺)，對應的預期價格為 167.3735(單位千美元)，可得預期價格對房屋面積的彈性($\hat{e}$) = 0.882。

```
[1] "當房屋面積為 2000 平方英尺時，價格對面積的彈性為： 0.882"
```

(f)

在線性回歸模型和二次回歸模型中，殘差的分佈都沒有完全隨機，而是呈現某種系統性的趨勢，代表可能違反同質變異數的假設。





(g)

線性回歸模型的殘差平方和(SSE)為 5,262,846.9，而二次回歸模型的殘差平方和為 4,222,356.3。二次回歸模型的 SSE 更低。若 SSE 越低，表示觀測點與回歸線的距離較小，故二次回歸模型的預測值比線性回歸模型更接近實際數據。

```
> # 輸出結果
> print(paste("線性回歸模型的 SSE:", round(SSE_linear, 4)))
[1] "線性回歸模型的 SSE: 5262846.9471"
> print(paste("二次回歸模型的 SSE:", round(SSE_quadratic, 4)))
[1] "二次回歸模型的 SSE: 4222356.3493"
```

（2.25 下一頁）

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.

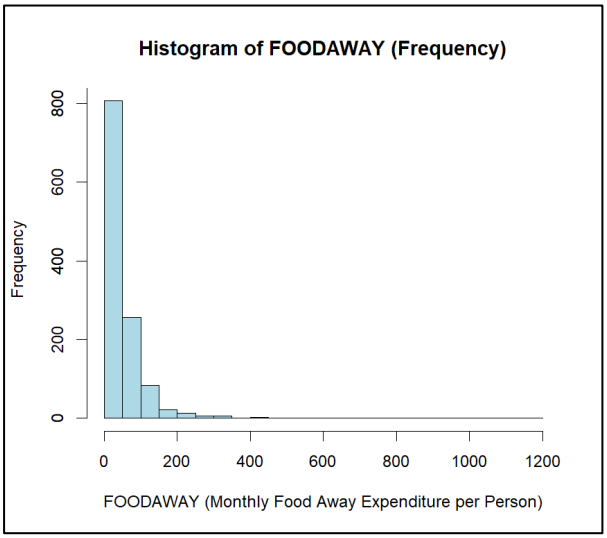
e. Plot ln(*FOODAWAY*) against *INCOME*, and include the fitted line from part (d).

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 of FOODAWAY	49.2709
Median of FOODAWAY	32.5550
25th percentile of FOODAWAY	12.0400
75th percentile of FOODAWAY	67.5075

```
> cat("Mean of FOODAWAY:", mean_foodaway, "\n")
Mean of FOODAWAY: 49.27085
> cat("Median of FOODAWAY:", median_foodaway, "\n")
Median of FOODAWAY: 32.555
> cat("25th Percentile of FOODAWAY:", percentile_25, "\n")
25th Percentile of FOODAWAY: 12.04
> cat("75th Percentile of FOODAWAY:", percentile_75, "\n")
75th Percentile of FOODAWAY: 67.5025
```



(b)

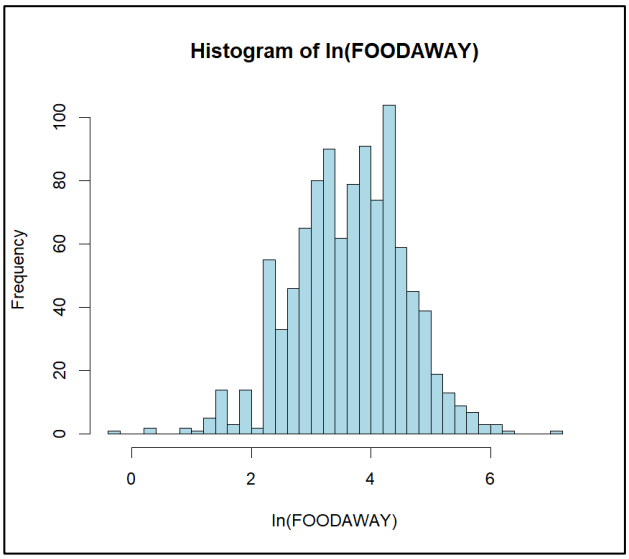
	Numbers	Mean	Median
ADVANCED = 1	257	73.1549	48.15
COLLEGE = 1	369	48.5972	36.11
NONE	574	39.0102	26.02

```
> # 顯示結果
> cat("FOODAWAY Statistics by Education Level:\n")
FOODAWAY Statistics by Education Level:
> cat("1. Households with an Advanced Degree (N =", n_advanced, "):\n")
1. Households with an Advanced Degree (N = 257 ):
> cat("   Mean:", mean_advanced, " Median:", median_advanced, "\n")
   Mean: 73.15494  Median: 48.15
>
> cat("2. Households with a College Degree (N =", n_college, "):\n")
2. Households with a College Degree (N = 369 ):
> cat("   Mean:", mean_college, " Median:", median_college, "\n")
   Mean: 48.59718  Median: 36.11
>
> cat("3. Households with No Advanced or College Degree (N =", n_no_degree,
"):\n")
3. Households with No Advanced or College Degree (N = 574 ):
> cat("   Mean:", mean_no_degree, " Median:", median_no_degree, "\n")
   Mean: 39.01017  Median: 26.02
```

(c)

ln(FOODAWAY)的數值比原本 FOODAWAY 少了 178 個，因為有 178 戶家庭回報每人外食支出為$0。而 ln(0)是沒有意義的，導致回歸分析無法使用這些觀測值。

Mean of ln(FOODAWAY)	3.6508
Median of ln(FOODAWAY)	3.6865
25th percentile of ln(FOODAWAY)	3.0759
75th percentile of ln(FOODAWAY)	4.2797
Maximum of ln(FOODAWAY)	7.0724
Minimum of ln(FOODAWAY)	-0.3011




```
> # 顯示統計摘要
> 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
```

(d)

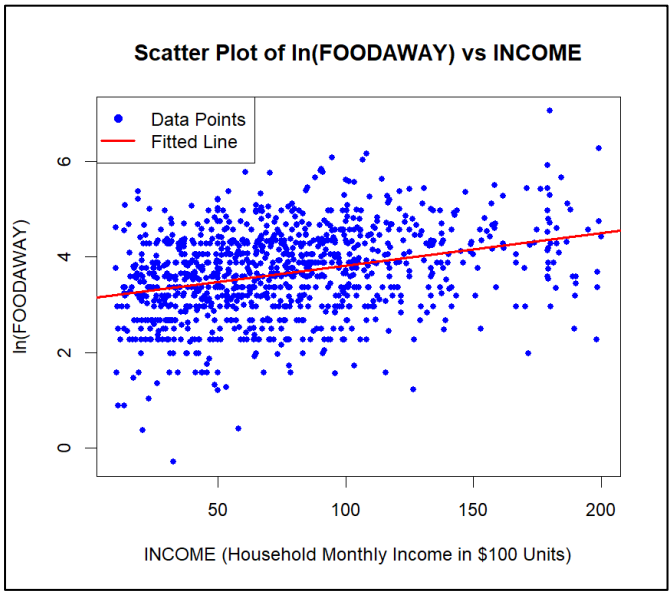
$$\ln(\widehat{FOODAWAY}) = 3.1293 + 0.0069 \cdot INCOME$$
$$(SE) \qquad (0.05655)\ (0.0007)$$

其他條件不變下，當 INCOME 增加 1(即收入增加 $100)時，ln(FOODAWAY) 平均增加 0.0069。若轉換回 FOODAWAY：$e^{0.0069017} - 1 \approx 0.00693 = 0.693\%$。即 INCOME 增加$100 時，外食支出會增加約 0.69%。

```
> # 顯示回歸結果
> summary(reg_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
```

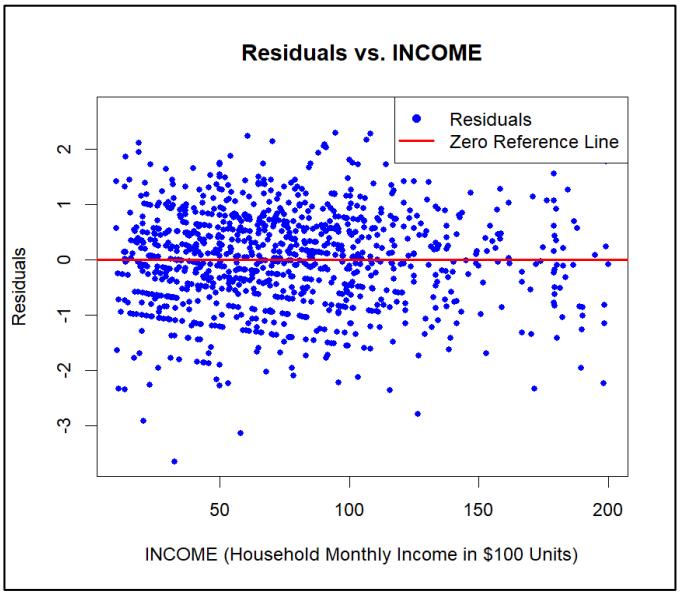
(e)

ln(FOODAWAY)和 INCOMES 大致呈正向關係。



(f)

殘差分佈大致隨機，且在紅線(殘差等於 0)上下，沒有明顯的特定趨勢，符合同質性變異數的假設。



(2.28 下一頁)

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.

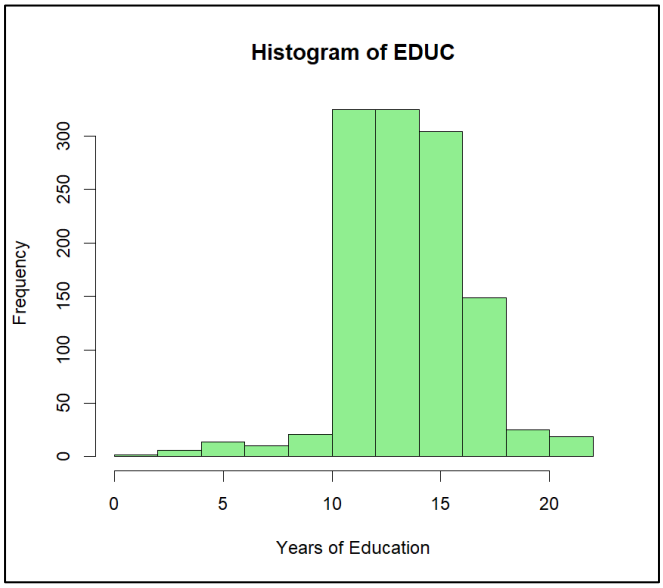
e. Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 16 years of education. Compare these values to the estimated marginal effect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

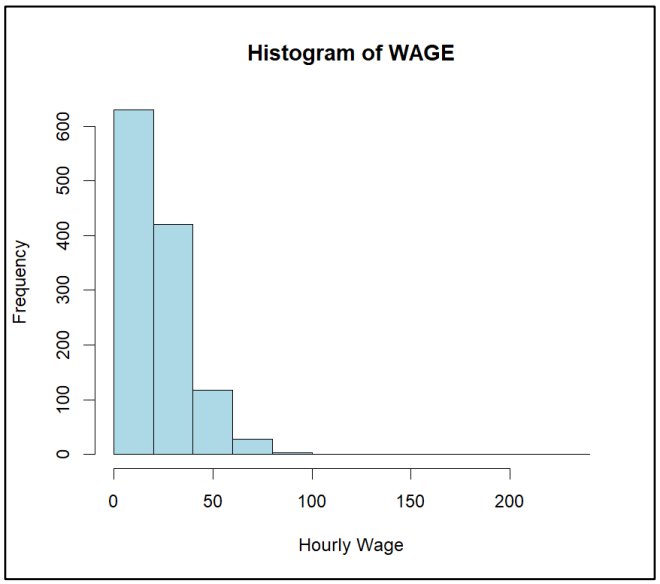
f. Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *WAGE* and *EDUC*. Which model appears to fit the data better?

(a)

WAGE 圖為右尾且最大值和最小值差距很大；EDUC 則呈些微左尾，表示僅有少部分人完全沒有或未完成義務教育。

```
# 取得 WAGE 和 EDUC 的摘要統計
summary(cps5_small$wage)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.94   13.00   19.30   23.64   29.80  221.10
summary(cps5_small$educ)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.0    12.0    14.0    14.2    16.0    21.0
```





(b)

$\widehat{WAGE} = -10.4000 + 2.3968 \cdot EDUC$

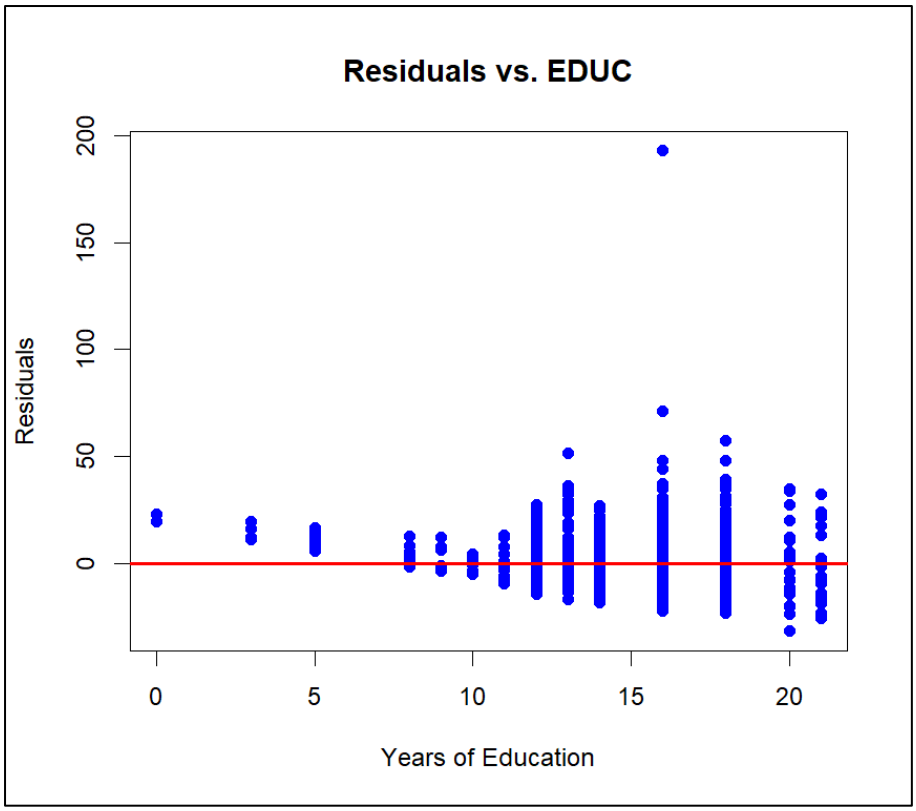
(SE) (1.9624) (0.1354)

其他條件不變下，每多增加一年受教育年分，WAGE 將會增加 2.3968 單位。

```
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
```

(c)

隨著教育年份增加，殘差的分佈越來越大，不符合同質性變異數的假設。



(d)

比較男性和女性的回歸模型，可以發現截距為負，不太符合現實。若看教育年份帶給兩個族群的薪資受益程度，女性的正影響比男性更大。另外，男性的殘差最大值與最小值差距遠大於女性，或許可能有薪資發放有性別不平等的情況。

```
> # 顯示回歸結果
> 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
```

下兩圖比較白人和黑人族群的薪資回歸模型，可以發現隨著受教育程度越高，對於兩個群體來說都會為薪資帶來正向影響，而白人族群的影響程度又比黑人更大。

```
> 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
```

```
> 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
```

(e)

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

計算邊際影響：$d(EDUC)/d(WAGE) = 2 \times \alpha_2 \times EDUC$。當受教育 12 年，每多增加一年受教育年份，薪資平均增加 2.1392 單位；當受教育 16 年，每多增加一年受教育年份，薪資平均增加 2.852288 單位。與(b)小題簡單線性回歸不同，不論當前受多少年教育，每增加一年受教育年份，其薪資皆增加 2.3968 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
```

```
> # 顯示結果
> 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
```

(f)

二次回歸(紅線)比線性回歸(藍線)稍微更好地配飾該資料集，捕捉到教育對薪資的影響是非線性的，並顯示隨著教育程度越高，薪資增長有一定程度的加速效應。