313707049 俞懷蕥

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.

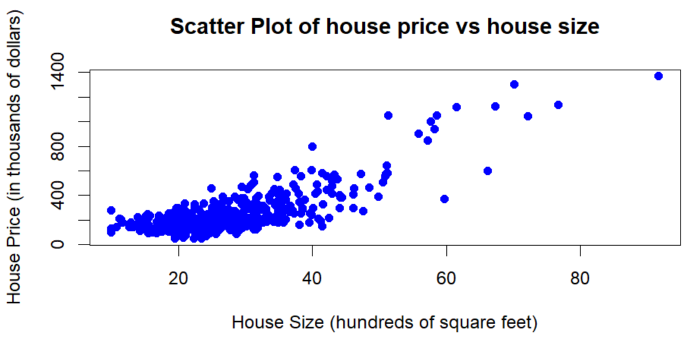
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

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)

$$PRICE = \beta_1 + \beta_2 \cdot SQFT + e$$
$$\widehat{PRICE} = -115.4236 + 13.4029 \cdot SQFT$$

其他條件不變，居住面積每增加一平方英尺，房價增加13.4029美元。當居住面積為零，房價為-115.4236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)

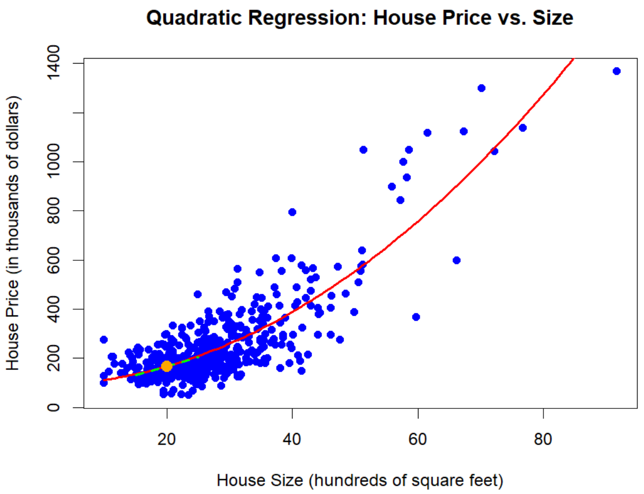
$$PRICE = \alpha_1 + \alpha_2 \cdot SQFT^2 + e$$
$$\widehat{PRICE} = 93.565854 + 0.184519 \cdot SQFT$$

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
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
其他條件不變，當居住面積達到2000平方英尺，居住面積每增加100平方英尺，房價增加7380.80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
```

(D)

紅色為C小題之二次曲線，淺綠為切線，橘色點為房屋面積2000平方英尺。

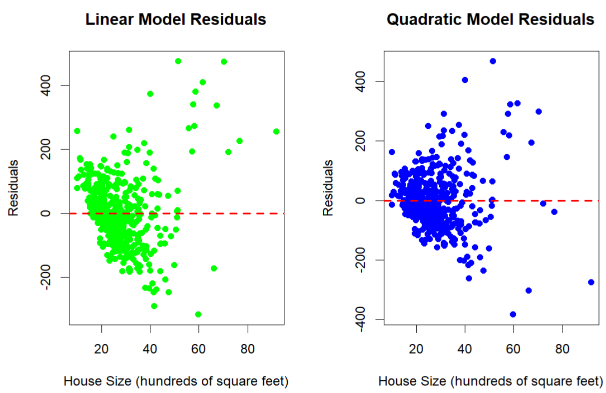


(E)

$d(PRICE)/d(SQEF) = 2\alpha_2 \cdot SQFT$。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$(\hat{e}) = 0.882$。

(F)

殘差成某種趨勢，違反同質變異數假設。



(g)

```
> # 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"
```

二次回歸模型SSE較小，代表觀測值與回歸線距離較小，更接近實際數據。

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.

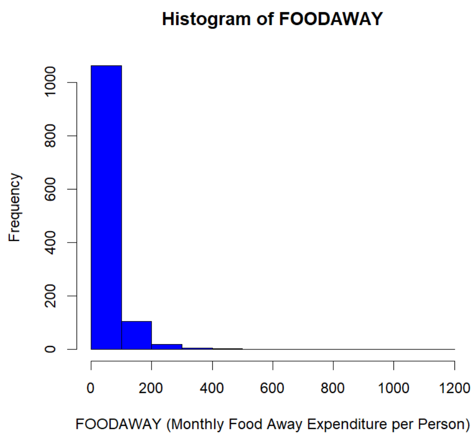
d. Estimate the linear regression $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

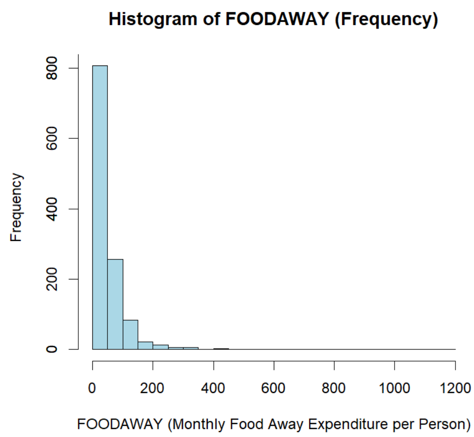
e. Plot ln(*FOODAWAY*) against *INCOME*, and include the fitted line from part (d).

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)

```
> # 輸出結果
> 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
```

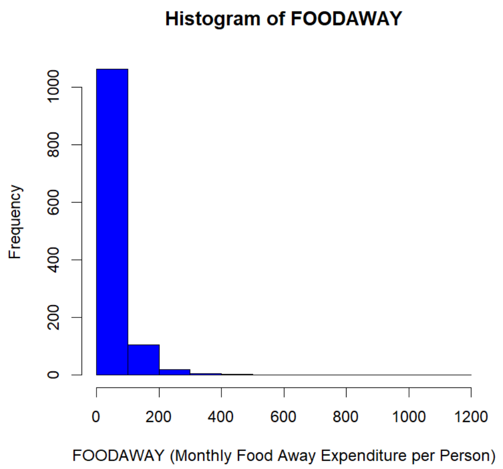


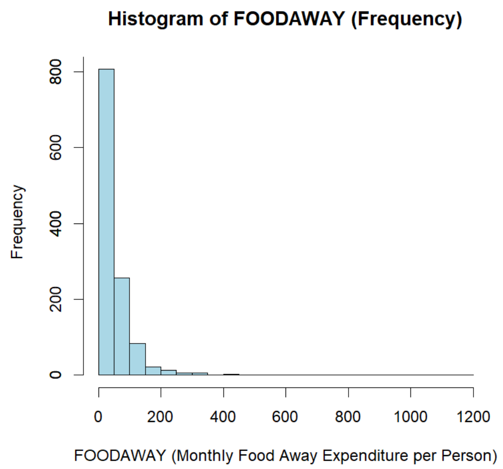


(右圖增加bins使圖形平滑)

(b)

```
> # 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
```

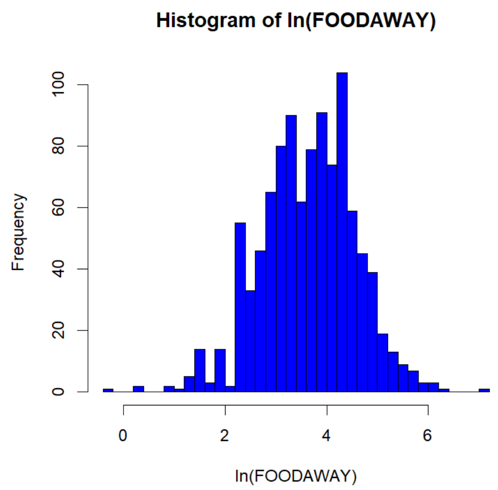




(右圖增加bins使圖形平滑)

(c)

ln(FOODAWAY)包含0，使得數值沒有意義。由結果可看出ln(FOODAWAY)的數值比FOODAWAY少178個。




```
> # 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
> cat("Number of Observations (FOODAWAY):", n_foodaway, "\n")
Number of Observations (FOODAWAY): 1200
```

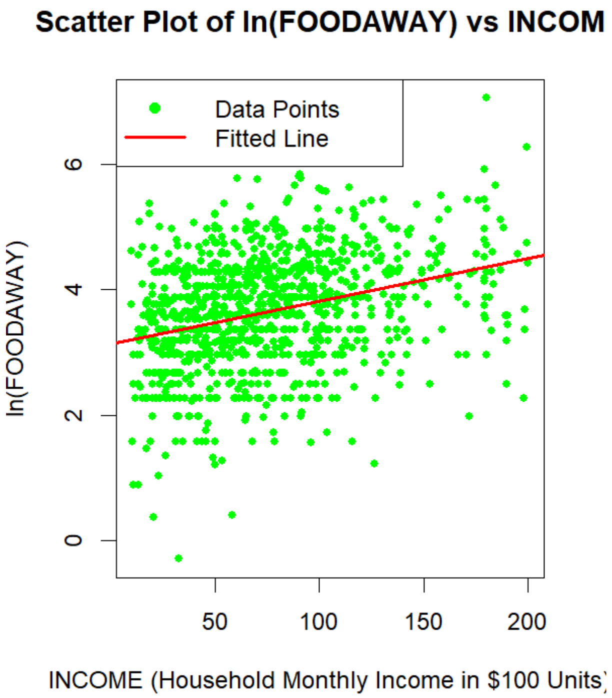
(d)

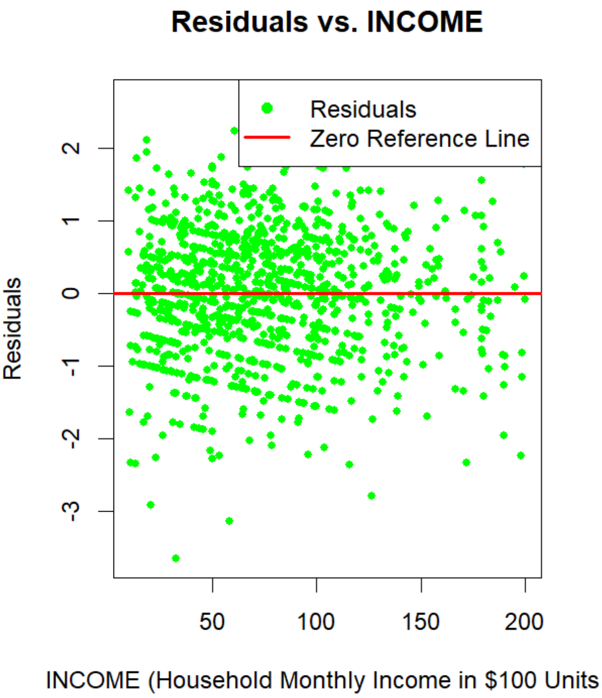
$$\ln(\widehat{FOODAWAY}) = 3.1293 + 0.0069 \cdot INCOME$$

其他條件不變之下，收入每增加一單位($100)，ln(FOODAWAY)平均增加0.0069單位。而e^0.0069-1 約為6.93%，即income增加$100，外食支出平均增加0.69%。

```
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)
左圖可看出大致呈正向關係
(f)右圖可看出殘差呈現隨機，無特定分布趨勢，符合同質變異數假設。





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).

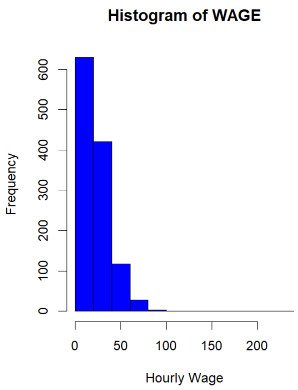
f. Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *WAGE* and *EDUC*. Which model appears to fit the data better?

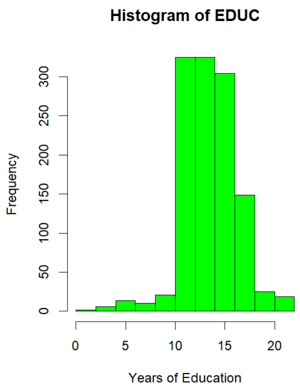
(a)

左圖:右偏態;右圖:左偏態

從右圖可知大部分的人教育年限長。

```
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)

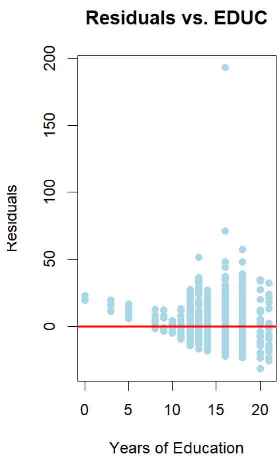
$$\widehat{WAGE} = -10.4000 + 2.3968 \cdot EDUC$$

平均每增加一單位教育年限，薪資增加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

(Intercept) ***
educ        ***
---
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)

隨著教育年限越多，殘差越大，不符合同質變異數假設。



(d)

比較男女:

兩子集的工資率對受教育年限都有顯著的正斜率,

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 。殘差的極大

質,男性是 191.328 ,遠高於女性的 49.502 ,雖然只是單筆數

據不足以代表,但可推測是工資率的發放有性別不平等的現象。

```
> 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
```

比較黑人和白人:兩者都顯著正斜率,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族群來說都會上升 。值得注意的是白人族群在這方面受惠程度較高,可能推測是因為種族歧視導致此項差異(即便黑人多受教育能帶來的邊際效益也較低)。又觀察兩者殘差的極大值也可以做出類似推論。

```
> 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)

```
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
```

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

```
> 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
```

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) quadratic model的殘差較小，更配適此資料集，也解決截距不可為負。