313707014 陳紀蓁

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.

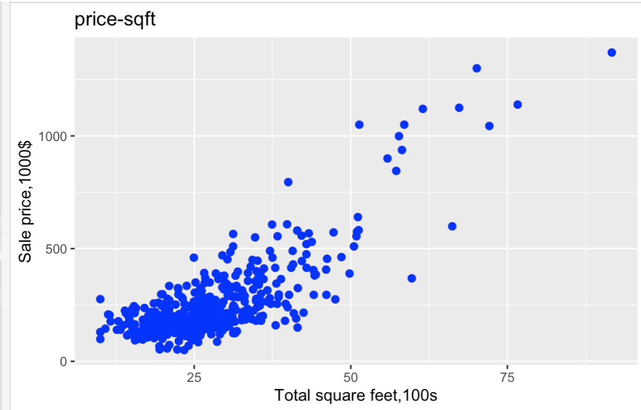
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

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.

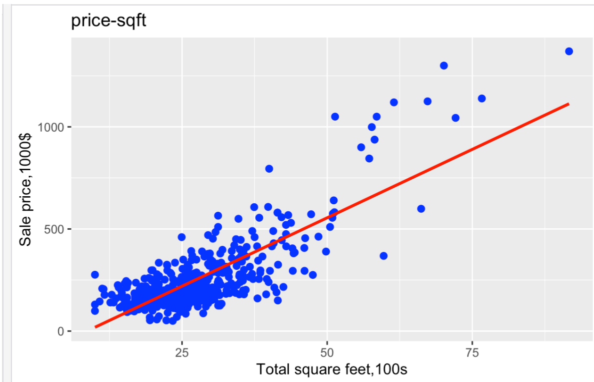


b. PRICE = −115.4236 + 13.4029 * SQFT

在其他條件不變的前提下，每增加100平方英尺，房屋的預期價格會增加13.4029 (1000美元)

截距為 −115.4236 ，代表當 SQFT = 0 時，房屋的預期價格為 −115.4236 (1000美元)

```
Call:
lm(formula = price ~ sqft)

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
```



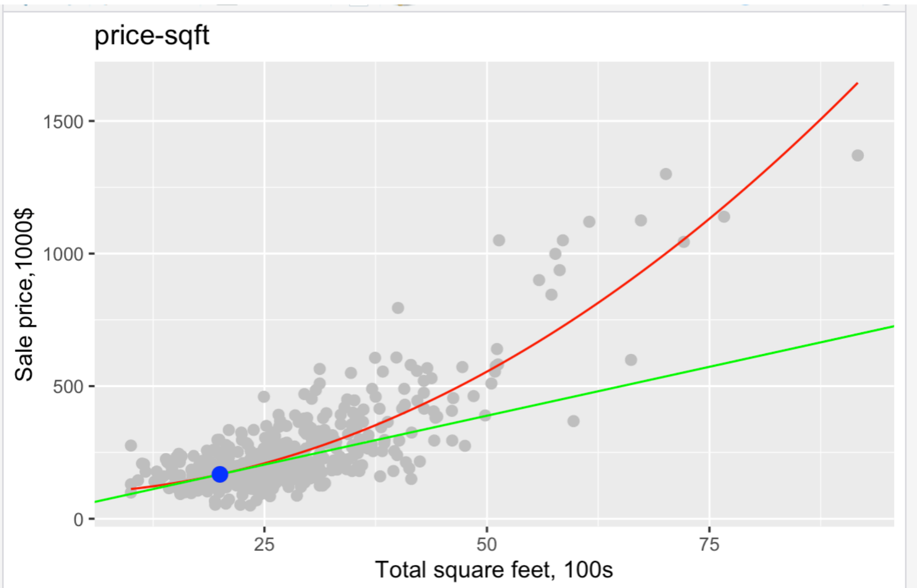
c. $PRICE = 93.565854 + 0.184519 * SQFT^2$

Margin effect : $2 * 0.184519 * SQFT$

SQFT = 20, margin effect : 7.38096 (1000美元)

在房屋面積2000平方英尺的前提下，每增加100平方英尺，房屋的預期價格會增加7.38096 (1000美元)

d.

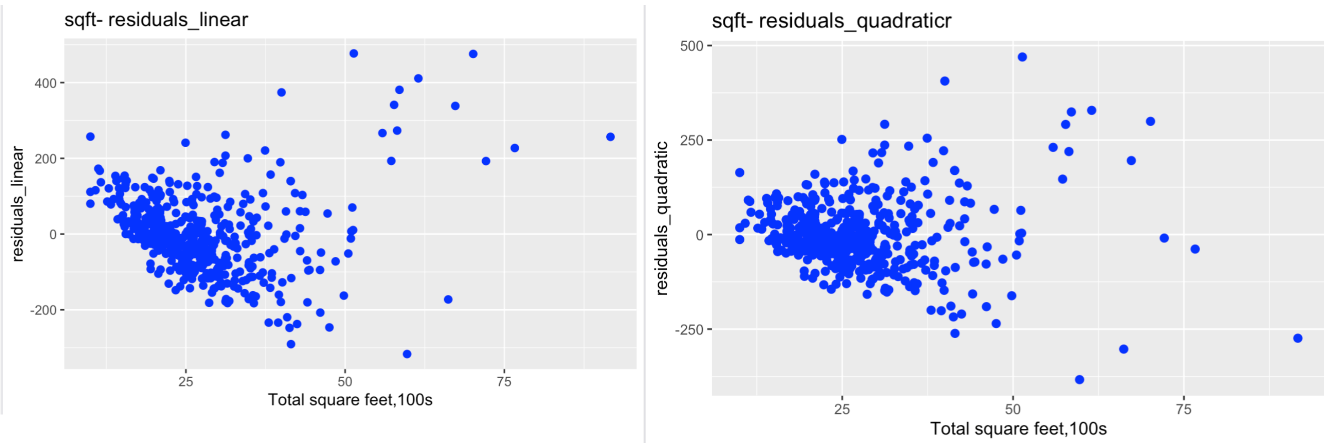


紅線：二次回歸線

綠線：當房屋面積在2000平方英尺的切線

e. elasticity = 0.8819511

f.



殘差在上面兩張圖中皆沒有呈現常態分佈，且隨著SQFT的增加而增加，因此違反了homoskedasticity 假設

g.
SSE (linear) : 5,262,847
SSE (quadratic) : 4,222,356

二次回歸的殘差平方和較低為4,222,356

較低的 SSE， 代表模型的預測值比較接近實際值，因此可以說二次模型相較於線性模型擬合度較高

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?

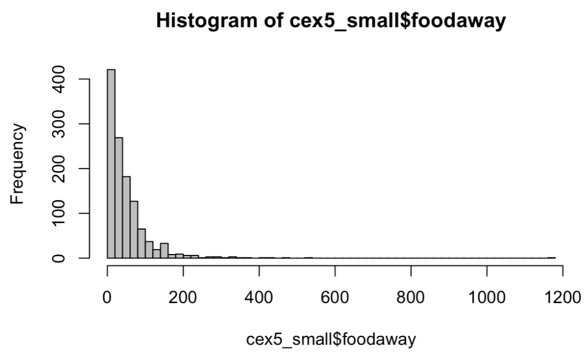
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

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: 49.27
Median 32.55
25th percentiles : 12.04
75th percentiles : 67.5
N = 1200

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

```
> describe(cex5_small$foodaway)
   vars    n  mean    sd median trimmed   mad min  max range skew kurtosis   se
X1    1 1200 49.27 65.28  32.56   38.35 33.99   0 1179  1179 6.25    81.48 1.88
```

b.

Advance :

N = 257

Mean = 73.15

Median = 48.15

```
> describe(filtered_data_ad$foodaway)
   vars   n  mean     sd median trimmed   mad min  max range skew kurtosis   se
X1    1 257 73.15 102.04  48.15    56.5 49.98   0 1179  1179 5.91    54.23 6.37
> 
```

College :

N = 369

Mean = 48.6

Median = 36.11

```
> describe(filtered_data_co$foodaway)
   vars   n mean    sd median trimmed   mad min    max  range skew kurtosis   se
X1    1 369 48.6 51.97  36.11   40.09 32.13   0 416.11 416.11 2.73    11.49 2.71
> 
```

None:

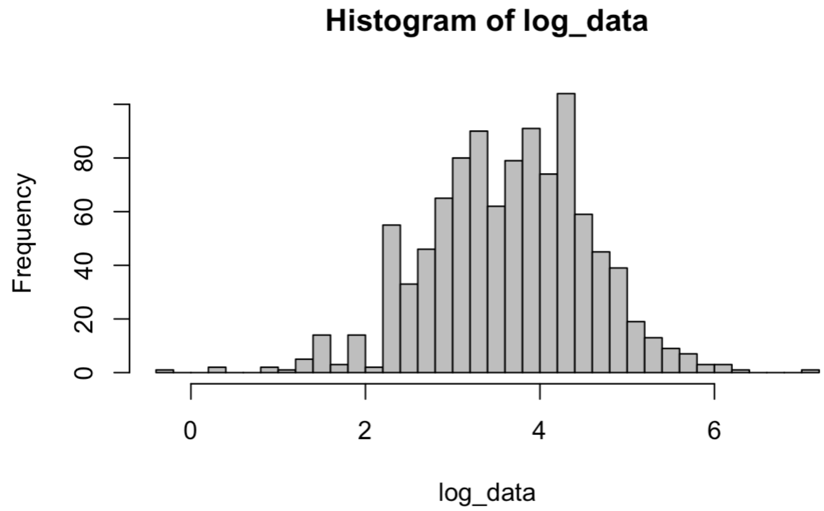
N = 574

Mean = 39.01

Median = 26.02

```
> describe(filtered_data_no$foodaway)
   vars   n  mean    sd median trimmed   mad min    max  range skew kurtosis   se
X1    1 574 39.01 46.58  26.02   30.94 32.81   0 437.78 437.78 3.06    15.95 1.94
> 
```

c.




```
> describe(log_data)
   vars    n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 1022 3.65 0.92   3.69    3.67 0.88 -0.3 7.07  7.37 -0.23     0.45 0.03
```

因為有178個家庭的外食花費為 0，當執行ln(foodaway) 時，會產生缺失值，因此 ln(foodaway) 的值相較於 foodaway 少了178個

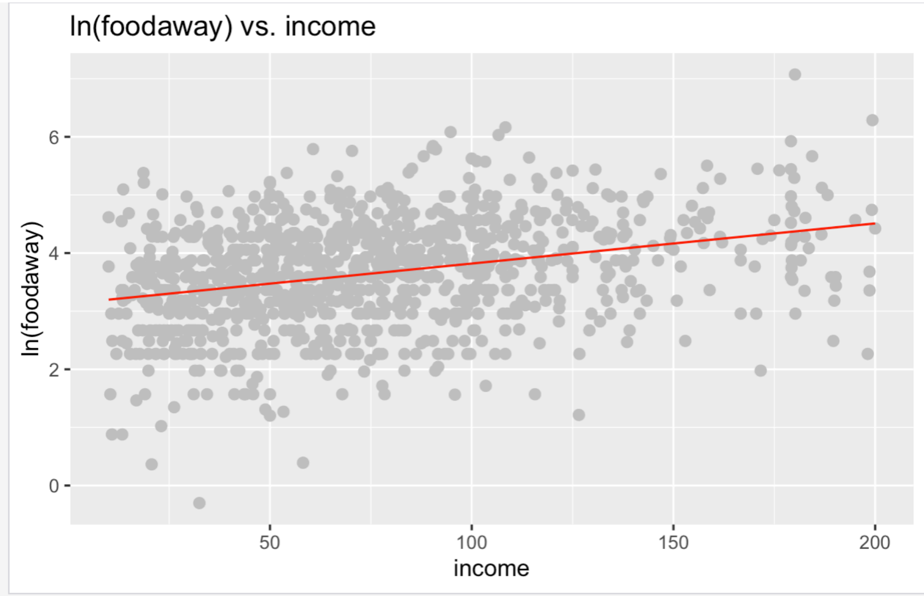
d. $ln(foodaway) = 3.1293004 + 0.0069017 * INCOME$

Slope : 0.0069017

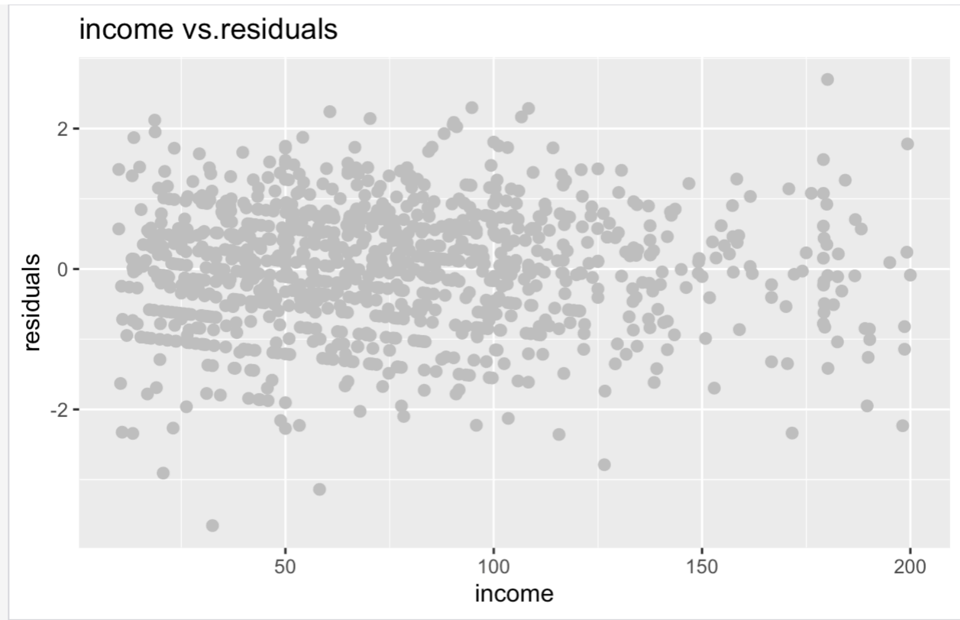
在其他條件不變的前提下，當 income 增加 100 units, 外食花費（per person）會增加 0.0069017

```
Call:
lm(formula = log_data ~ income_data, data = comb_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_data 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. 圖為：ln(foodaway) 和 income 的關係



f. 從圖中可以看到沒有特別的分佈狀態，應該可以推論為隨機分佈



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).

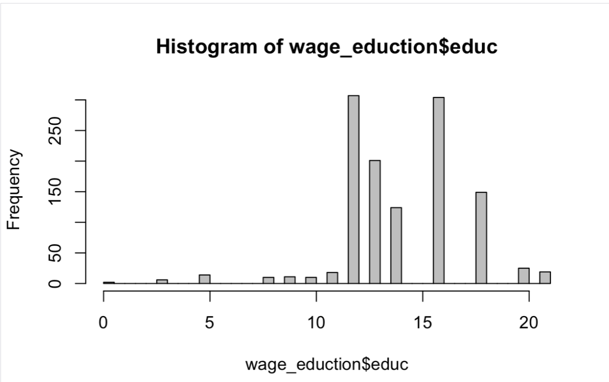
f. Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *WAGE* and *EDUC*. Which model appears to fit the data better?

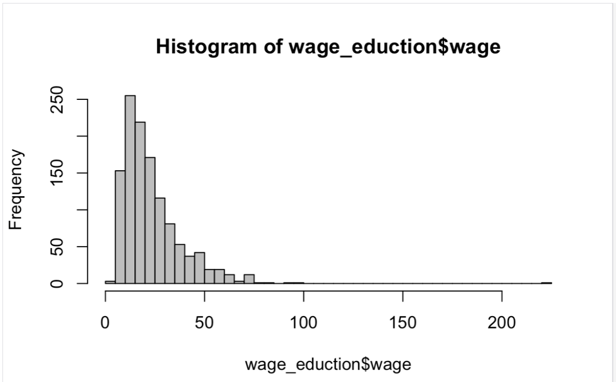
a.

summary statistics of WAGE and EDUC

```
> summary(wage_eduction$wage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3.94   13.00   19.30   23.64   29.80  221.10 
> summary(wage_eduction$educ)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0    12.0    14.0    14.2    16.0    21.0 
```

histograms for the variables EDUC and WAGE.





EDUC：多落在12~16年間，分佈較平均

WAGE:整體呈現右偏，代表可能有少數極端高薪的可能

b. 回歸線預測

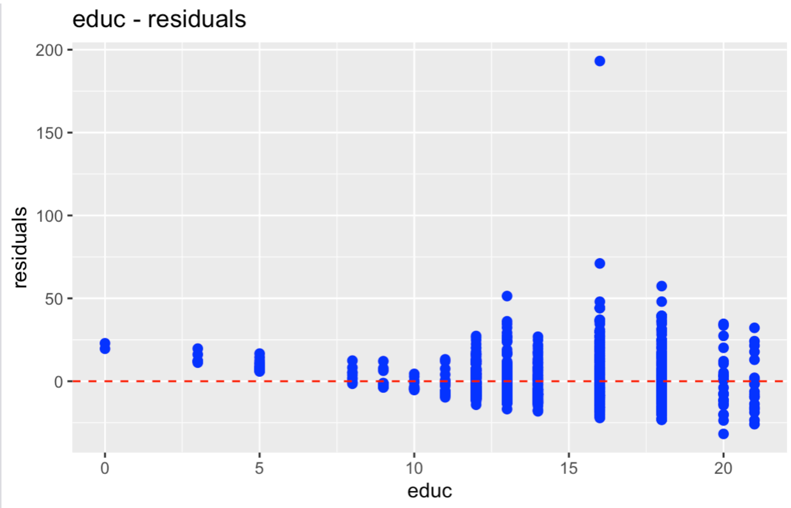
```
Call:
lm(formula = wage ~ educ, data = wage_eduction)

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
```

得出結果為：

$$WAGE = -10.4 + 2.3968EDUC$$

截距為 –10.4 代表當教育年份為 0 時的預期薪資，但是教育年份不可能為 0
斜率為 2.3968 代表當教育年份每增加一，預期薪資會增加 2.3968

c. least squares residuals
Sum of residuals : 7.642775e–13



殘差隨著教育年份增加，變異增大
違反同質變異 (homoskedasticity)

若SR1–SR5成立，理想的殘差圖應：近似隨機散佈在 0 上下，不隨 EDUC 系統性地變大或變小

d.

```
$blacks

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

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
```

```
$whites

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

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
```

```
$male

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

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
```

```
$female

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

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
```

黑人： WAGE = −6.2541 + 1.9233 * EDUC
白人： WAGE = −10.475 + 2.418 * EDUC
男性： WAGE = −8.2849 + 2.3785 * EDUC
女性： WAGE = −16.6028 + 2.6595 * EDUC

不同種族間的比較（黑人vs.白人）
當EDUC 為 0 時，白人的薪資較低，但是隨著教育年數增加，白人的平均工資會高過於黑人，因為白人的教育係數較高

不同性別間的比較 (男性 vs. 女性)
當 EDUC 為 0 時，女性的薪資較低，但是隨著教育年數增加，可以發現女性的增長幅度是可以超過男性的，因為女性的教育係數較高

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
```

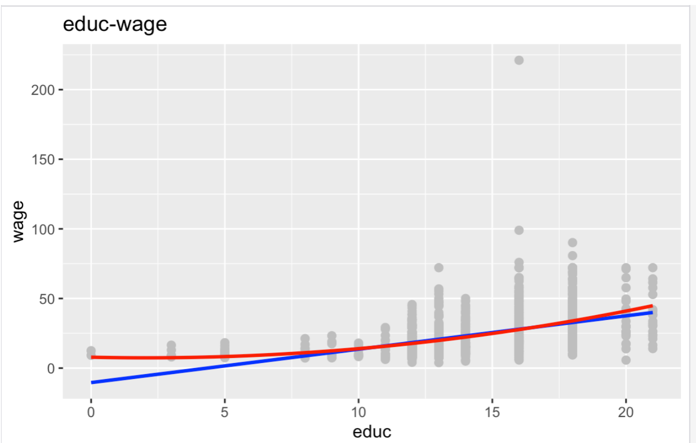
$WAGE = 4.916477 + 0.089134 * EDUC^2$

Margin effect : $2 * 0.089134 * EDUC$

Year = 12的邊際效果：2.139216
Year = 16的邊際效果：2.852288

在（b.）小題的假設下，不管 Year 是 12 還是 16 邊際效果皆為 2.3968 ，代表線性模型的邊際報酬是固定的，而非線性模型的邊際報酬率隨著教育水準變化。可以發現 Year 16 的邊際效果大於 Year 12 的，代表在較高教育水準之下，每增加一年的教育可以產生更顯著的薪資增長效果。

f.



紅線（二次回歸）
藍線（線性回歸）

從圖中可以發現紅線更加地貼近資料分佈狀態