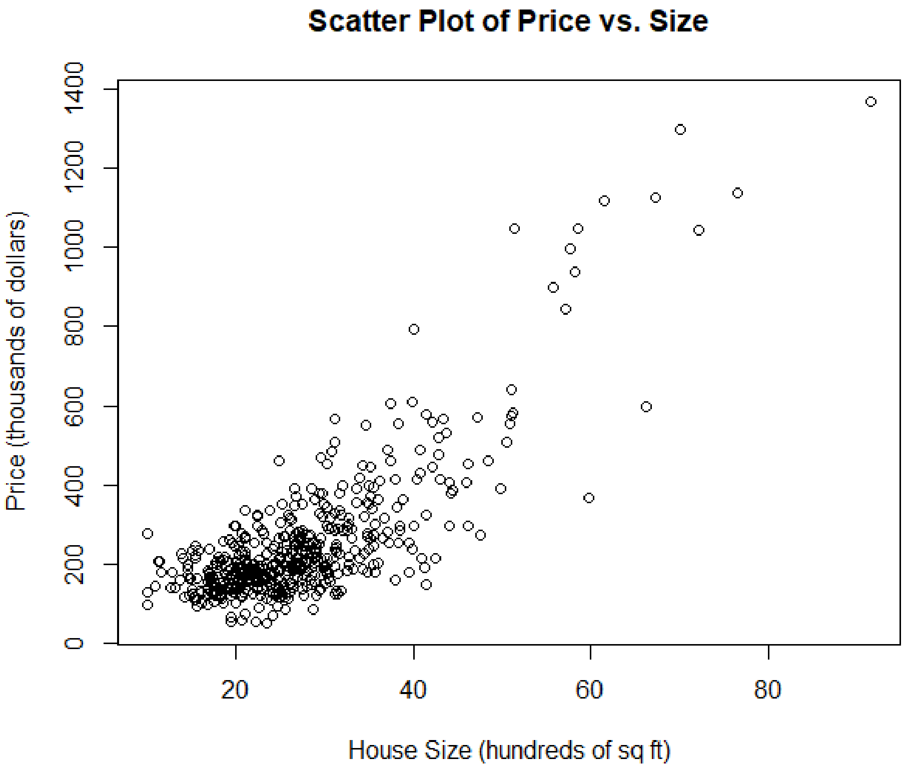
2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Plot house price against house size in a scatter diagram.

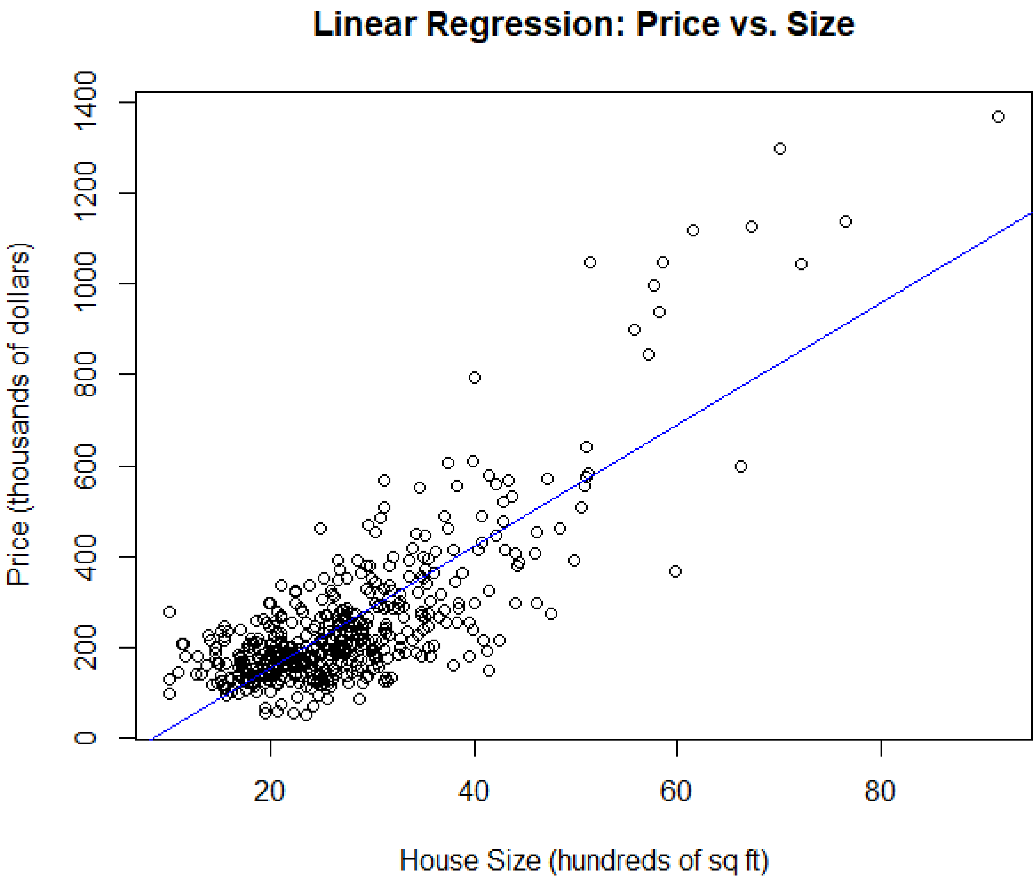


b. Estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.

```
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
```

$$Price = -115.4236 + 13.4029 * SQFT$$

➔若$SQFT$增加一單位，價格增加 13.4029，若$SQFT = 0$,價格為-115.4236



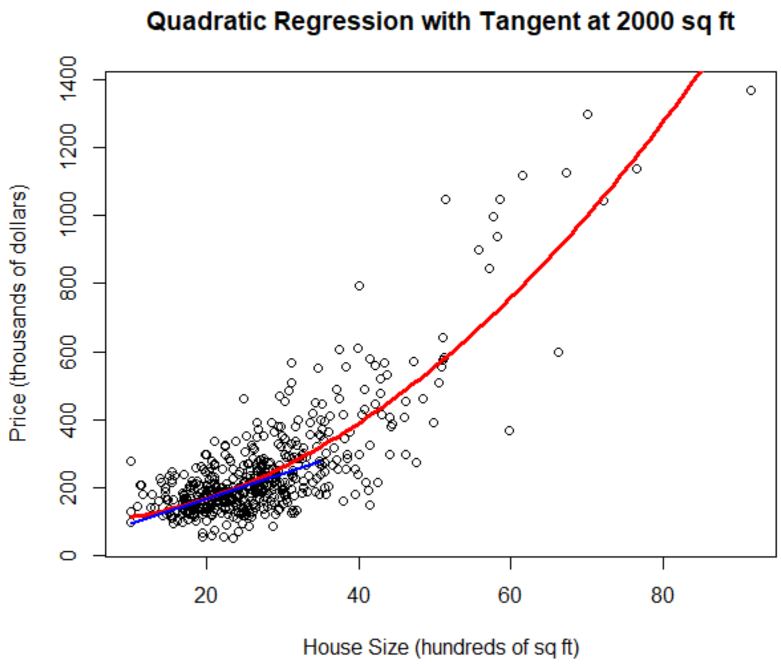
c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

```
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 ***
sqft2        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
```

$$Price = 93.565854 + 0.184519 * SQFT^2$$

$$Margin\ effect = \frac{dPrice}{dSQFT} = 2 * 0.184519 * 20 = 7.38$$

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.

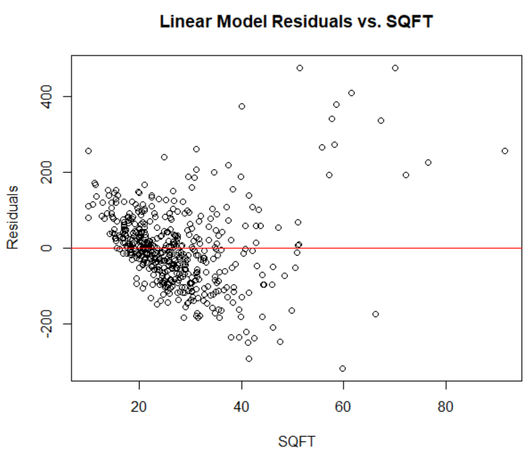


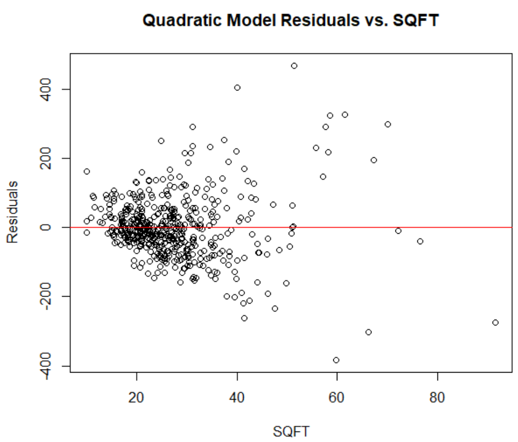
e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

```
> # At SQFT = 20
> price_at_20 <- alpha1 + alpha2 * 20^2
> elasticity <- (2 * alpha2 * 20) * (20 / price_at_20)
> cat("Elasticity at 2000 sqft:", elasticity, "\n")
Elasticity at 2000 sqft: 0.8819511
>
```

➔SQFT 上升 1%, 價格上升 0.8819511%

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?





➔殘差並非 random 分布，違反 assumptions

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?

```
> cat("SSE Linear:", SSE_linear, "\n")
SSE Linear: 5262847
> cat("SSE Quadratic:", SSE_quad, "\n")
SSE Quadratic: 4222356
```

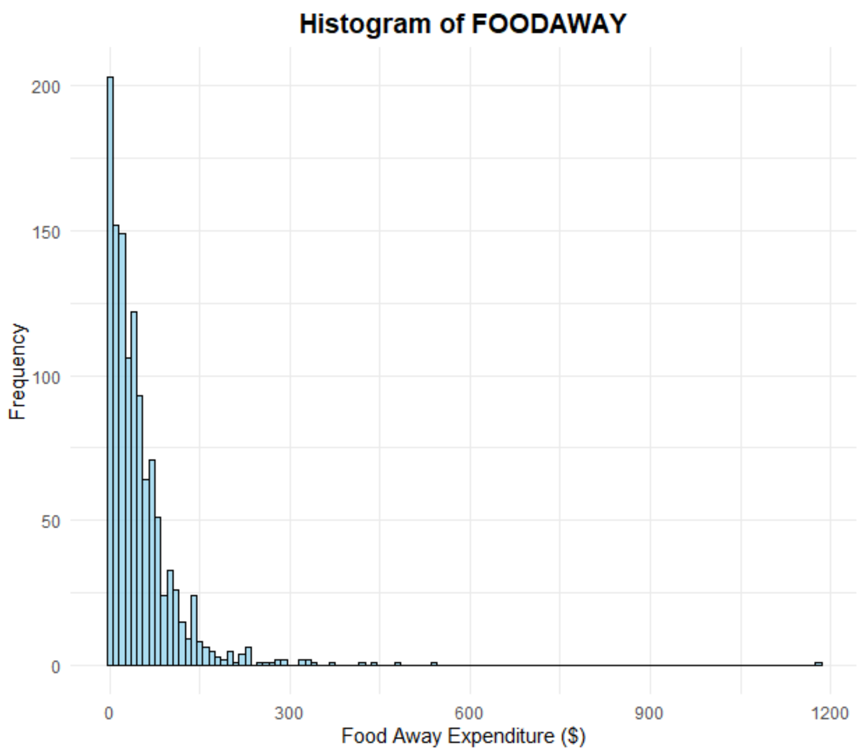
➔ Quadratic model has lower SSE➔better-fitting

更低的 SSE 表示預測值更接近真實值

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?

```
> cat("Mean FOODAWAY:", mean_foodaway, "\n")
Mean FOODAWAY: 49.27085
> cat("Median FOODAWAY:", median_foodaway, "\n")
Median FOODAWAY: 32.555
> cat("25th Percentile:", quantiles[1], "\n")
25th Percentile: 12.04
> cat("75th Percentile:", quantiles[2], "\n")
75th Percentile: 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?

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

b.

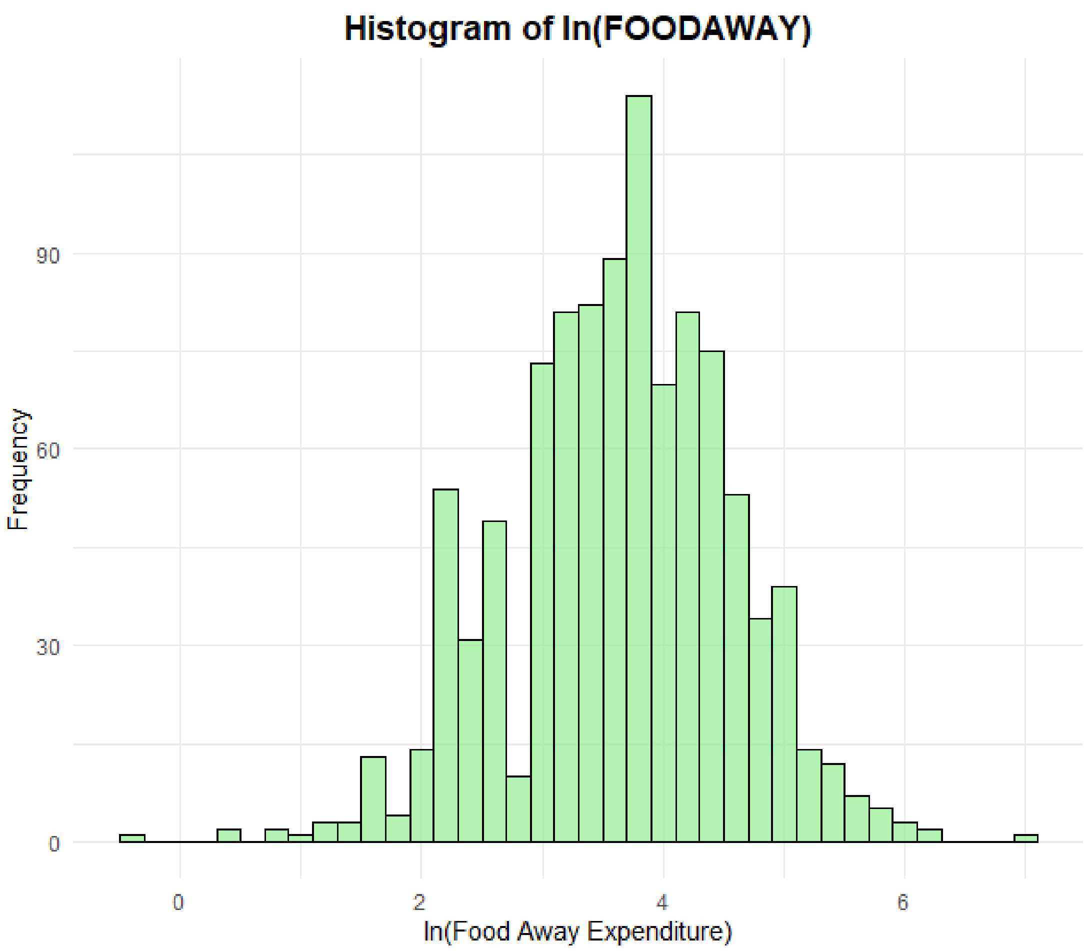
```
> cat("Advanced Degree - Mean:", mean_adv, "Median:", median_adv, "\n")
Advanced Degree - Mean: 73.15494 Median: 48.15
> cat("College Degree - Mean:", mean_college, "Median:", median_college, "\n")
College Degree - Mean: 48.59718 Median: 36.11
> cat("No Degree - Mean:", mean_none, "Median:", median_none, "\n")
No Degree - Mean: 39.01017 Median: 26.02
```

c.

因為 log(N)在 N≤0 時無值，故取樣時會排除在外，因此兩個樣本數不同

```
> cat("Mean ln(FOODAWAY):", mean_ln, "\n")
Mean ln(FOODAWAY): -Inf
> cat("Median ln(FOODAWAY):", median_ln, "\n")
Median ln(FOODAWAY): 3.482878
```

```
> cat("Number of observations in ln(FOODAWAY):", length(na.omit(cex5_small$ln_FOODAWAY)), "\n")
Number of observations in ln(FOODAWAY): 1200
```



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?

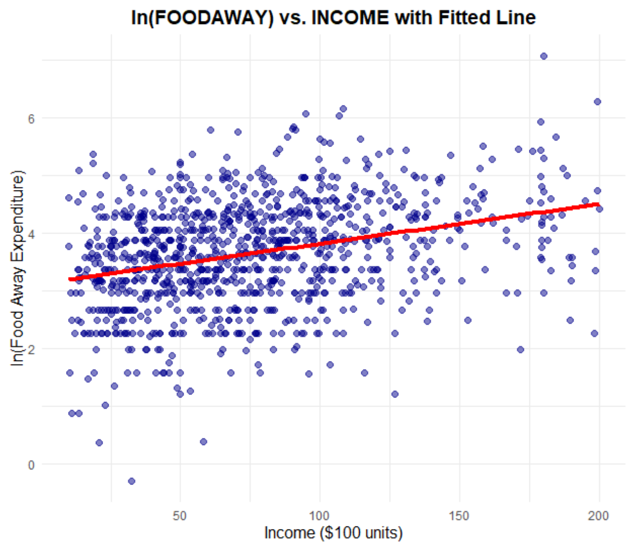
d.

```
> cat("Intercept (β1):", beta1, "\n")
Intercept (β1): 3.1293
> cat("Slope (β2):", beta2, "\n")
Slope (β2): 0.006901748
```

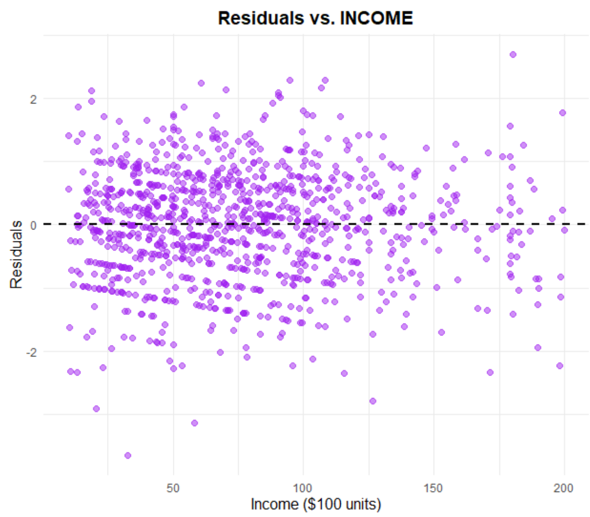
$$\ln(Foodaway) = 3.1293 + 0.0069017 * income$$

→come 上升 100 時，Foodaway 上升 0.69017%

e.



f.



→殘差沒趨勢為水平線→OLS 模型有效

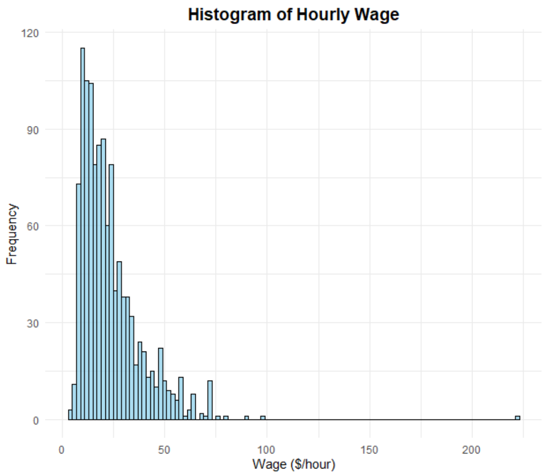
2.28 How much does education affect wage rates? The data file *cps5_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version.]

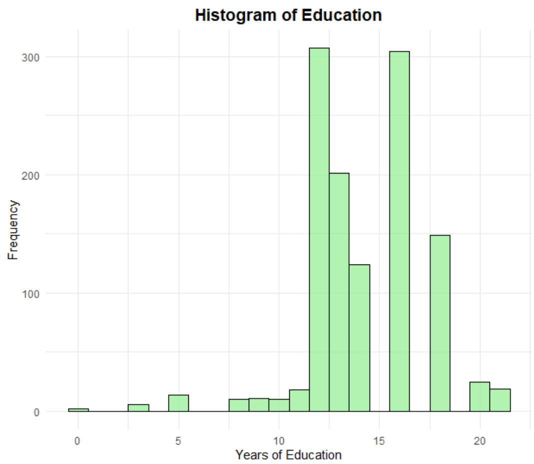
a. Obtain the summary statistics and histograms for the variables *WAGE* and *EDUC*. Discuss the data characteristics.

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

a.

```
> cat("Summary Statistics for WAGE:\n", summary_wage, "\n")
Summary Statistics for WAGE:
 3.94 13 19.3 23.64004 29.8 221.1
> cat("Summary Statistics for EDUC:\n", summary_educ, "\n")
Summary Statistics for EDUC:
 0 12 14 14.2025 16 21
```





b.

```
> cat("Intercept (β1):", beta1, "\n")
Intercept (β1): -10.39996
> cat("Slope (β2):", beta2, "\n")
Slope (β2): 2.396761
```

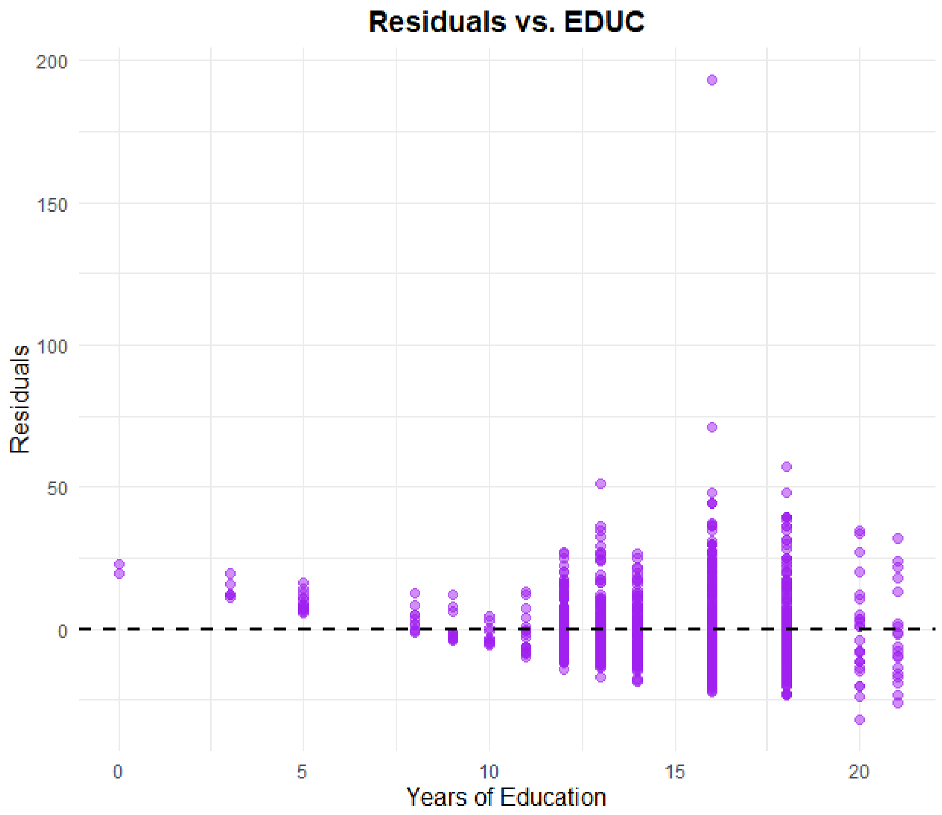
$$wage = -10.39996 + 2.396761 * EDUC$$

EDUC 上升 1，wage 上升 2.396761

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.

c.



➔可看見殘差在高 EDUC 時範圍擴大，違反 SR4(Homoscedasticity)
若 SR1-SR5 成立，不應有任何 pattern

d.

```
Regression Equations:
Males:    WAGE = -8.28 + 2.38*EDUC
Females: WAGE = -16.6 + 2.66*EDUC
Blacks:   WAGE = -6.25 + 1.92*EDUC
Whites:   WAGE = -10.47 + 2.42*EDUC
>
```

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?

e.

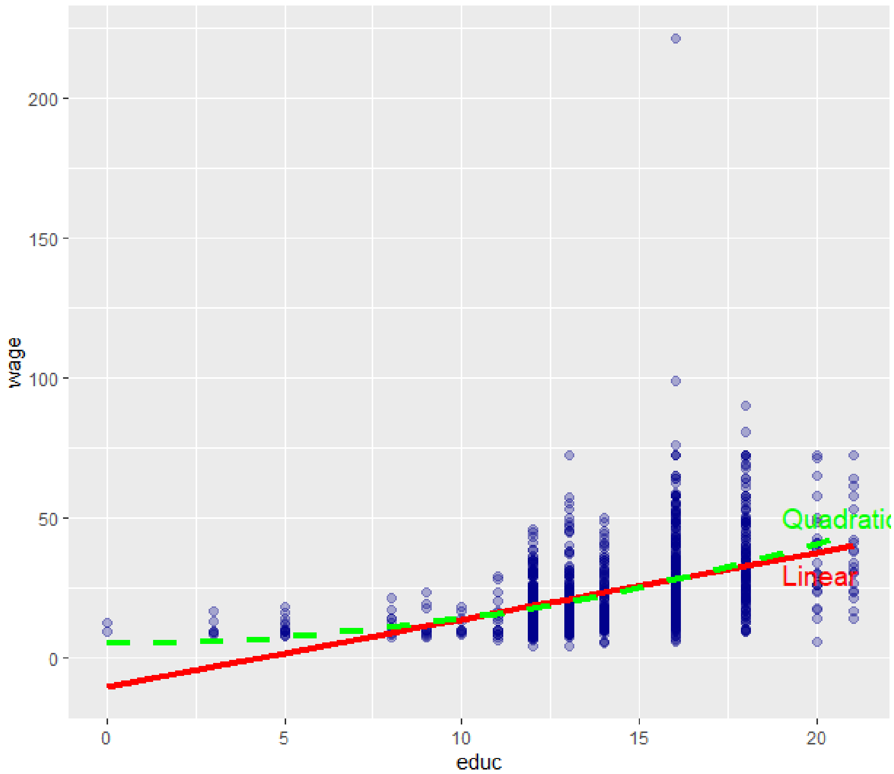
```
Residuals:
    Min      1Q  Median      3Q     Max 
-34.820  -8.117  -2.752   5.248 193.365 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.916477   1.091864   4.503 7.36e-06 ***
EDUC2       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
```

$$wage = 4.916477 + 0.089134 * EDUC^2$$

```
> 
> cat("Marginal effect at EDUC = 12:", marginal_12, "\n")
Marginal effect at EDUC = 12: 2.139216 
> cat("Marginal effect at EDUC = 16:", marginal_16, "\n")
Marginal effect at EDUC = 16: 2.852288 
> cat("Marginal effect from linear model (β2):", beta2, "\n")
Marginal effect from linear model (β2): 2.396761 
```

B 小題之 marginal effect 固定, e 則是隨 EDUC 上升而上升

f.



➔quadratic is better