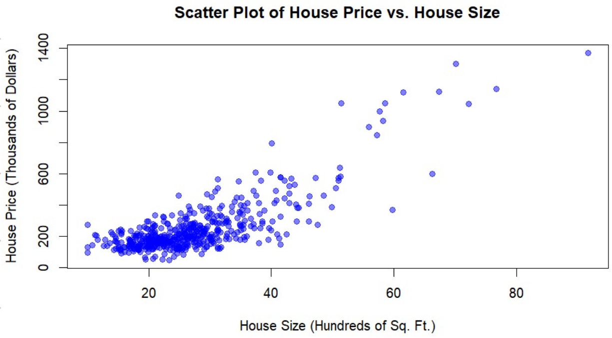
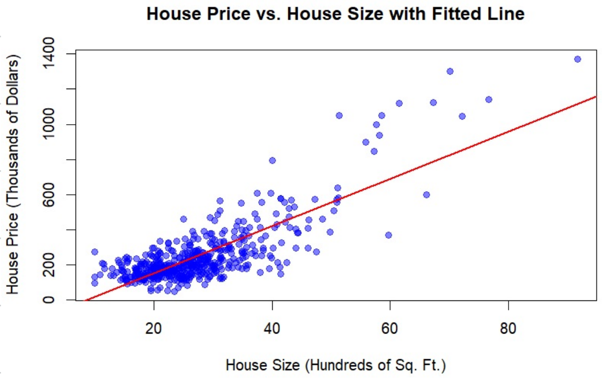
HW0303

17.

a.



b.




```
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 β1  13.0882  -8.819   <2e-16 ***
sqft          13.4029 β2   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
```

The estimated linear regression model for house price is : $\widehat{PRICE} = -115.4236 + 13.4029\, SQFT$

→ $\beta_1 \approx -115.4236$ (intercept) : a house with 0 square feet is estimated as \$-115,423.6

$\beta_2 \approx 13.4029$ (slope) : on average, for every additional 100 square feet of house size, the house price is expected to increase by about \$13,402.9

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 β1 6.072226   15.41   <2e-16 ***
I(sqft^2)    0.184519 β2 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
```

The estimated quadratic regression model for house price is :

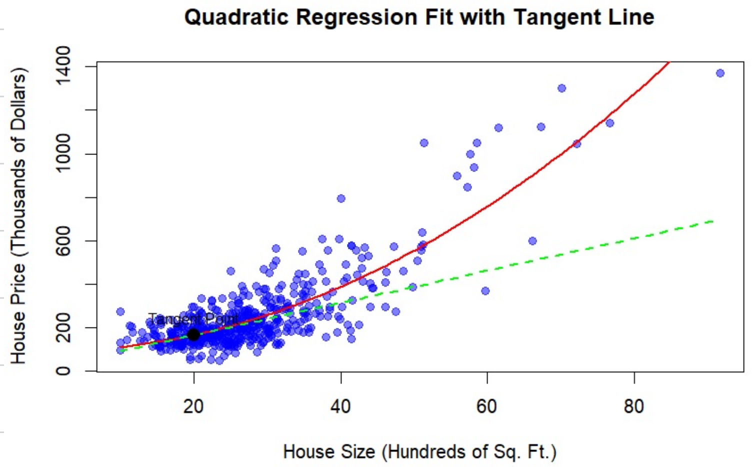
$\widehat{PRICE} = 93.565854 + 0.184519\, sqft^2$

→ Marginal Effect at 2000 square feet :

$\left.\frac{d(PRICE)}{d(SQFT)}\right|_{SQFT=20} = 2 \times 0.184519 \times 20 = 7.38076$

→ An additional 100 square feet of living area in a home with 2000 square feet of living space, the house price is expected to increase by about \$7,380.76

d.

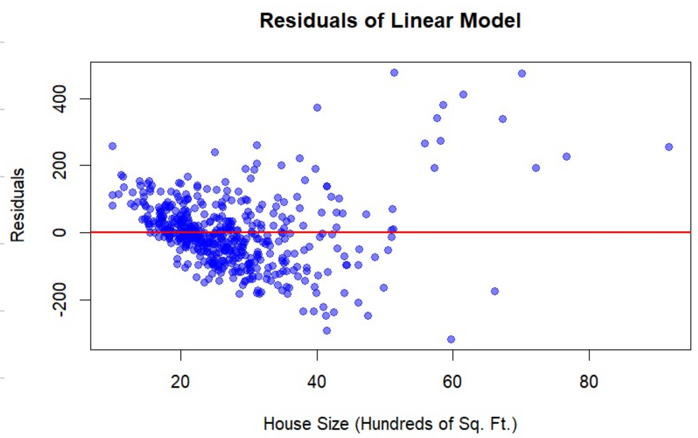


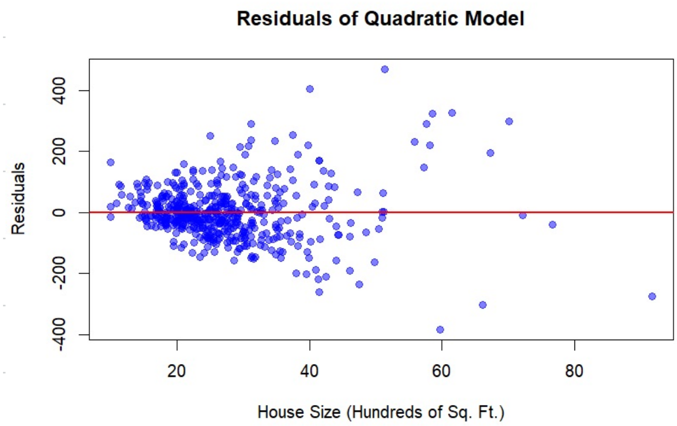
e.

$$\text{Elacity} = \frac{d(\text{PRICE})}{d(\text{SQFT})}\bigg|_{\text{SQFT}=20} \times \frac{\text{SQFT}}{\widehat{\text{PRICE}}}$$

$$= 2 \times 0.184519 \times 20 \times \frac{20}{93.565854 + 0.184519 \times 20^2} = 0.8819511$$

f.





In both the linear and quadratic models, the spread of residuals seems not random. This shows that the homoscedasticity assump. is violated.

g.

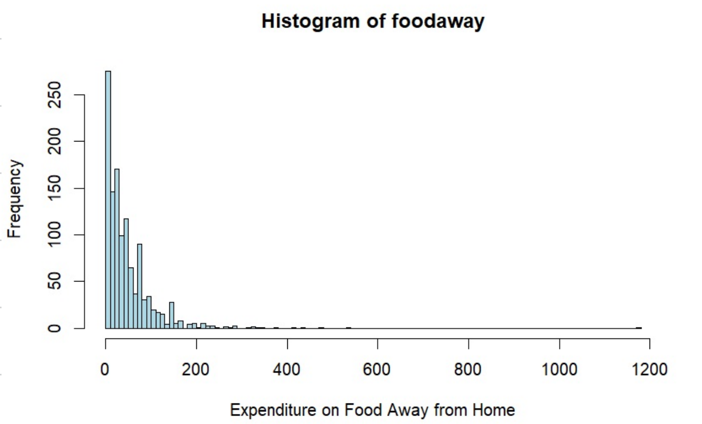
SSE for { linear = 5,262,847
quadratic = 4,222,356

→ The quadratic model has a lower SSE.

→ The model with the lower SSE is the one that better fits the data with less error value.

25.

a.



→ Mean : 49.27085

→ Median : 32.555

→ 25^{th} percentile : 12.04

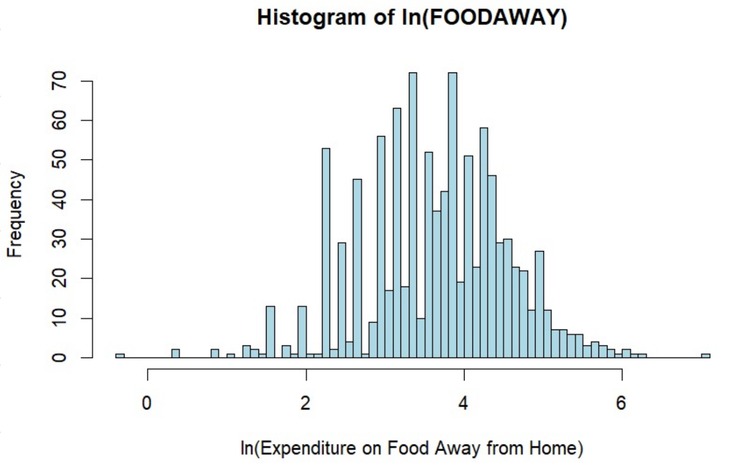
→ 75^{th} percentile : 67.5025

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00   12.04   32.55   49.27   67.50 1179.00
```

b.

	Mean	Median
Advanced degree	73.15494	48.15
College degree	48.59718	36.11
No degree	39.0107	26.02

c.




```
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
-0.3011   3.0759  3.6865  3.6508   4.2797  7.0724
```

→ Mean: 3.650804

→ Median: 3.686499

→ 25th percentile: 3.075929

→ 75th percentile: 4.279717

→ The number of observations of ln(foodway) is fewer by 178 since these 178 original data is less or equal to zero.

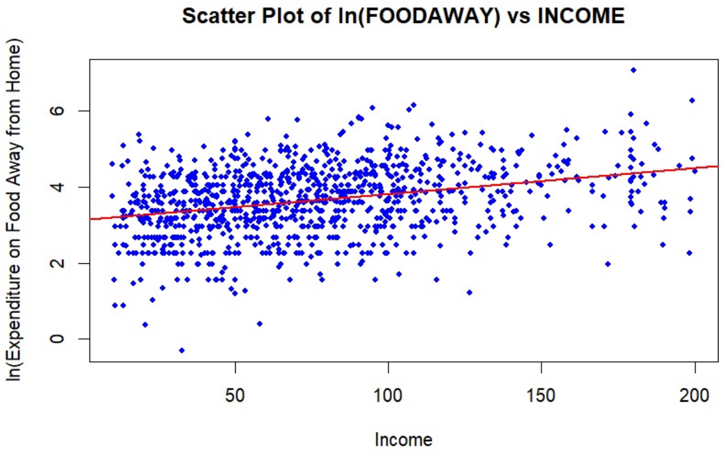
d.

```
Call:
lm(formula = log_foodaway ~ income, data = cex5_cleaned)

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
```

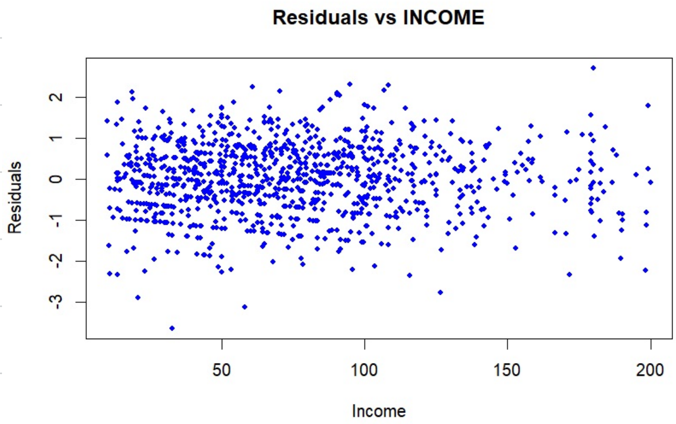
The estimated regression model is: $\widehat{\ln(FOODAWAY)} = 3.1293004 + 0.0069017\,INCOME$

The slope is 0.0069017, meaning that on average, for every additional \$100 in household income, the food away from home expenditure is expected to increase by about 0.69% per one.

e.



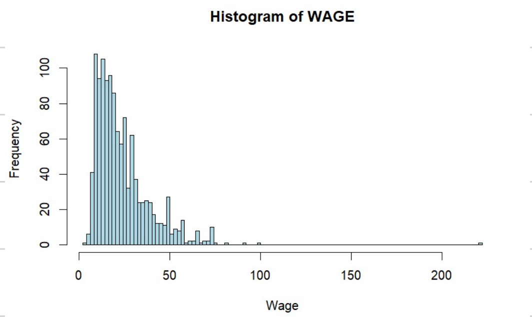
f.

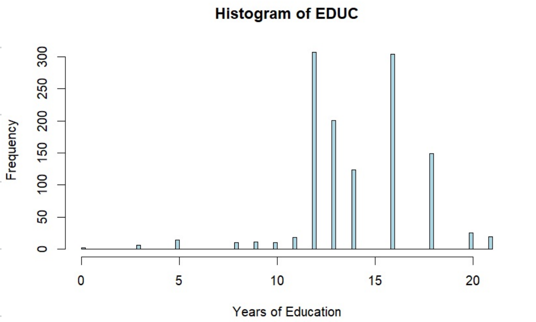


→ The residuals do not show a clear pattern.
They seem to be randomly distributed.

28.

a.





Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.94	13.00	19.30	23.64	29.80	221.10

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	12.0	14.0	14.2	16.0	21.0

→ The histogram of wage is left-skewed, with many low-income earners and a few high-income individuals.

→ The histogram of education is clustered between 12-16 years, indicating structured schooling systems.

→ If we analyze the relationship between wage and education, we may expect the positive relationship. However, the histograms doesn't show as expectation. Therefore, there may be other factors playing a role.

b.

```
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
```

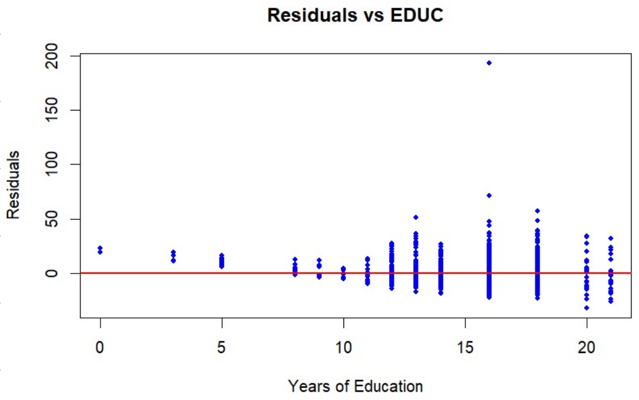
→ $\widehat{WAGE} = -10.4 + 2.3968\ EDUC$

→ Slope (2.3968):

For each additional year of education, the expected wage increases by approximately 2.4 units, showing the education strongly correlated with wages.

→ However, the value of adjusted R-squared is low, indicating there may be other factors influencing the model.

c.



→ The spread of residuals increases as EDUC increases, the higher education with more diverse wages, indicating Heteroskedasticity (SR3).

→ If SR1 ~ SR5 hold, there shouldn't be any pattens evident. The residuals should be distributed uniformly.

d.

```
> print(summary_male)
```

→ $\widehat{WAGE} = -8.2849 + 2.3785EDUC$

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

```
> print(summary_white)
```

→ $\widehat{WAGE} = -10.475 + 2.418EDUC$

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

```
> print(summary_female)
```

→ $\widehat{WAGE} = -16.6028 + 2.6595EDUC$

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

```
> print(summary_black)
```

→ $\widehat{WAGE} = -6.2541 + 1.9233EDUC$

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

The summary :

The expected wage when EDUC = 0 is at least.

Group	Intercept	Education Coefficient	R^2	p-value
Male	-8.2849	2.3785	0.1927	< 2.2e-16
Female	$\boxed{-16.6028}$	$\boxed{2.6595}$	0.2764	< 2.2e-16
White	-10.475	2.418	0.2072	< 2.2e-16
Black	-6.2541	$\boxed{1.9233}$	0.1846	4.79E-06

Education has influenced the female at most and influenced the black at least.

e.

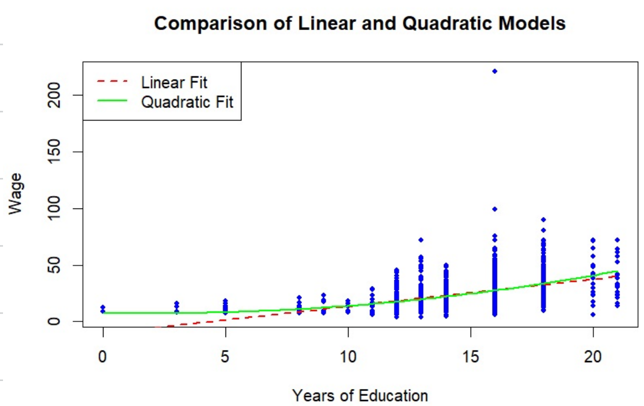
```
Call:
lm(formula = wage ~ educ2, data = cps5_small)

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 ***
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
```

→ $\widehat{WAGE} = 4.961477 + 0.089134\, EDUC^2$

→ $\begin{cases} EDUC = 12, & ME = 2 \times 0.089134 \times 12 = 2.139216 \\ EDUC = 16, & ME = 2 \times 0.089134 \times 16 = 2.852288 \end{cases}$

→ Compared to (b), the marginal effect will variate in (e) instead of a fixed number. In (e), the influence of education increases as the wage increases according to the increasing marginal effect.

f.



→ The quadratic line is more fit with the data.