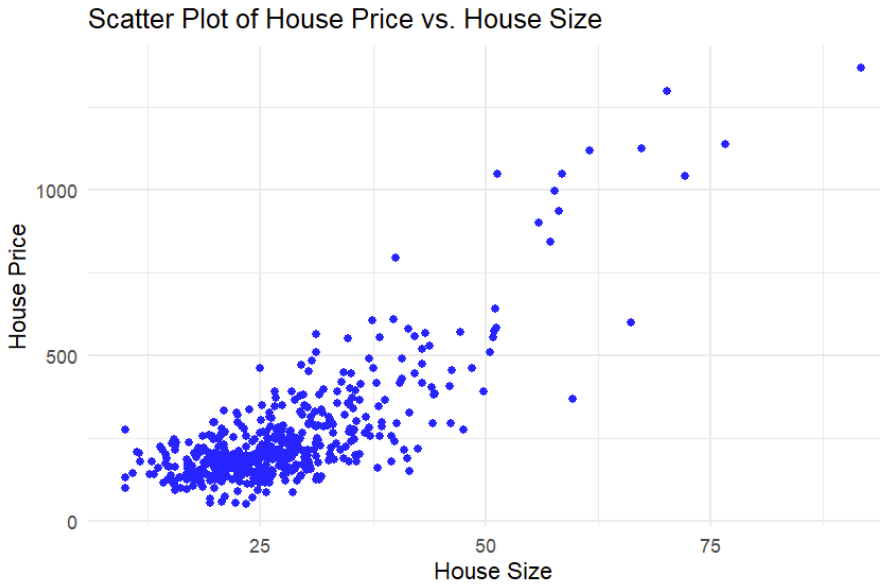
Le Thi Phuong Thao

Student: 413707007

HW0303

Question 2.17

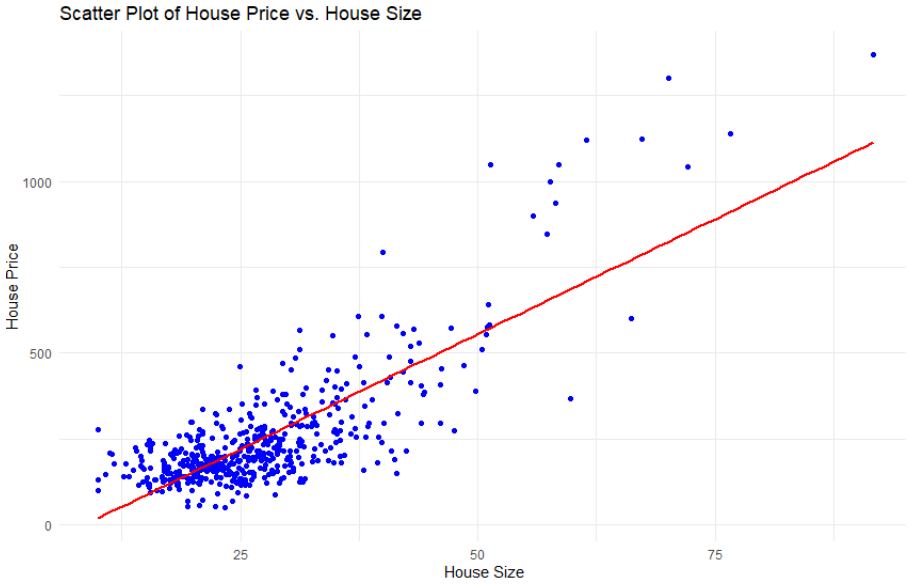
a) *Plot house price against house size in a scatter diagram*



b) *Estimate the linear regression model, draw a sketch of the fitted line*

```
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
```

We estimate that an additional 100 square feet of living area will increase the expected home price by $13,402.94 holding all else constant. The estimated intercept −115.4236 would imply that a house with zero square feet has an expected price of $−115,423.60.



c) *Estimate the quadratic regression model*

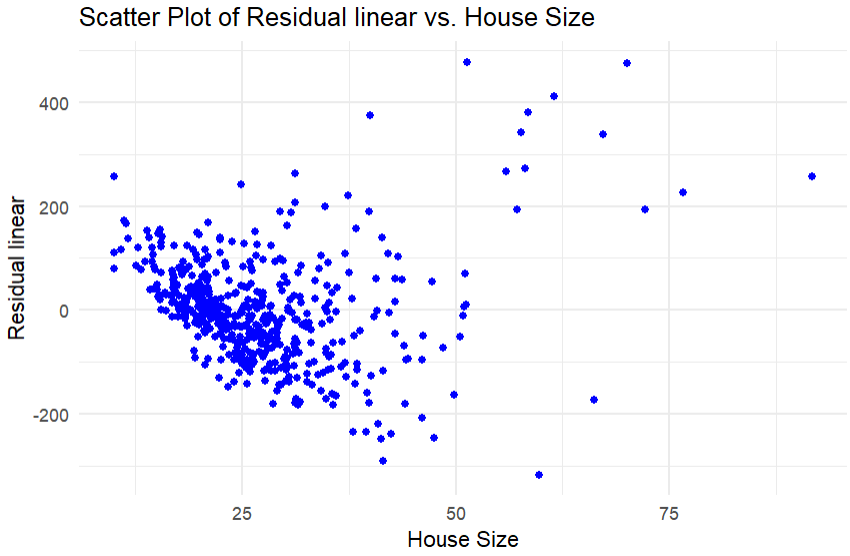
```
Call:
lm(formula = price ~ I(sqft^2), data = collegetown)

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
```

We estimate that an additional 100 square feet of living area for a 2000 square foot home will increase the expected home price by $7,380.80 holding all else constant

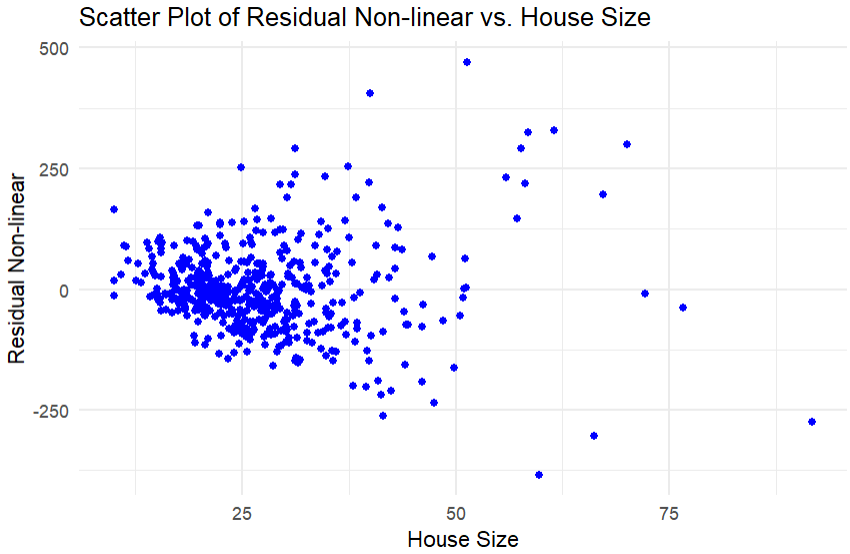
d) *Graph the fitted curve for the model*

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*
Ans: 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*
Linear



Non-linear



In both models, the residual patterns do not appear random. The variation in the residuals increases as SQFT increases, suggesting that the homoskedasticity assumption may be violated.

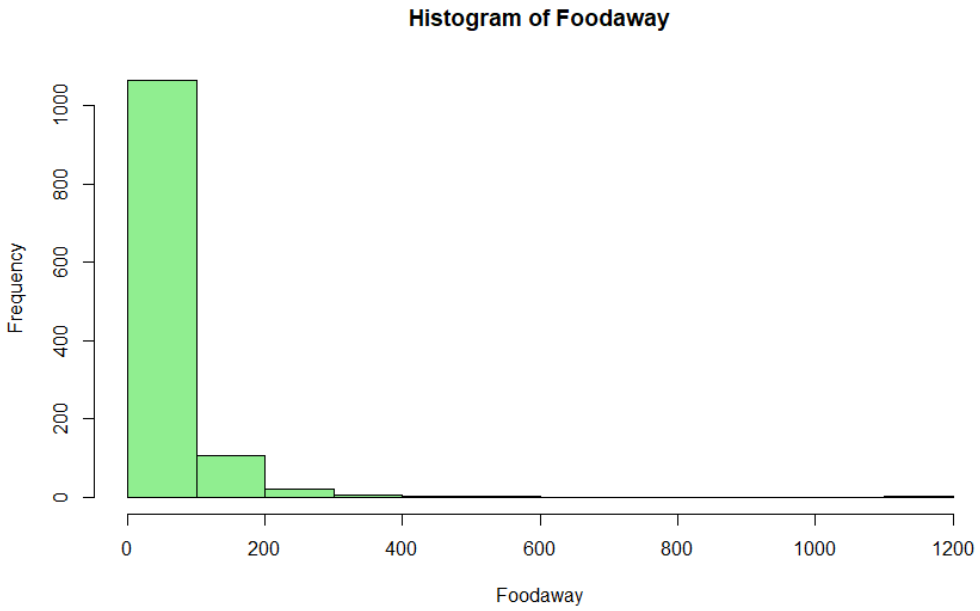
g) *One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (SSE) from the models in (b) and (c). Which model has a lower SSE? How does having a lower SSE indicate a "better-fitting" model*

```
> SSE_linear
[1] 5262847
> SSE_quad
[1] 4222356
```

The sum of square residuals linear relationship is 5,262,847. The sum of square residuals for the quadratic relationship is 4,222,356. In this case the quadratic model has the lower SSE. The lower SSE means that the data values are closer to the fitted line for the quadratic model than for the linear model

Question 2.25

a) *Construct a histogram of FOODAWAY and its summary statistics. What are the mean and median values? What are the 25th and 75th percentiles*




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

P25 = 1st quarter = 12.04

P75 = 3rd quarter = 67.50

Mean = 49.27

Median = 32.55

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*

The answers are respectively

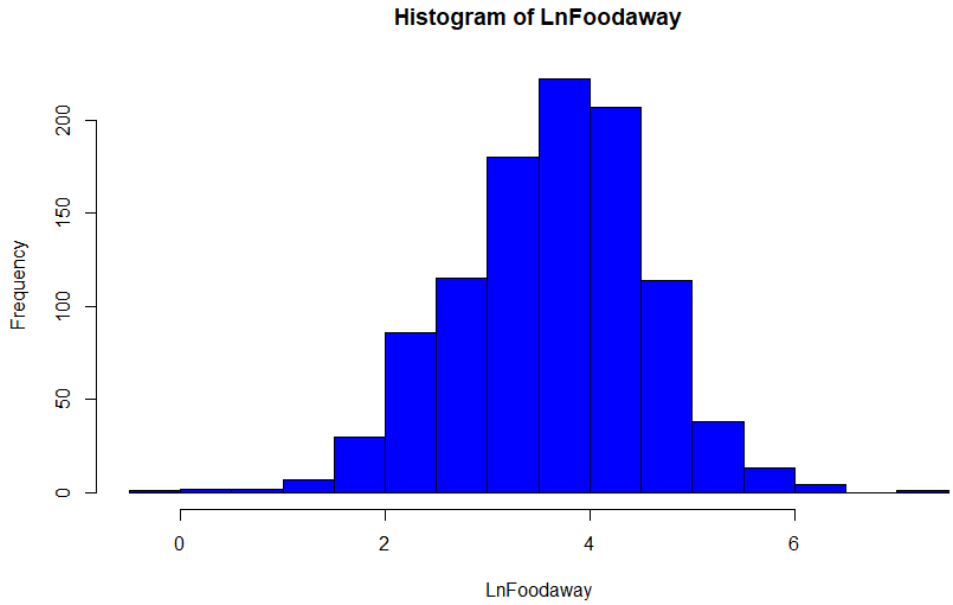
73.15 & 48.15

48.60 & 36.11

39.01 & 26.02

```
> summary(subset(cex5_small,advanced == 1)$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   21.67   48.15   73.15   90.00 1179.00
> summary(subset(cex5_small,college == 1)$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   14.44   36.11   48.60   68.67  416.11
> summary(subset(cex5_small,(advanced == 0 & college == 0))$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.63   26.02   39.01   52.65  437.78
```

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




```
> summary(cex5_small$lnfoodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-0.3011  3.0759  3.6865  3.6508  4.2797  7.0724     178
```

```
> length(cex5_small$lnfoodaway)
[1] 1022
```

There are 1200 observations in raw foodaway data and 1022 observations for its logarithm data since Foodway has 0 value, ln(0)= ∞ so log(foodway) is not computable

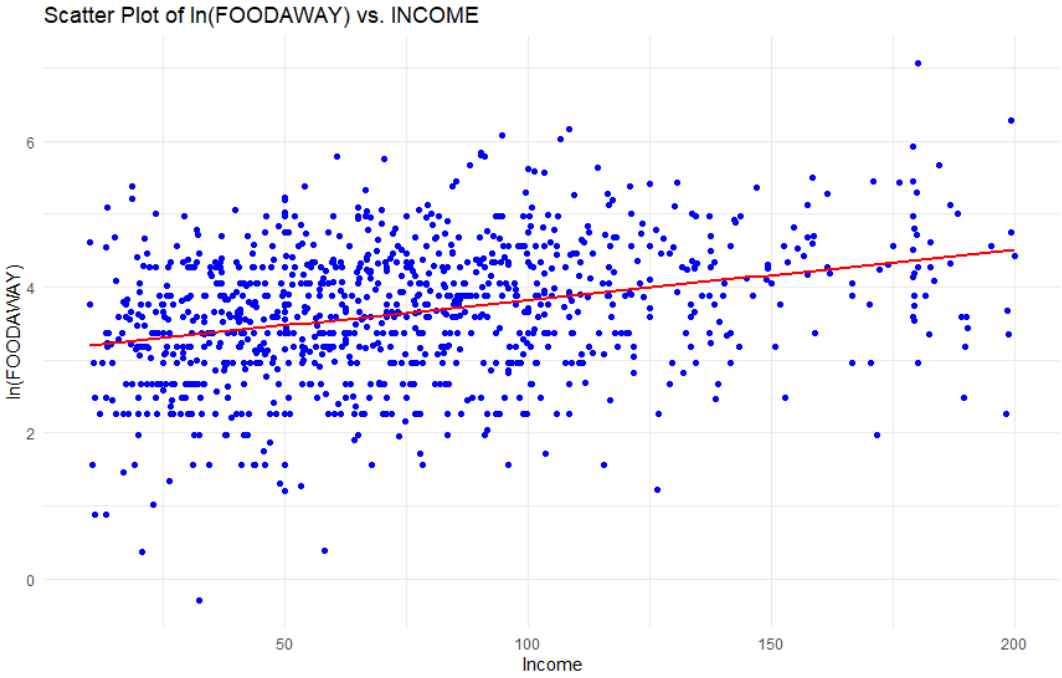
d) *Estimate the linear regression*

Slop b2

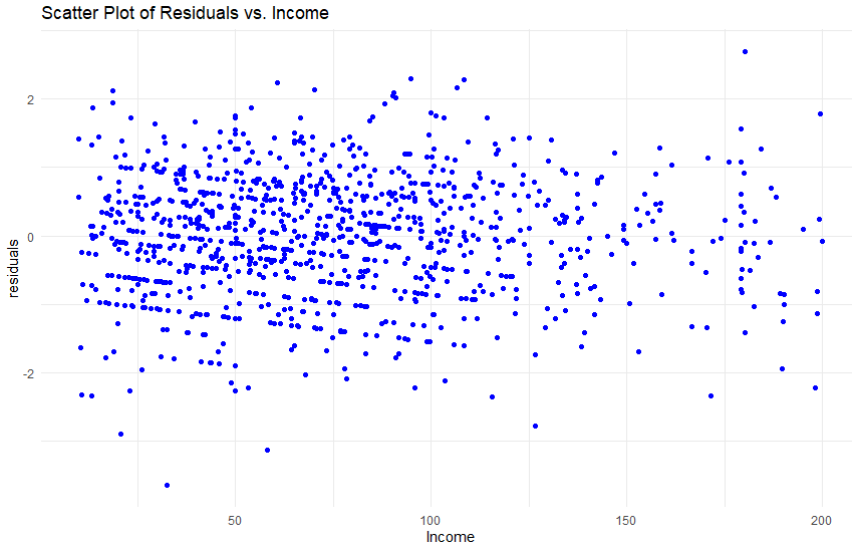
```
> b2 <- coef(d_regression) [[2]]
> print(b2)
[1] 0.006901748
```

A unit increase in INCOME is associated with a 0.69% increase in FOODAWAY spending.

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?*

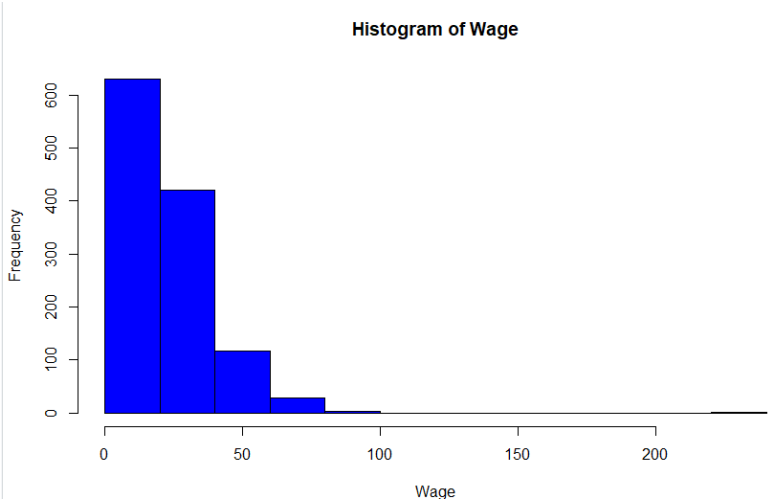


The residual is quite randomly distributed, there is more whitespaces of the residual plot when income is increased (there are a few high-income observations)

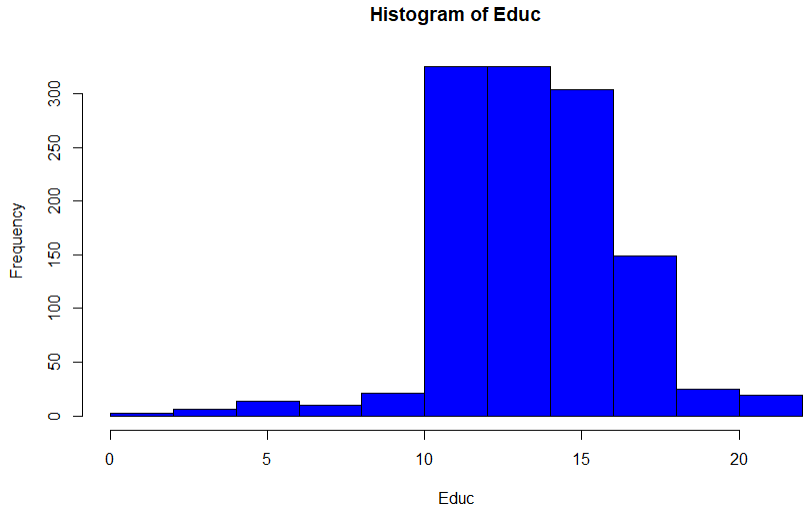
Question 2.28

a) Summary statistics

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



The wage histogram is right-skewed, meaning most people earn relatively low wages, while a small number earn very high wages.



The education histogram is more centered, with many people having around 10 to 16 years of education. It appears more balanced (less skewed) compared to wage.

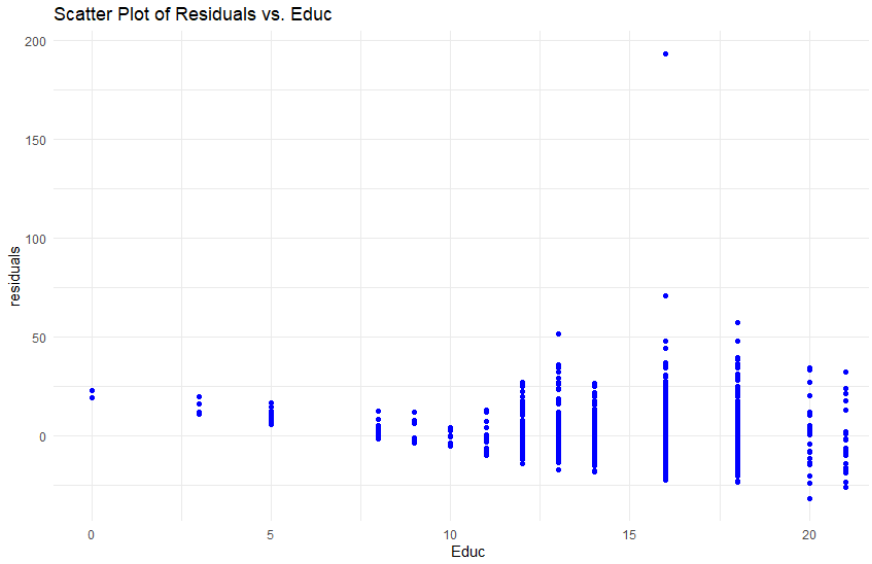
Overall, wage is heavily skewed, while education is more normally distributed

b) Estimate regression

```
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
```

Education has a positive and significant impact on wage. For each additional year of education, wage increases by about $2.15 on average. R squared means around 20.7% of the variation in wage is explained by education alone.

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



The magnitude of the residuals increases with the large of Educ, suggesting that the error variance is larger for larger values of EDUC.

d) Estimate the quadratic regression by group and compare results

Regression Results by Group

	Dependent variable:			
	Wage			
	Females (1)	Males (2)	Blacks (3)	Whites (4)
educ	2.659***	2.378***	1.923***	2.418***
	(0.188)	(0.188)	(0.398)	(0.143)
Constant	-16.603***	-8.285***	-6.254	-10.475***
	(2.784)	(2.674)	(5.554)	(2.081)
Observations	528	672	105	1,095
R2	0.276	0.193	0.185	0.207
Adjusted R2	0.275	0.192	0.177	0.206
Residual Std. Error	11.504 (df = 526)	14.706 (df = 670)	10.506 (df = 103)	13.792 (df = 1093)
F Statistic	200.914*** (df = 1; 526)	159.967*** (df = 1; 670)	23.319*** (df = 1; 103)	285.669*** (df = 1; 1093)
Note:				*p<0.1; **p<0.05; ***p<0.01

In all sub-sample, education has a positive and statistically significant effect on wages. Females show the highest estimated increase in education, while blacks have the lowest among these groups. Comparing gender and race, education has more influence on wages than race factor. Overall, education alone explains only part of wage variation, indicating other factors also play a role (R square is less than 30%)

e) 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?

The quadratic model (green line) appears fit the data better