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.

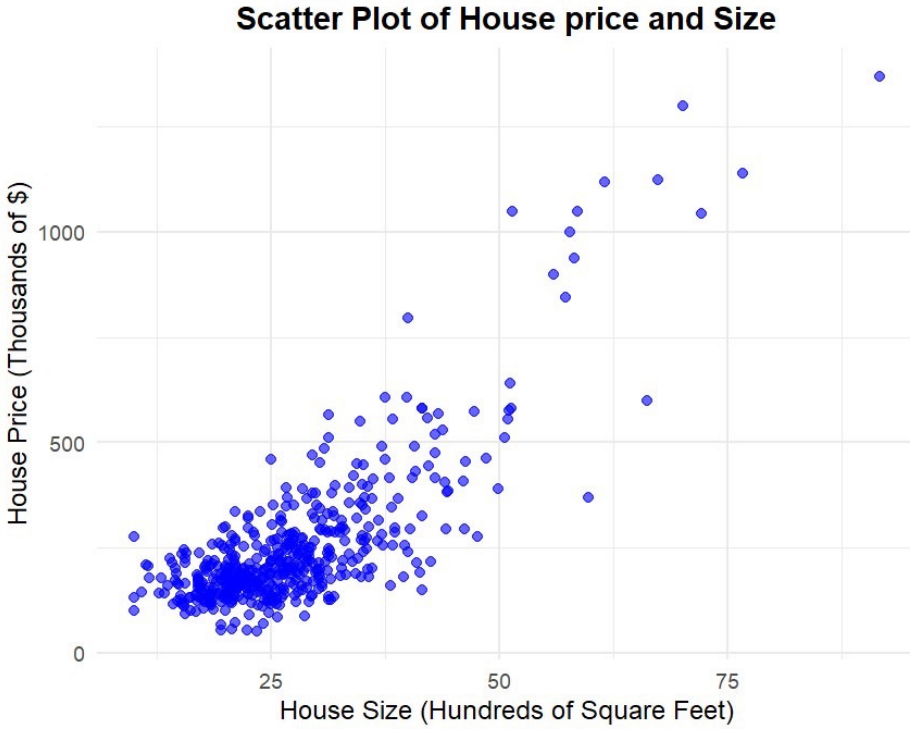
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

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)

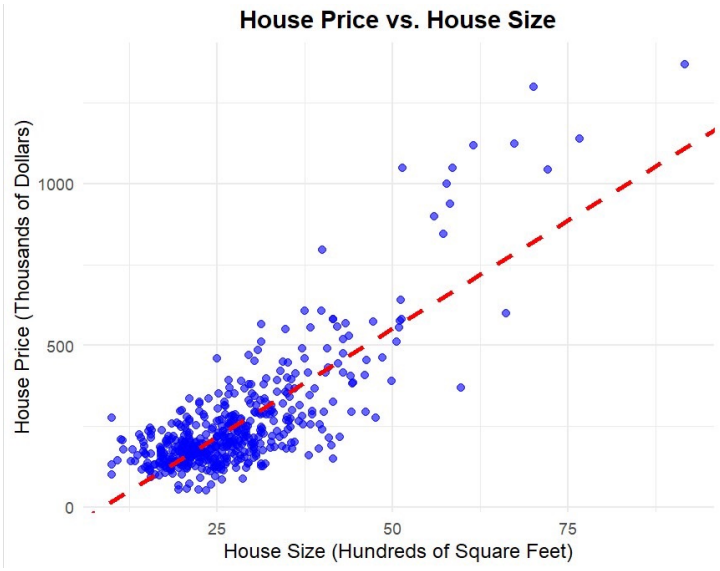


(b)

```
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 ***
```

Price = −115.4236 + 13.4029 SQFT

1. It shows that if SQFT = 0, than the costumer need to pay −115,423.6 dollars.
2. In addition, for increasing 100 of square feet, the house price goes up by 13,402.9 $



(C)

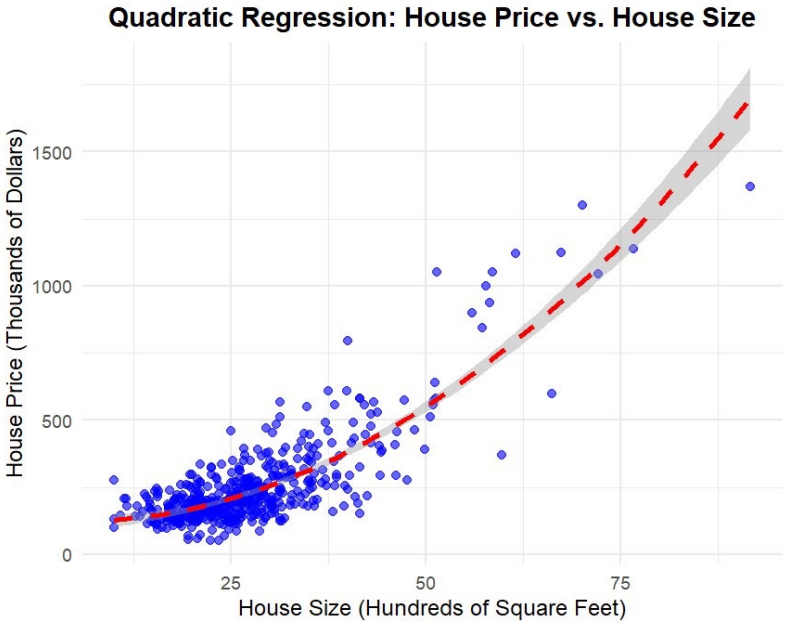
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.565854   6.072226   15.41   <2e-16 ***
I(sqft^2)    0.184519   0.005256   35.11   <2e-16 ***
---
```

unit : 100 sqft

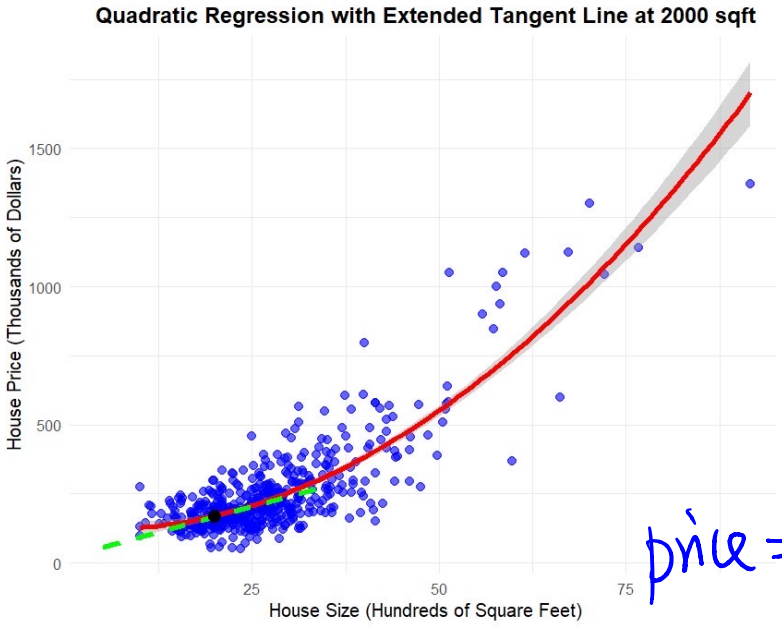
$$\text{Price} = 93.5659 + 0.1845\,SQFT^2$$

$$\text{Marginal effect} = \frac{d(\text{Price})}{d(SQFT)} = 2(0.1845)SQFT = 0.369\,SQFT$$

$0.369\,(20) = 7.38$ (thousand \$) $= 7{,}380$ \$



(d)



$price = \alpha_1 + \alpha_2 SQFT^2$

(e) when $SQFT = 20$, $price = 93.5659 + 0.1845(20^2)$

$= 167.3659$

$$elasticity = \frac{d(price)}{d(SQFT)} \cdot \frac{SQFT}{price}$$

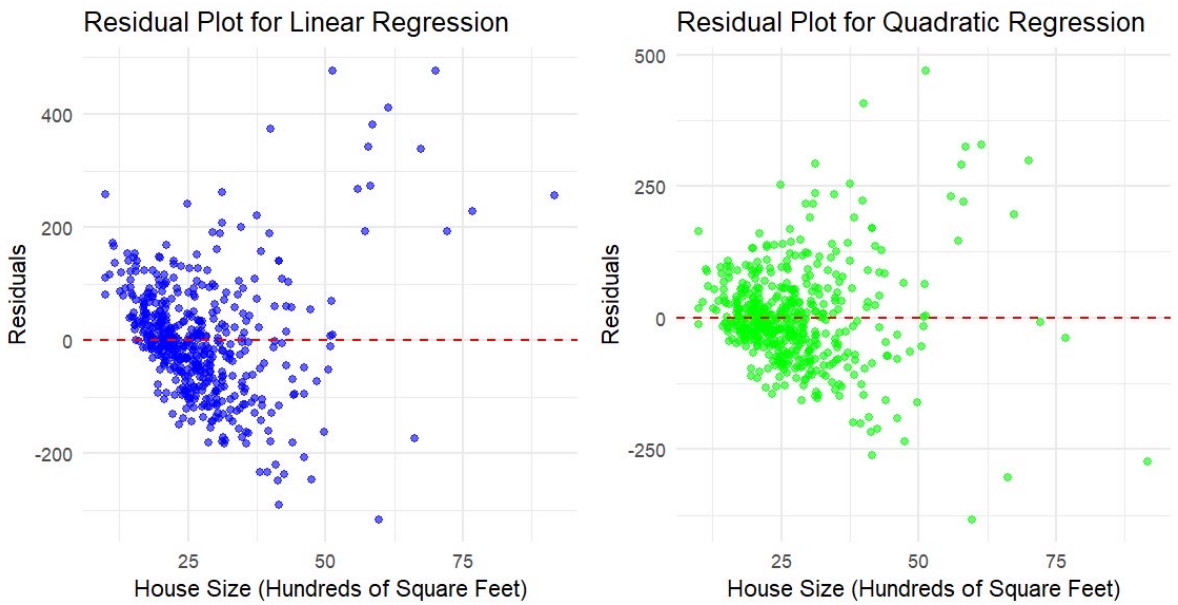
$$= 2\alpha_2 SQFT \times \left(\frac{SQFT}{price}\right)$$

$$= 2 \times 0.1845 \times 20^2 \times \frac{1}{167.3659}$$

$$= 0.8819$$

It implies that SQFT ↑1%, then price ↑0.8819%

(f)



Since the residual is non-random, both two assumptions may be violated.

(g)

(b) SSE of linear regression = 5,262,846.95

(c) SSE of quadratic regression = 4,222,356.35

∵ c < b ∴ quadratic regression is a better-fitting model in this case.

→ lower SSE means the forecast is closer to the real data.

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.

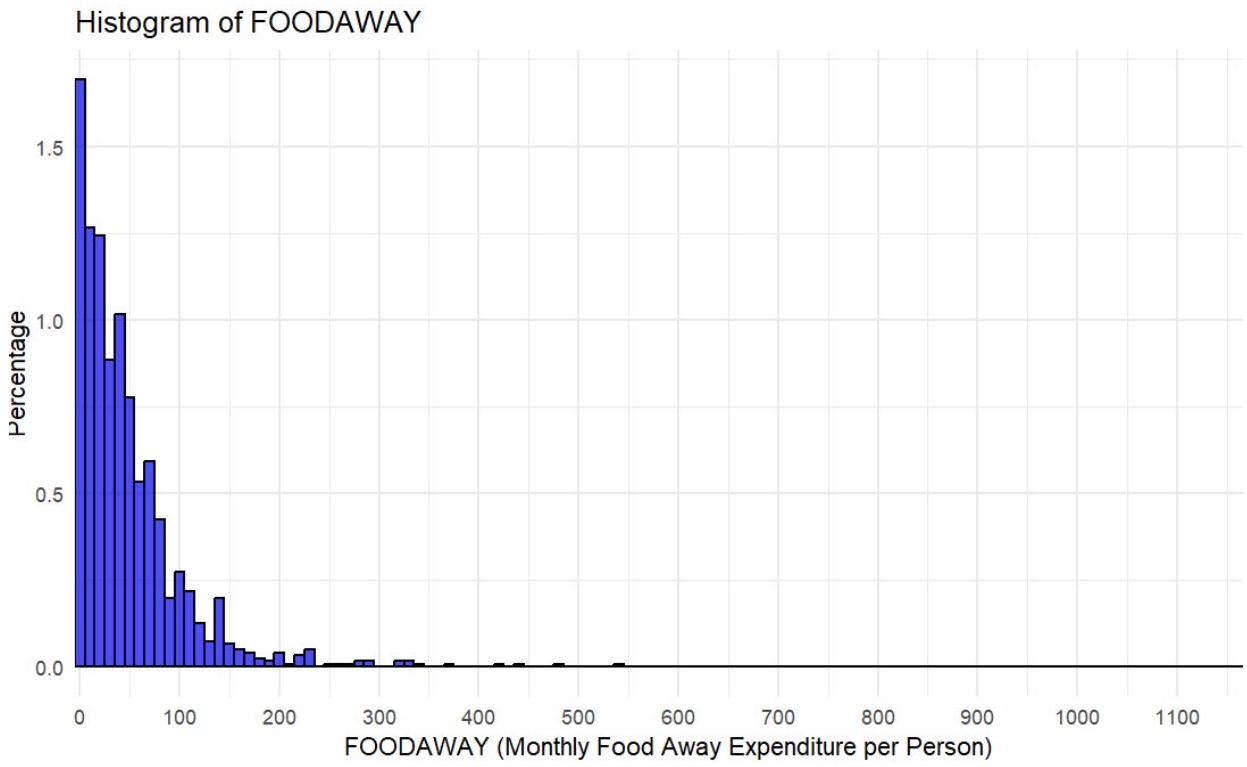
d. Estimate the linear regression $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

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)

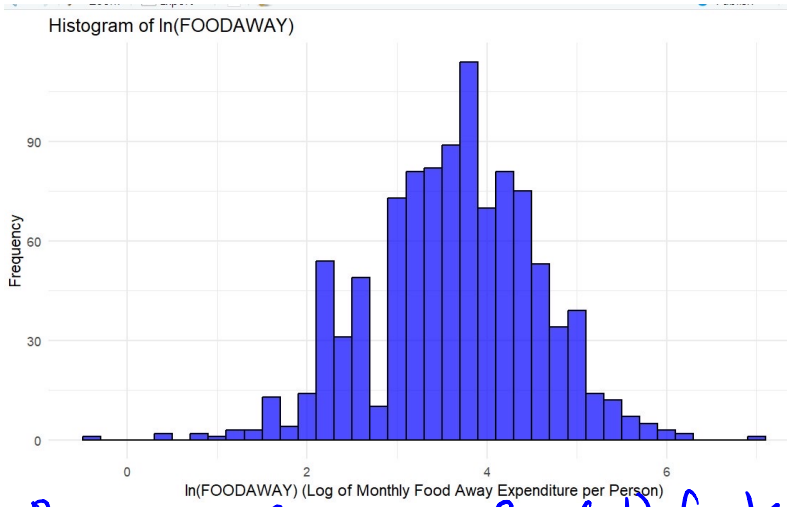
```
> cat("Mean:", mean_foodaway, "\n")
Mean: 49.27085
> cat("Median:", median_foodaway, "\n")
Median: 32.555
> cat("25th Percentile:", percentile_25, "\n")
25th Percentile: 12.04
> cat("75th Percentile:", percentile_75, "\n")
75th Percentile: 67.5025
```



(b)

```
Advanced Degree - Mean: 73.15494 Median: 48.15
> cat("College Degree - Mean:", mean_college, "Median:", me
dian_college, "\n")
College Degree - Mean: 48.59718 Median: 36.11
> cat("No Degree - Mean:", mean_no_degree, "Median:", media
n_no_degree, "\n")
No Degree - Mean: 39.01017 Median: 26.02
```

(c)



Because log(0) or log(N), for N<0 is not define, some obv. will not be included in our sample, so obv. of ln(Foodaway) is less than Foodaway.

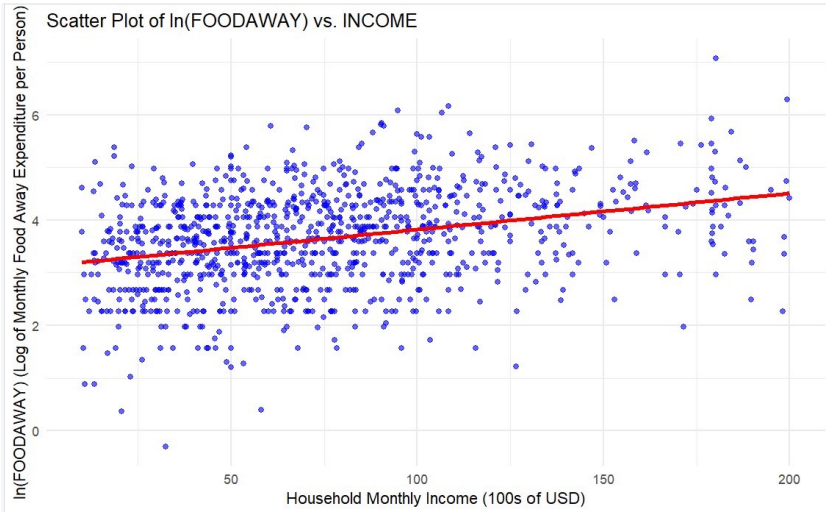
(d)

```
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 ***
```

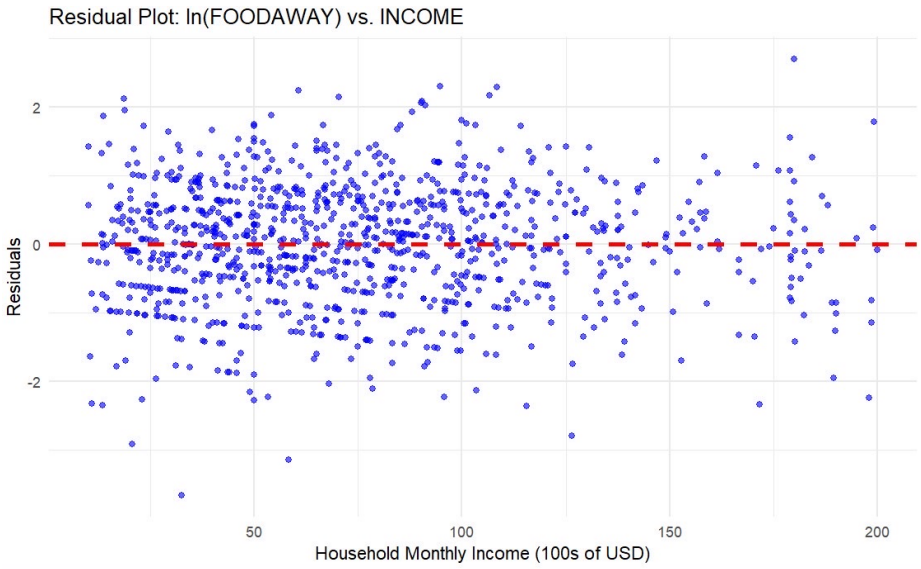
$\ln(\text{Foodaway}) = 3.1293 + 0.0069\,\text{income}$

Because it's ln linear model, β_1 can be explained as increase in 1 unit ($100) of income, the foodaway will change 0.69%

(e)



(f)



From the graph, we can observe that there is no linear trend for the data. Hence the OLS model works.

residual

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?

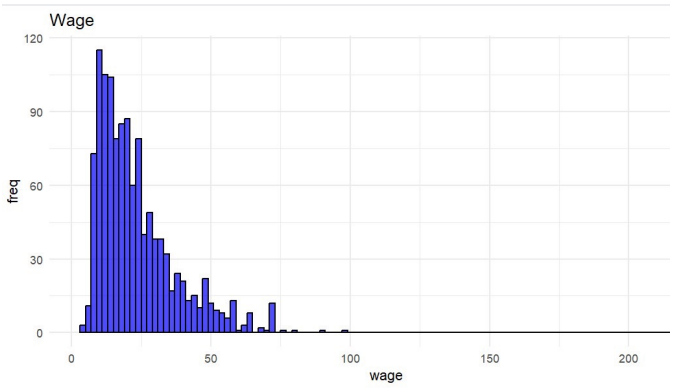
d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

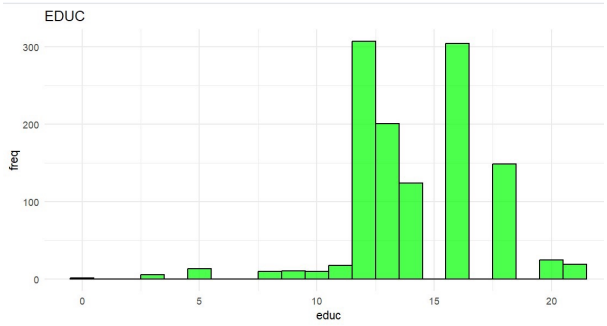
e. Estimate the quadratic regression $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$ and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 16 years of education. Compare these values to the estimated marginal effect 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?

2.28

(a)





The wage histogram is right-skewed distribution, it means that there are lots of low wage people.

The eduction histogram indicates that the eduction degree is concentrate in 12 ~ 16 years.

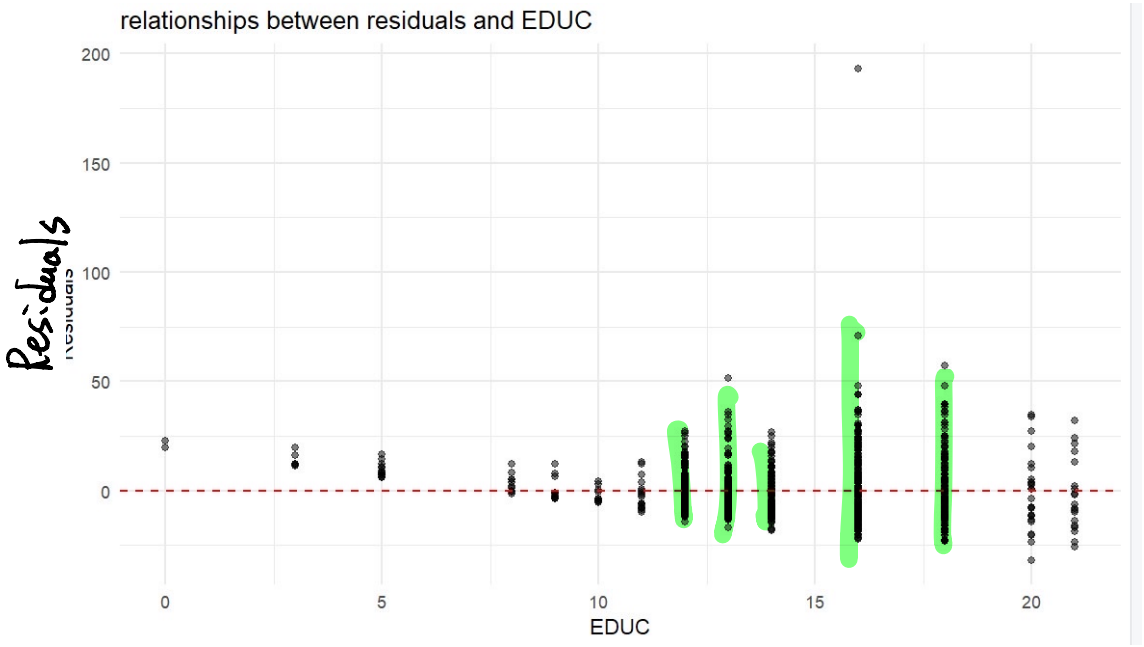
(b)

```
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 ***
---
```

linear regression: WAGE = -10.4000 + 2.3968 EDUC

It means that increase EDUC in 1 year will increase WAGE by 2.3968 units

(c)



The graph indicates that there is more diverse with highly educated people. It means that the regression violates SR4 (Homoscedasticity) → Heteroscedasticity

If SR1 ~ SR5 holds, there shouldn't be any pattern of residuals.

(d) According to the R code results:

Male: Wage = −8.3849 + 2.3785EDUC

Female: Wage = −16.6028 + 2.6595EDUC

↳ The education influences female the most.

White: Wage = 10.475 + 2.418EDUC

Black: Wage = −6.3541 + 1.9233EDUC

The expect value when EDUC=0, female is the least.

(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 ***
I(educ^2)   0.089134   0.004858  18.347  < 2e-16 ***
---
```

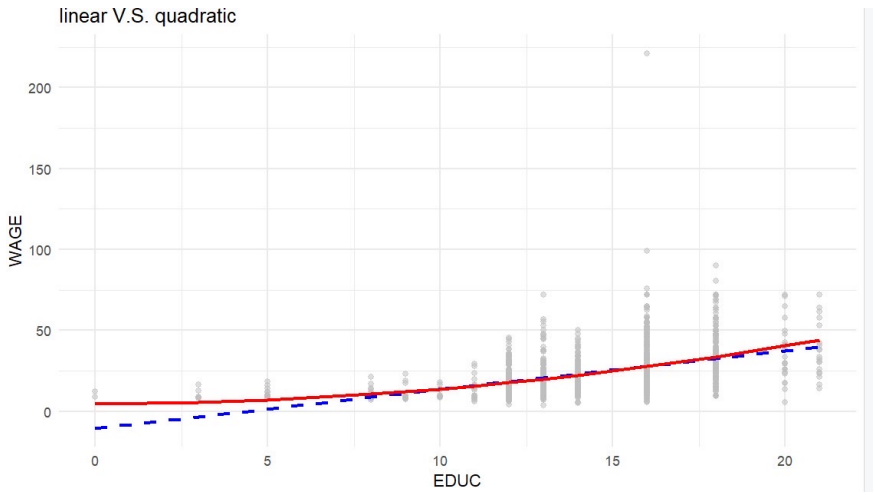
Estimate regression: $Wage = 4.9165 + 0.0891 EDUC^2$

when EDUC = 12: Marginal effect $= 2 \times 0.0891 \times 12 = 2.1392$

when EDUC = 16: Marginal effect $= 2 \times 0.0891 \times 16 = 2.8522$

The marginal effect of (b) is a fixed num. (2.3968)
while ME of (e) increase more when EDUE↑.

(f)



The quadratic line is better fitted with the data.