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.

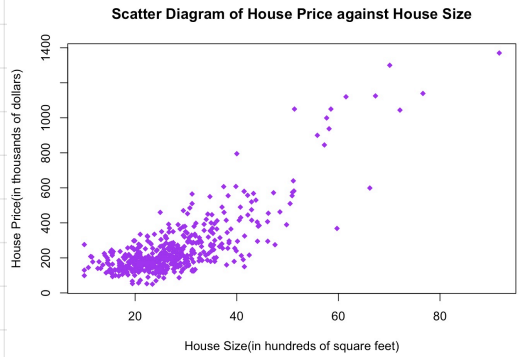
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

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?

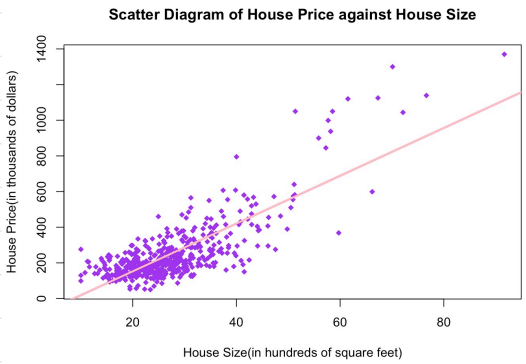
g. One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?

a、



b、 $\widehat{price} = -115.4236 + 13.4029\ SQFT + e$

每增加 100 square feet in size, house price 平均上升 $13.4029 thousands。



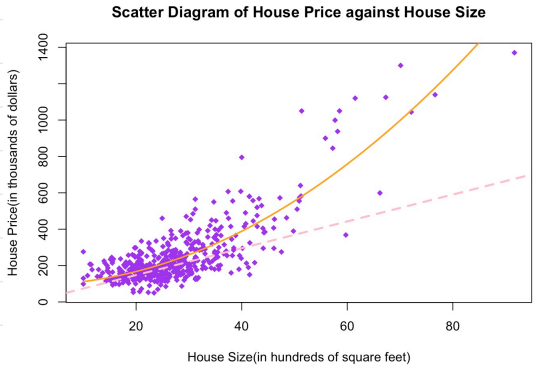

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -115.4236    13.0882  -8.819   <2e-16 ***
x             13.4029     0.4492  29.840   <2e-16 ***
```

c、 $\widehat{price} = 93.5658 + 0.1845\ SQFT^2 + e$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.565854   6.072226   15.41   <2e-16 ***
x2           0.184519   0.005256   35.11   <2e-16 ***
```

$$\frac{\partial E(\widehat{price} \mid SQFT = 20)}{\partial SQFT} = 2 \times 0.1845 \times 20 = 7.381$$

d.

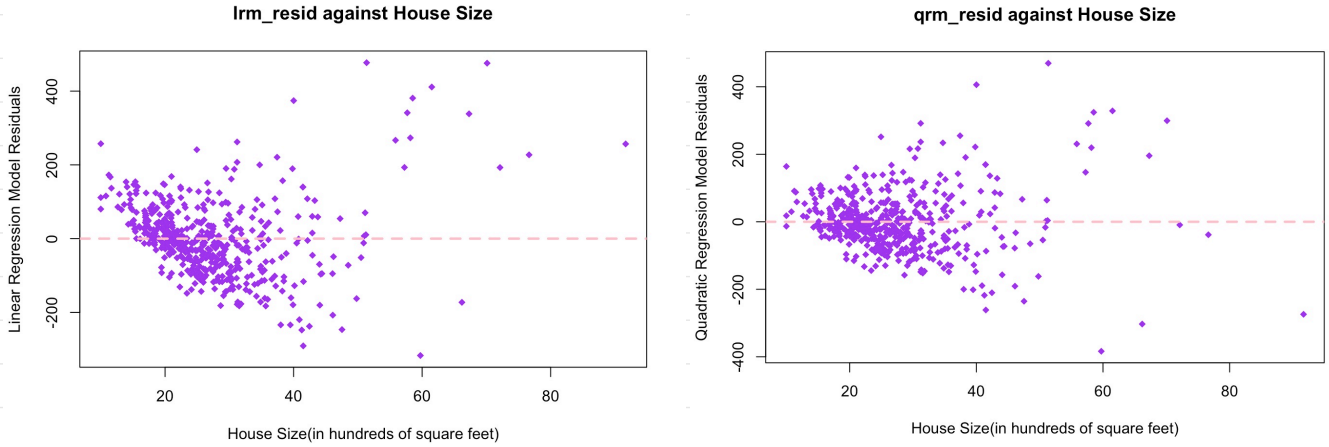


e.

$$\varepsilon = \frac{\partial \widehat{price} / \widehat{price}}{\partial SQFT / SQFT} = \frac{\partial \widehat{price}}{\partial SQFT} \times \frac{SQFT}{\widehat{price}}$$

$$= 7.381 \times \frac{20}{93.565854 + 0.184519 \times 20^2} = 0.88198$$

f.



The residual plot of the linear regression model shows that as SQFT increases, the variance of the residuals also increases, indicating a violation of the homoscedasticity assumption.
The residuals from the quadratic regression model are more randomly distributed, with a more consistent variance, suggesting that the homoscedasticity assumption is more reasonable in this case.
Overall, the quadratic regression model better satisfies the basic assumptions of regression when explaining the relationship between PRICE and SQFT, particularly by improving the issue of heteroscedasticity.

g.

```
> lrm_sse = sum(lrm_resid^2)
> lrm_sse
[1] 5262847
>
> qrm_sse = sum(qrm_resid^2)
> qrm_sse
[1] 4222356
```

The Sum of Squared Errors (SSE) for the quadratic regression model is lower (4,222,356) compared to the linear regression model (5,262,847). A lower SSE indicates that the quadratic model provides a better fit to the data because it reduces the discrepancy between the predicted values and the actual observed values. This suggests that the quadratic model explains the relationship between PRICE and SQFT more effectively than the linear model.

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?

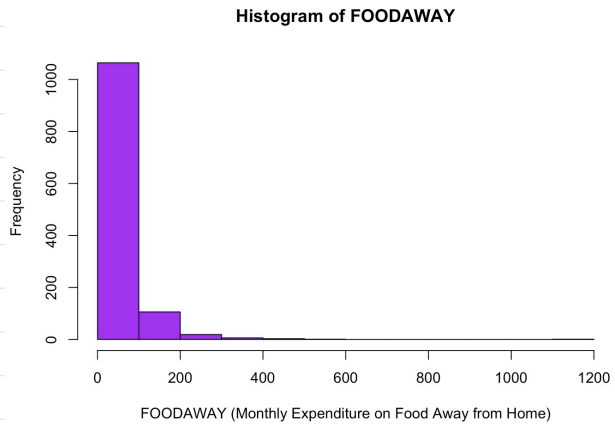
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

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 = 49.27085

median = 32.555

Q_1 = 12.04

Q_3 = 67.5025

b. advanced degree: mean = 73.15494

median = 48.15

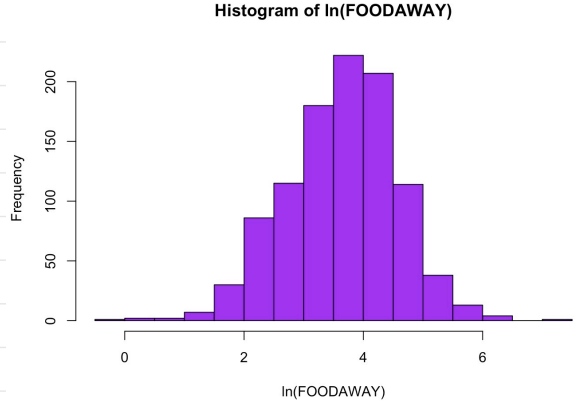
college degree: mean = 48.59718

median = 30.11

no advanced or college degree: mean = 39.01017

median = 26.02

c.



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

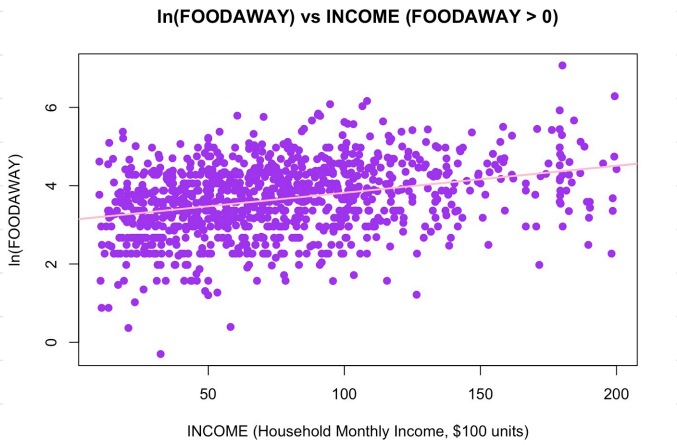
There are 1200 observations for FOODAWAY, but only 1022 finite observations for ln(FOODAWAY). This discrepancy occurs because 178 households reported zero expenditures for FOODAWAY. Taking the natural logarithm of zero results in negative infinity (-Inf) in R. These -Inf values are not considered valid for further statistical analysis, reducing the number of usable observations in ln(FOODAWAY).

d.

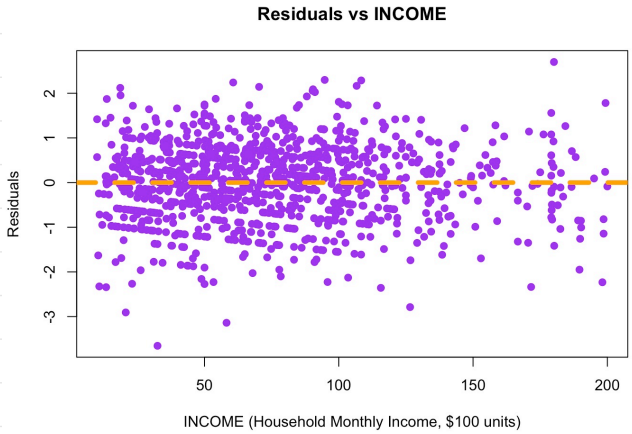
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.1293004  0.0565503   55.34   <2e-16 ***
x_income    0.0069017  0.0006546   10.54   <2e-16 ***
```

$\widehat{\ln(FOODAWAY)} = 3.1293 + 0.0069\,INCOME + e$

e.



f.



In the residual plot from part (f), the residuals are fairly evenly scattered around zero, with no clear systematic patterns. This suggests that the errors appear to be randomly distributed and there is no strong evidence of heteroscedasticity or non-linearity in the relationship between ln(FOODAWAY) and INCOME.
Although there is slightly more variation in residuals at higher income levels, overall the assumption of homoscedasticity seems reasonable.

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.

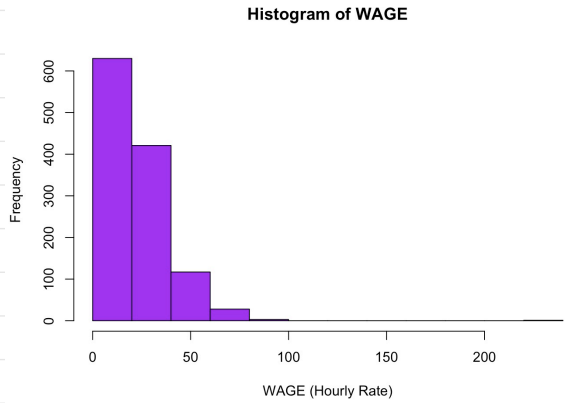
c. Calculate the least squares residuals and plot them against *EDUC*. Are any patterns evident? If assumptions SR1–SR5 hold, should any patterns be evident in the least squares residuals?

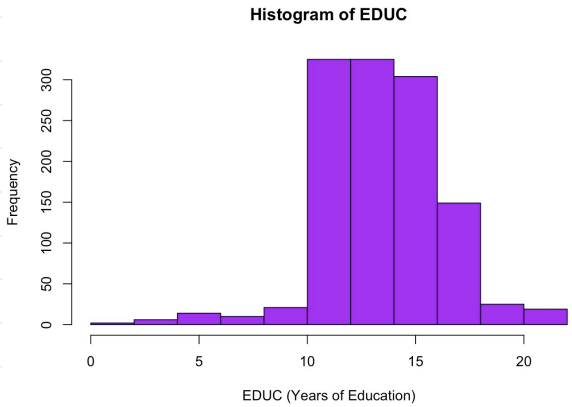
d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

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(cps5_small$wage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.94   13.00   19.30   23.64   29.80  221.10
> summary(cps5_small$educ)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    12.0    14.0    14.2    16.0    21.0
```

The summary statistics and histograms for WAGE and EDUC reveal distinct distribution patterns. WAGE (hourly wage rates) is highly right-skewed. Most of the observations are concentrated between $0 and $50 per hour, with a mean of $23.64 and a median of $19.30. The maximum wage of $221.10 suggests the presence of outliers or a long-tail distribution. EDUC (years of education) is more symmetrically distributed, mostly ranging between 12 and 16 years. The mean education level is 14.2 years, indicating that most individuals have at least a high school diploma and many have completed higher education. Overall, the wage distribution shows considerable inequality, while the education distribution is relatively concentrated around common education milestones.

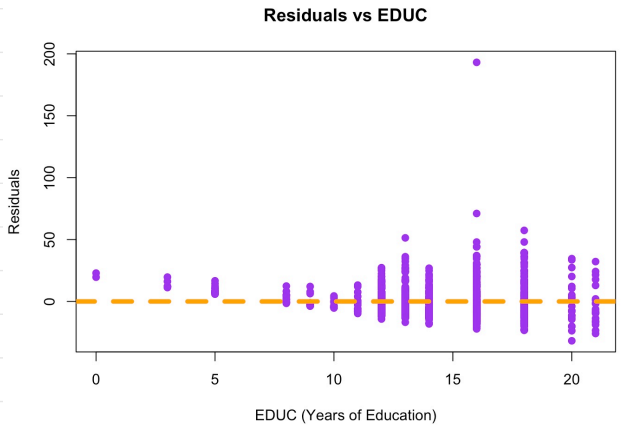
b.

```
Call:
lm(formula = cps5_small$wage ~ cps5_small$educ)

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
cps5_small$educ   2.3968     0.1354    17.7  < 2e-16

(Intercept)     ***
cps5_small$educ ***
---
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
```

The regression results indicate a significant positive relationship between years of education and hourly wage.
The estimated regression equation is:
`WAGE = -10.4000 + 2.3968 * EDUC`
This suggests that each additional year of education is associated with an increase of approximately $2.40 in hourly wage.

The R-squared value is 0.2073, meaning that education explains about 20.7% of the variation in wages.
While the model shows a significant relationship, other factors likely contribute to wage differences beyond education.

c.



In part (c), the residual plot shows a clear funnel-shaped pattern. The variance of residuals increases as EDUC increases, indicating heteroscedasticity.
This violates assumption SR4, which requires homoscedasticity (constant variance of the error term).
The model appears to perform worse for individuals with higher levels of education, as shown by the greater spread and extreme residual values for EDUC greater than 15 years.
This suggests the need for alternative modeling strategies, such as weighted least squares or exploring a quadratic relationship between WAGE and EDUC.

d.

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

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

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

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

Female
- Regression equation: WAGE = -16.60 + 2.66 * EDUC
- Marginal return on education: 2.66
- R-squared: 27.64%

Male
- Regression equation: WAGE = -8.28 + 2.38 * EDUC
- Marginal return on education: 2.38
- R-squared: 19.27%

Black
- Regression equation: WAGE = -6.25 + 1.92 * EDUC
- Marginal return on education: 1.92
- R-squared: 18.46%

White
- Regression equation: WAGE = -10.48 + 2.42 * EDUC
- Marginal return on education: 2.42
- R-squared: 20.72%

Conclusion
Education has the largest impact on female wages and the smallest impact on black workers. The model explains wage variation best among females. These disparities may reflect structural inequalities and different returns to education across demographic groups.

e.

```
Call:
lm(formula = cps5_small$wage ~ x2)

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 ***
x2          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
```

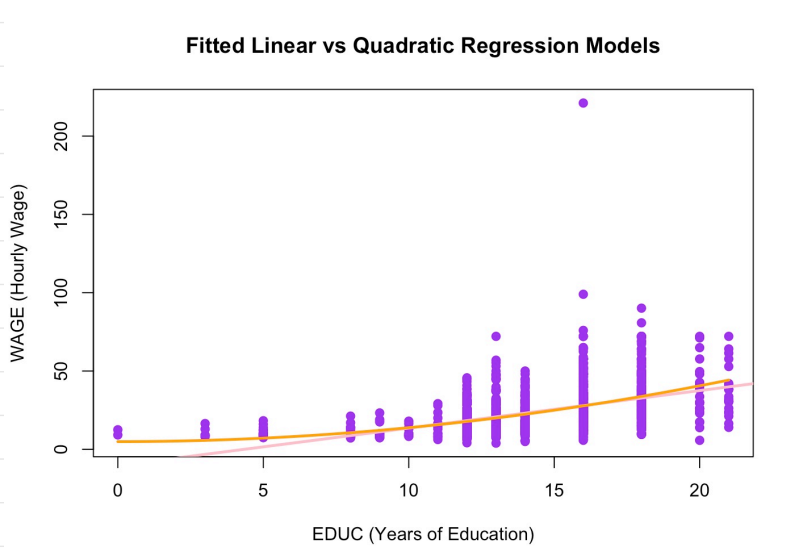
The quadratic model shows that as the years of education increase, the marginal return on wages also increases.
The marginal effect of education on wage is: - 2.14 dollars at 12 years of education - 2.85 dollars at 16 years of education Compared to the linear regression model from part (b), where the marginal effect was constant at 2.40 dollars, the quadratic model suggests increasing returns to education. The marginal effect increases as the years of education rise. Furthermore, the quadratic model has a slightly higher R-squared (0.2194) than the linear model (0.2073), indicating a better fit to the data.

$$\widehat{WAGE} = 4.916477 + 0.089134\, EDUC^2 + e$$

$$\frac{\partial E(\widehat{WAGE} \mid EDUC = 12)}{\partial EDUC} = 2 \cdot 0.089134 \cdot 12 = 2.139216$$

$$\frac{\partial E(\widehat{WAGE} \mid EDUC = 16)}{\partial EDUC} = 2 \cdot 0.089134 \cdot 16 = 2.852288$$

f.



From the graph, it is evident that the quadratic regression model (the pink line) fits the data better than the linear model in the range of higher years of education.
The quadratic regression model improves upon the linear model's underestimation of wages for individuals with higher education levels, reflecting the phenomenon of increasing marginal returns to education.
Although the R-squared values for both models are not high (both are below 25%), the quadratic model still performs slightly better and provides a better fit to the data.
Conclusion:
The quadratic model appears to fit the data better than the linear model.
If a more complex nonlinear relationship between education and wages needs to be captured, the quadratic model should be used.