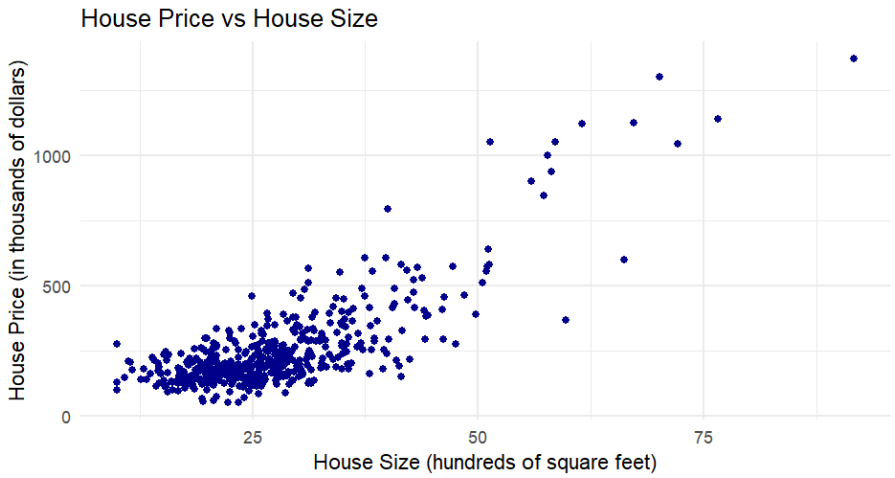
HW0303_Pinyo 312712017

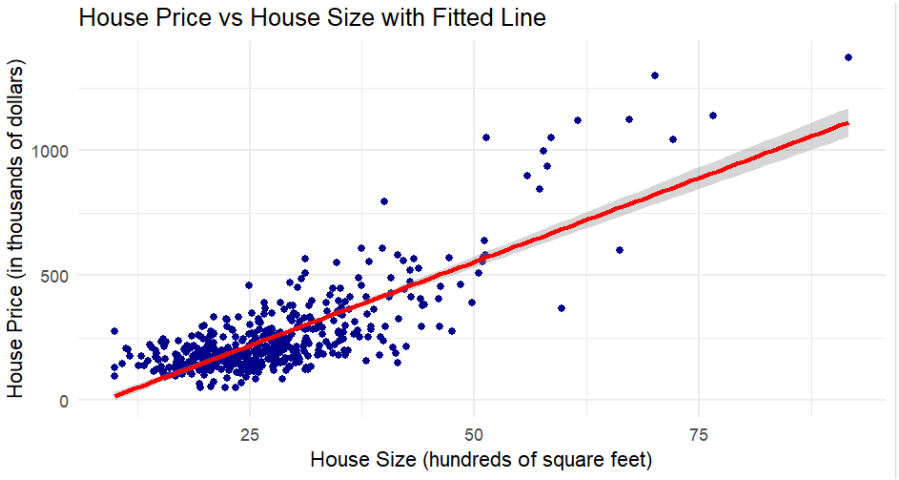
HW0303Q1 (C02Q17)

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*.

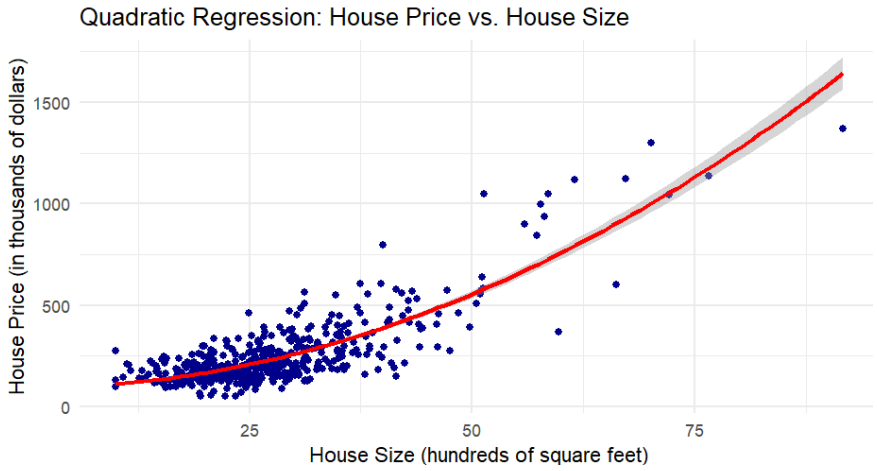
a.



b.

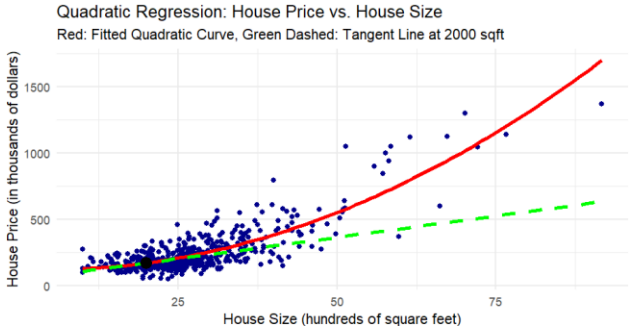


c.




```
> # Print the marginal effect
> cat("Marginal Effect at 2000 sqft:", marginal_effect, "thousands of dollars per 100 sqft\n")
Marginal Effect at 2000 sqft: 738.076 thousands of dollars per 100 sqft
```

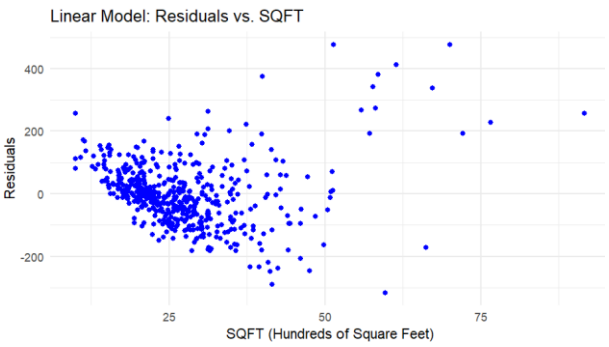
d.

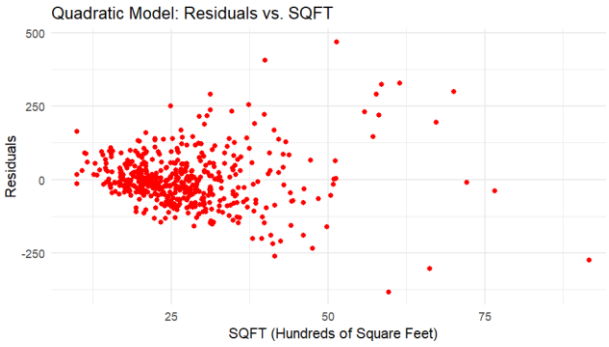


e.

```
>   # Output the elasticity value
>   elasticity
I(sqft^2)
0.8819511
```

f.





Checking Assumptions:

- If the residuals show **patterns** (e.g., a funnel shape, increasing/decreasing spread), this could suggest **heteroscedasticity** (violation of constant variance assumption).
- If the residuals show **curvature** or **non-random patterns**, this could indicate that the **model does not fit well** and assumptions about the functional form may be violated.
- Randomly scattered residuals without any pattern would suggest that the model fits the data well, and the assumptions are likely valid.

From the result, both models provide the pattern of residual, it means that the result violates the constant variance assumption.

g.

```
SSE for the linear model:  5262847
>   cat("SSE for the quadratic model: ", SSE_quadratic, "\n")
SSE for the quadratic model:  4222356
>
>   # Compare the SSE values
>   if(SSE_linear < SSE_quadratic) {
+     cat("The linear model has a lower SSE and fits the data better.")
+   } else if(SSE_quadratic < SSE_linear) {
+     cat("The quadratic model has a lower SSE and fits the data better.")
+   } else {
+     cat("Both models have the same SSE.")
+   }
The quadratic model has a lower SSE and fits the data better.
```

How Lower SSE Indicates a Better-Fitting Model:

- The **Sum of Squared Errors (SSE)** is a measure of the total squared difference between the observed values and the predicted values. A lower SSE indicates that the model's predictions are closer to the actual observed values.
- In general, a model with a lower SSE is considered a **better fit**, as it minimizes the discrepancy between the predicted and actual values of the dependent variable.

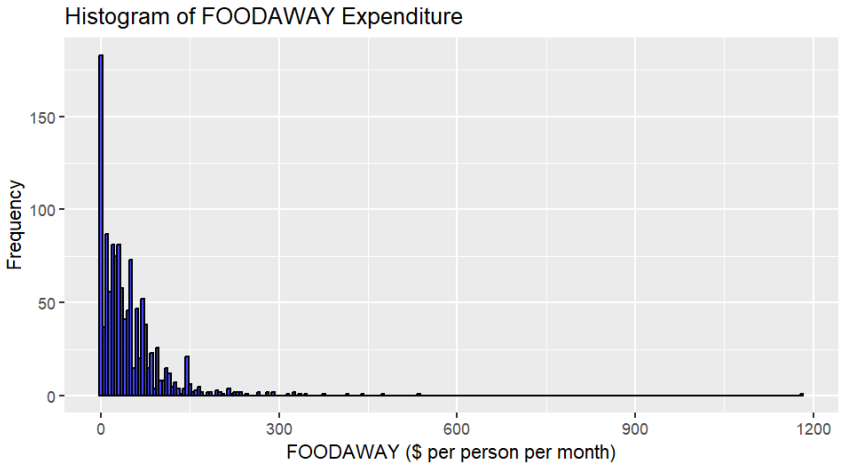
HW0303Q1 (C02Q25)

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.

```
>   # Output summary statistics
>   print(summary_stats)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.04   32.55   49.27   67.50 1179.00
>   cat("\nMean of FOODAWAY:", mean_foodaway, "\n")

Mean of FOODAWAY: 49.27085
>   cat("Median of FOODAWAY:", median_foodaway, "\n")
Median of FOODAWAY: 32.555
>   cat("25th percentile of FOODAWAY:", percentiles[1], "\n")
25th percentile of FOODAWAY: 12.04
>   cat("75th percentile of FOODAWAY:", percentiles[2], "\n")
75th percentile of FOODAWAY: 67.5025
```



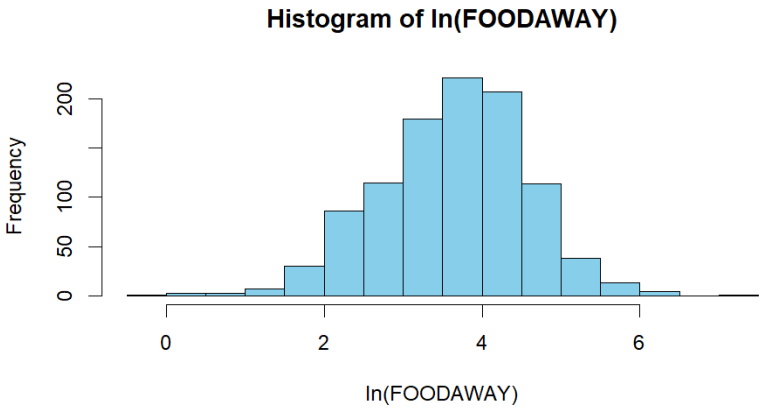
b.

```
Mean and Median of FOODAWAY for households with an Advanced Degree member:
>   cat("Mean:", mean_advanced, "\n")
Mean: 73.15494
>   cat("Median:", median_advanced, "\n\n")
Median: 48.15

>
>   cat("Mean and Median of FOODAWAY for households with a College Degree member:\n")
Mean and Median of FOODAWAY for households with a College Degree member:
>   cat("Mean:", mean_college, "\n")
Mean: 48.59718
>   cat("Median:", median_college, "\n\n")
Median: 36.11

>
>   cat("Mean and Median of FOODAWAY for households with no advanced or college degree membe
r:\n")
Mean and Median of FOODAWAY for households with no advanced or college degree member:
>   cat("Mean:", mean_no_degree, "\n")
Mean: 39.01017
>   cat("Median:", median_no_degree, "\n")
Median: 26.02
```

c.



ln(foodaway) has the number of observations less than foodaway because it does not count the data value equal to zero

d.

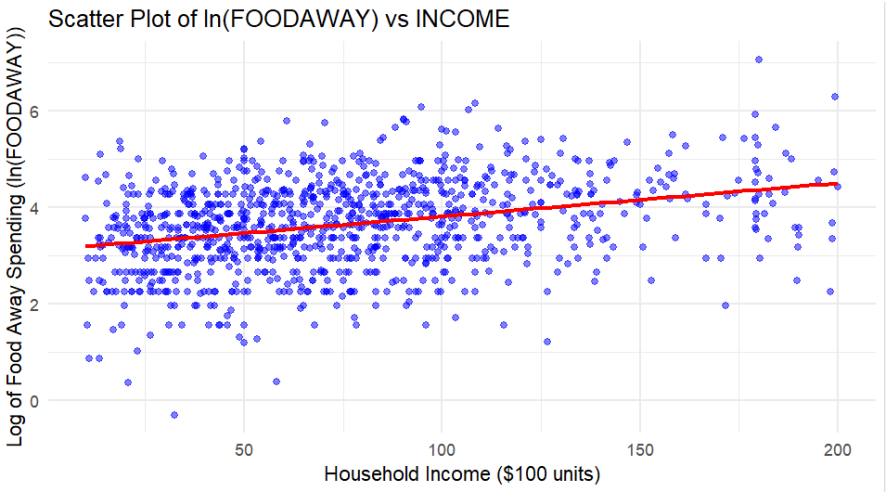
Slope (0.0069)

- The **coefficient of INCOME (0.0069)** means that for every **$100 increase in monthly income**, the **log of food away from home spending increases by 0.0069**.
- In percentage terms, since this is a log-linear model:

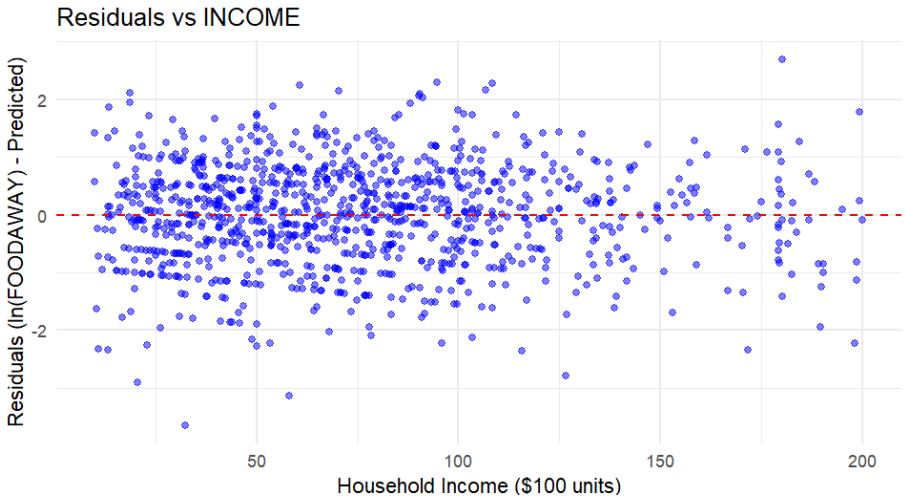
100×(e^0.0069−1)≈0.69%100 \times (e^{0.0069} - 1) \approx 0.69\%100×(e0.0069−1)≈0.69%

So, for every **$100 increase in income, food away from home spending increases by approximately 0.69%**.

e.



f.

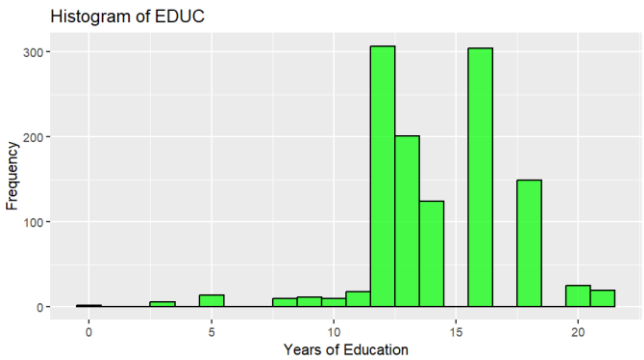


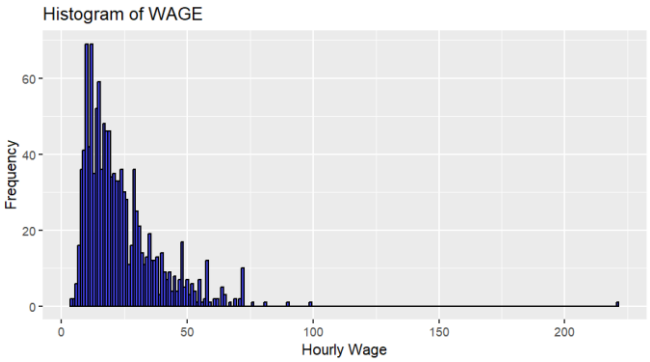
There is a pattern that the residuals seem to spread more when the household income more than 10,000$. This means standard errors of the regression coefficients might be unreliable, potentially leading to incorrect inferences.

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.

```
>
>   # 3. Summary statistics for wage and educ
>   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
```





b.

```
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
```

Result Interpretation

1. Regression Equation

The estimated regression equation is:

wage^=−10.40+2.40×educ\hat{\text{wage}} = -10.40 + 2.40 \times \text{educ}wage^ =−10.40+2.40×educ

- **Intercept (-10.40):** This means that if a person has **zero years of education**, their predicted wage would be -10.40. However, this doesn't make practical sense because education years cannot be zero in this context.
- **Slope (2.40):** Each additional year of education is associated with an **increase of 2.40 in hourly wage**.

2. Statistical Significance

- The **p-value for EDUC (< 2e-16)** is extremely small, indicating that education has a **highly significant** impact on wage.
- The **t-value for EDUC (17.7)** is very large, reinforcing that education is strongly related to wage.

3. Model Fit

- **R-squared = 0.2073 (20.73%)**
    - This means that about **20.73% of the variation in wage is explained by education**.
    - This is relatively low, suggesting that other factors also influence wages.
- **Adjusted R-squared = 0.2067 (20.67%)**

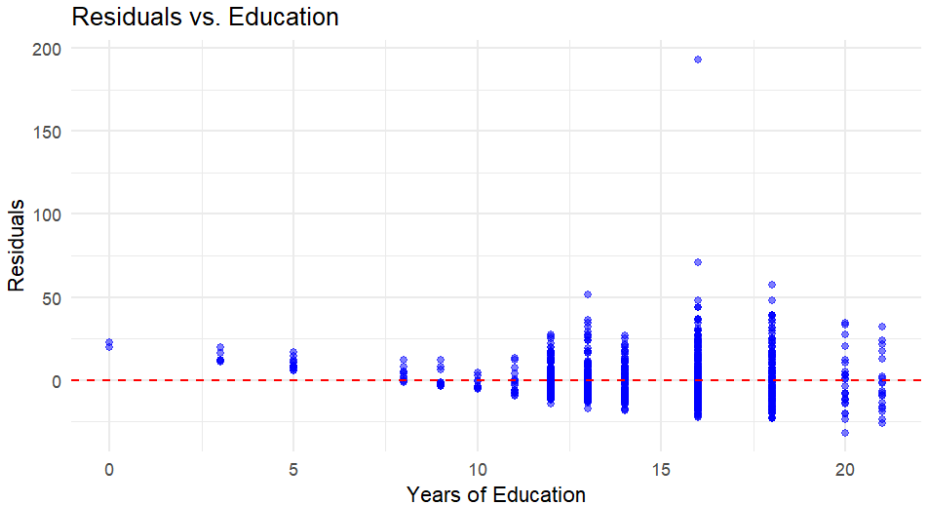
- This is slightly lower than R-squared, adjusting for the number of predictors in the model.

- **Residual Standard Error = 13.55**
    - The average deviation of actual wages from the predicted wages is about **$13.55 per hour**.

4. Overall Model Significance

- **F-statistic = 313.3, p-value < 2.2e-16**
    - The model is **statistically significant** overall, meaning that education contributes to predicting wages

c.



There is a pattern which the residuals will be expanded wider when year of education increases.

If the assumptions **SR1–SR5** (Standard Regression Assumptions) hold, the residuals should be **randomly scattered around zero** with no clear pattern.

d.

```
> summary(male_model)

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

```
> summary(female_model)

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

Beta-1 and Beta-2

Male: Beta-1 = -8.2849, Beta-2 = 2.3785

Female: Beta-1 = -16.6028, Beta-2 = 2.6595

P-value

Male and Female are both significant to wage

RSE

Male: 14.71

Female: 11.5

Female has lower RSE

R-square and Adjusted R-square

Male: 0.1927, 0.1915

Female: 0.2764, 0.275

Female can explain the wage more than Male

```
> summary(black_model)

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

```
> summary(white_model)

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

Beta-1 and Beta-2

Black: Beta-1 = -6.2541, Beta-2 = 1.9233

White: Beta-1 = -10.475, Beta-2 = 2.418

P-value

Black and White parameter are both significant to wage

RSE

Black: 10.51

White: 13.79

Black model has lower RSE than White model

R-square and Adjusted R-square

Black: 0.1846, 0.1767

White: 0.2072, 0.2065

White can explain the wage more than Black

e.

```
>   # Print results
>   cat("Marginal Effect at EDUC = 12:", marginal_12, "\n")
Marginal Effect at EDUC = 12: 2.074698
>   cat("Marginal Effect at EDUC = 16:", marginal_16, "\n")
Marginal Effect at EDUC = 16: 2.909434
>   cat("Linear Regression Marginal Effect:", linear_effect, "\n")
Linear Regression Marginal Effect: 2.396761
```

The **quadratic model suggests that the effect of education increases** at higher education levels.

- At 12 years → Wage increase is **2.07**
- At 16 years → Wage increase is **2.91**

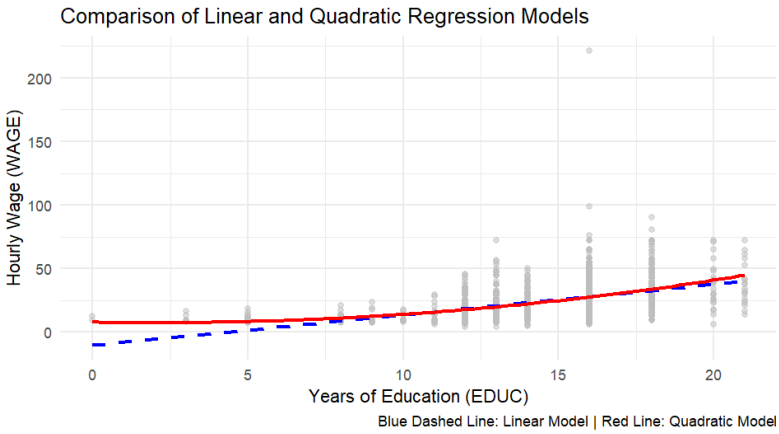
The **linear model assumes a fixed increase of 2.40**, which does not capture the variation seen in the quadratic model.

Since the quadratic model allows for **non-linear changes**, it suggests that **higher education might lead to greater wage increases** (increasing returns)

If **education truly has increasing returns**, policymakers and individuals should **invest more in higher education** as it leads to larger wage gains.

If the **quadratic term is statistically insignificant**, then the **linear model is sufficient**, and education has a **constant return** on wages.

f.



Quadratic model (red line) tends to fit the data better