Name: Nguyen Quoc Nhan (Tom)

ID: 413707009

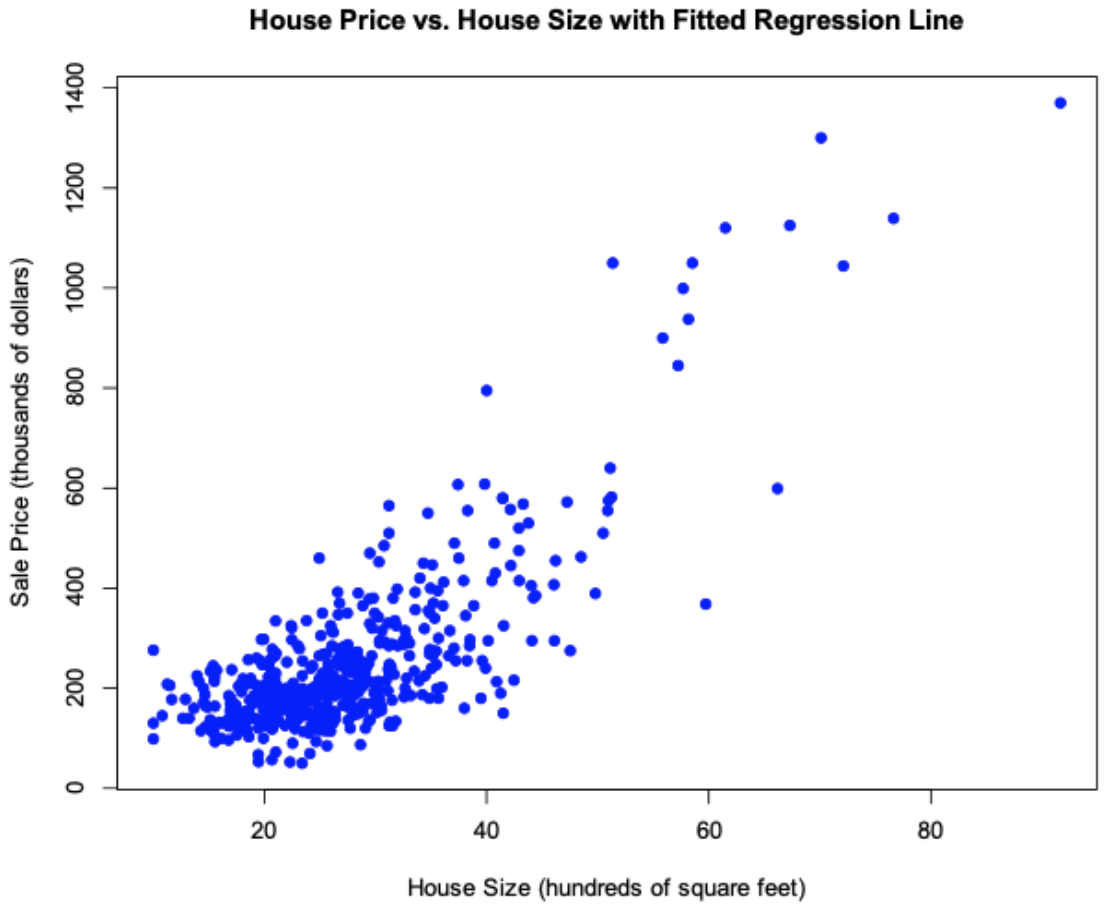
Department: DIF

Email: nqnnguyenquocnhan@gmail.com

HW0303

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.

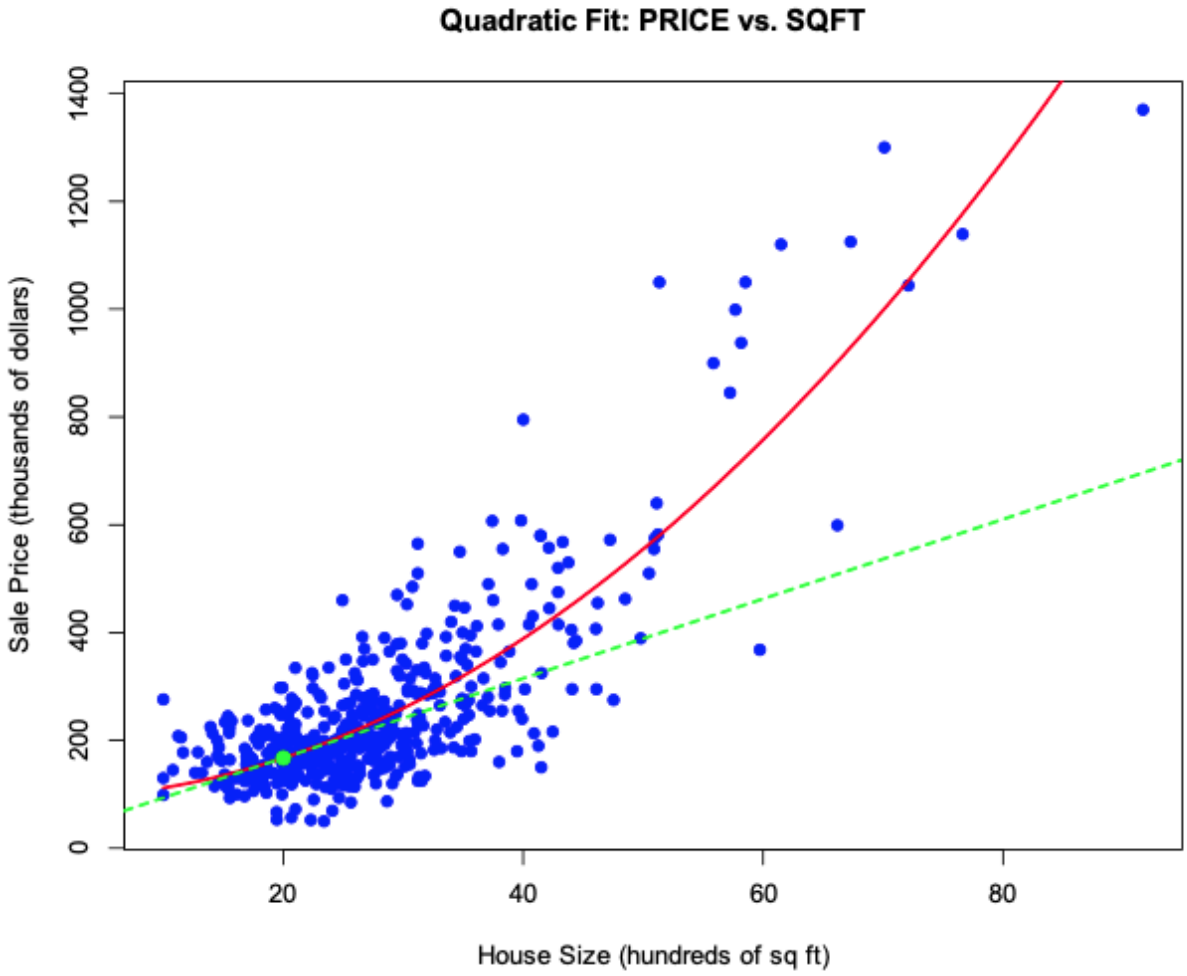


b. Estimate the linear regression model PRICE = β1 + β2SQFT + e. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic regression model PRICE = α1 + α2 $SQFT^2$ + e. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

The marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space is $7,380.76

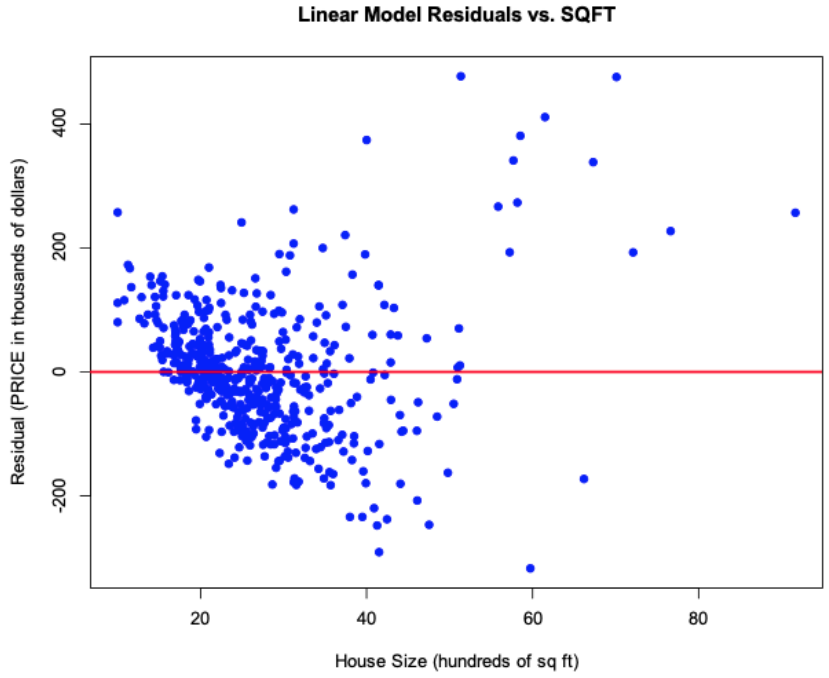
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

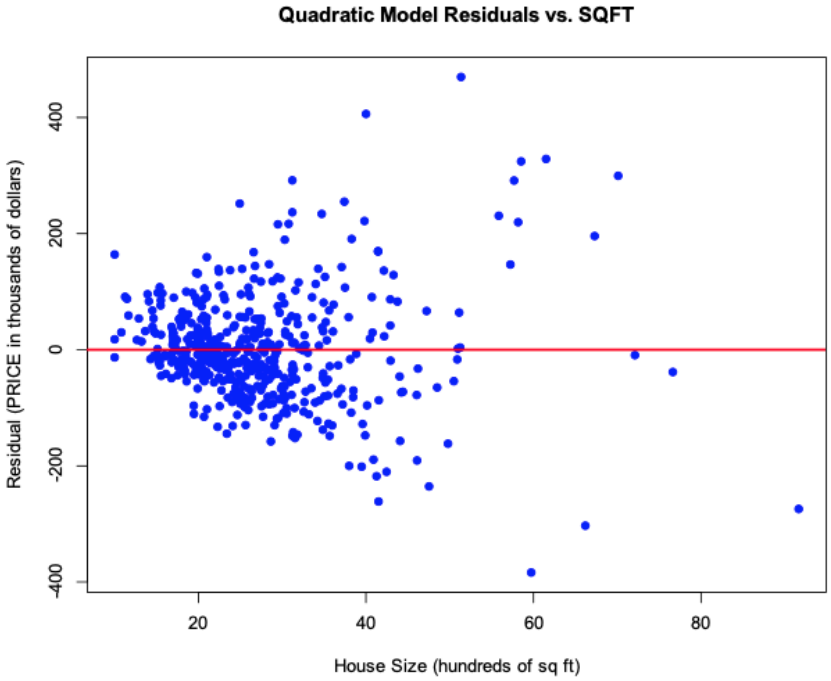


e. For the model in part (c), compute the elasticity of PRICE with respect to SQFT for a home with 2000 square feet of living space.

The elasticity of PRICE with respect to SQFT for a home with 2000 square feet of living space is 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?





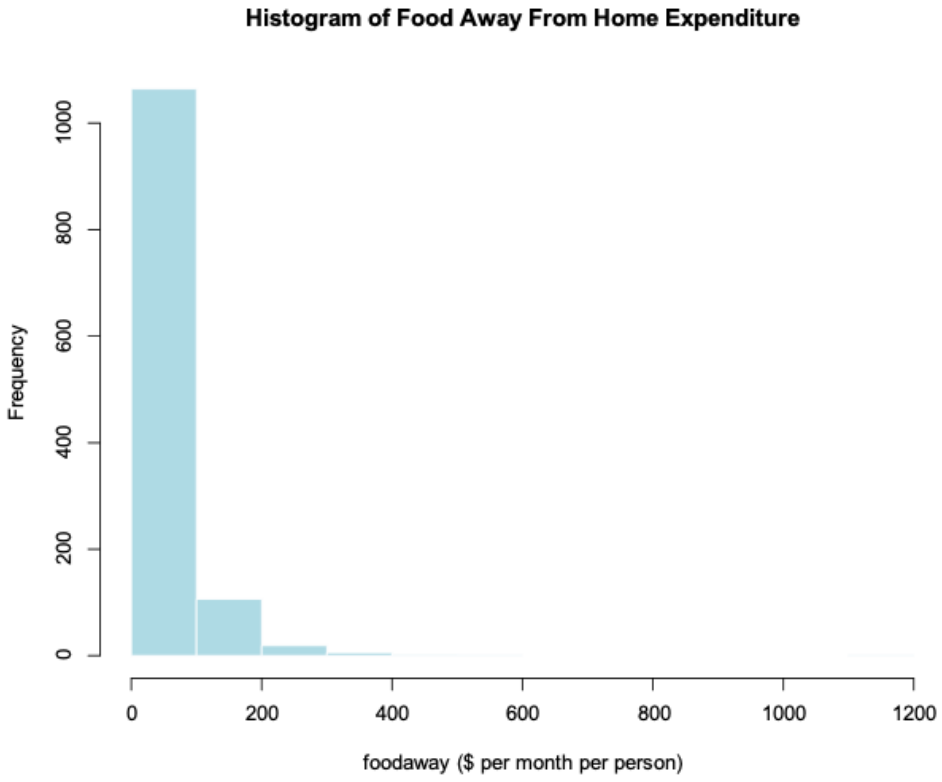
Upon reviewing the residual plots for both models, the patterns do not appear to be random. Moreover, the spread of the residuals grows as SQFT increases, indicating that the assumption of homoskedasticity may be breached.

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 of linear model is 5,262,847 and SSE of quadratic model is 4,222,356. With the SSE becomes smaller in quadratic model, resulting in quadratic model is a "better-fitting" model. Since the quadratic model has the lower SSE, the data points fit more closely to the curve of the quadratic model than to the straight line of 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?



Mean = 49.27085; Median = 32.555

```
> summary(cex5_small$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.04   32.55   49.27   67.50 1179.00
```

The 25th percentiles = 12.04; 75th percentiles = 67.5025

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?

With an advanced degree:

- Mean = 73.15
- Median = 48.15

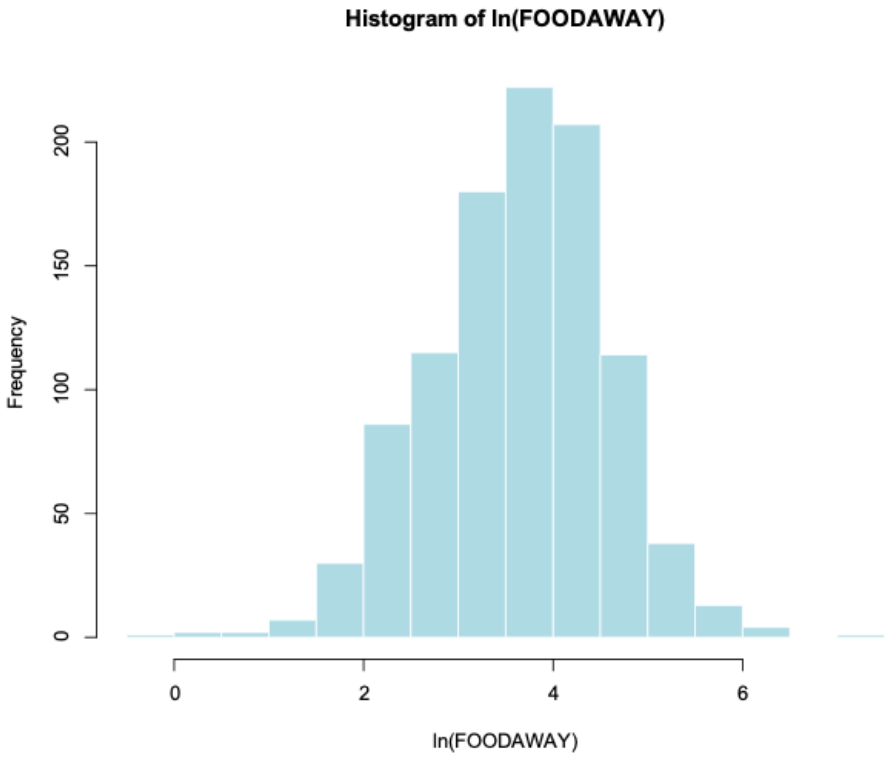
With a college degree member:

- Mean = 48.59
- Median = 36.11

With no advanced or college degree member:

- Mean = 0
- Median = 26.02

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$ln_foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -Inf   2.488   3.483    -Inf   4.212   7.072
```

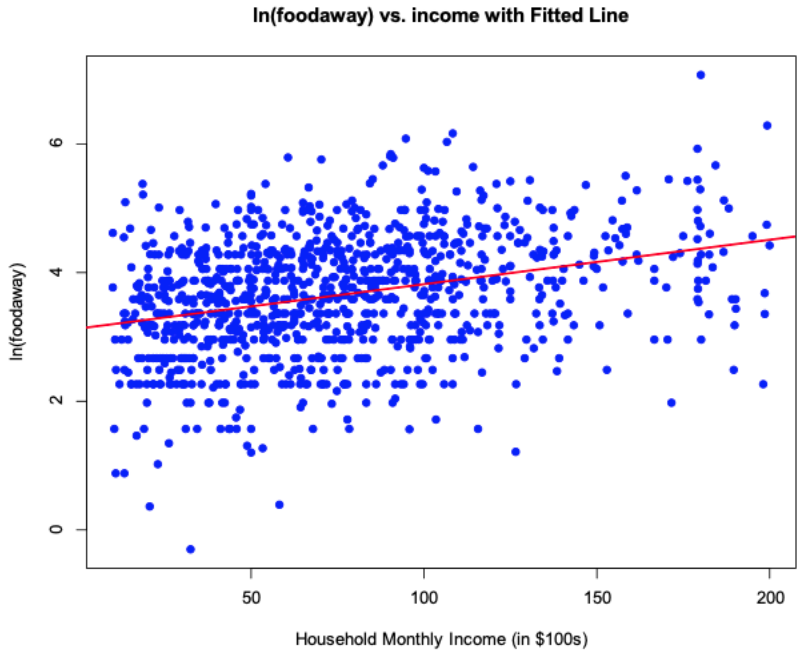
It is different because *ln_foodaway* requires the value larger than zero, while *foodaway* does not consider that situation.

d. Estimate the linear regression ln(FOODAWAY) = β1 + β2INCOME + e. Interpret the estimated slope.

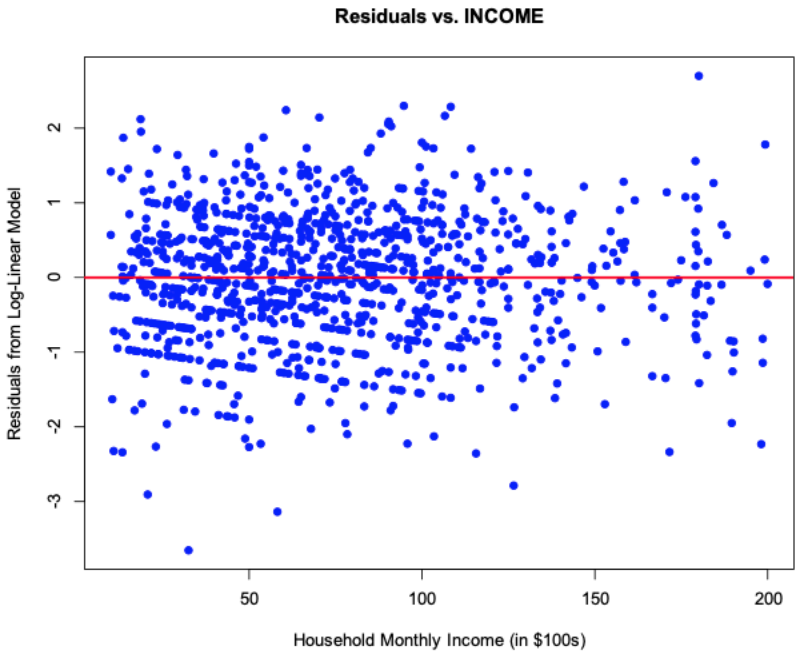
An increase of $100 of income is associated with approximately increase 0.6% of food away.

```
lm(formula = ln_foodaway ~ income, data = cex5_small_clean)

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
```

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?

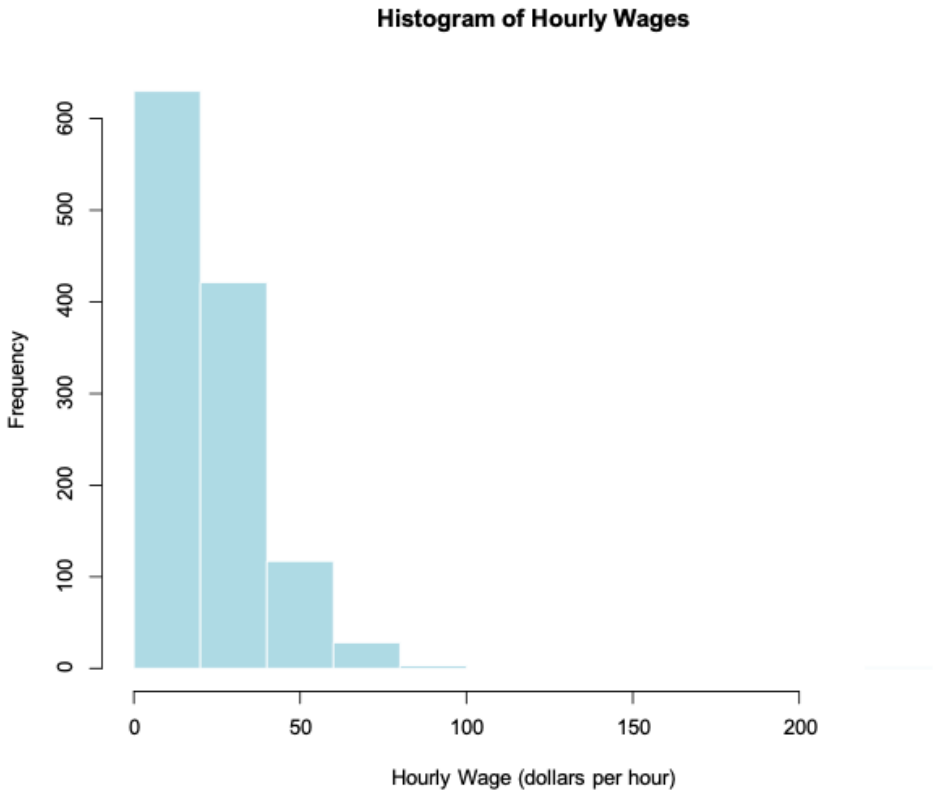


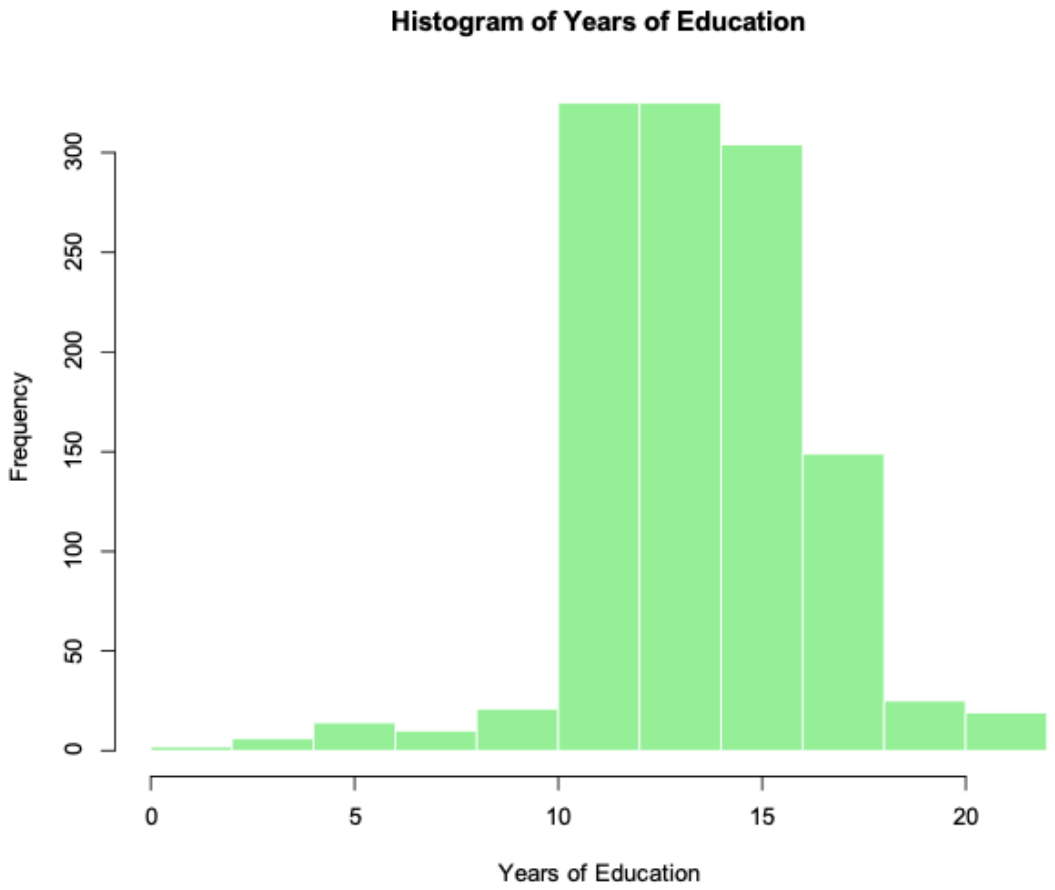
It looks random without any patterns.

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.

```
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 statistics for EDUC
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 wage distribution is notably skewed to the right, with the bulk of observations under $30 and a handful of extreme outliers pushing the upper bound to $221.10. Reflecting this skew, the median wage is around $19.30, whereas the mean sits at roughly $23.64. In contrast, years of education mostly fall between 12 and 16, forming a unimodal distribution near typical schooling durations, with very few cases at the extreme lower or upper ends.

Given these patterns, wages may benefit from a log transformation to address their skew, while education levels appear suitable for standard analytical methods.

b. Estimate the linear regression WAGE = β1 + β2EDUC + e and discuss the results.

```
> summary(model_lin)

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
```

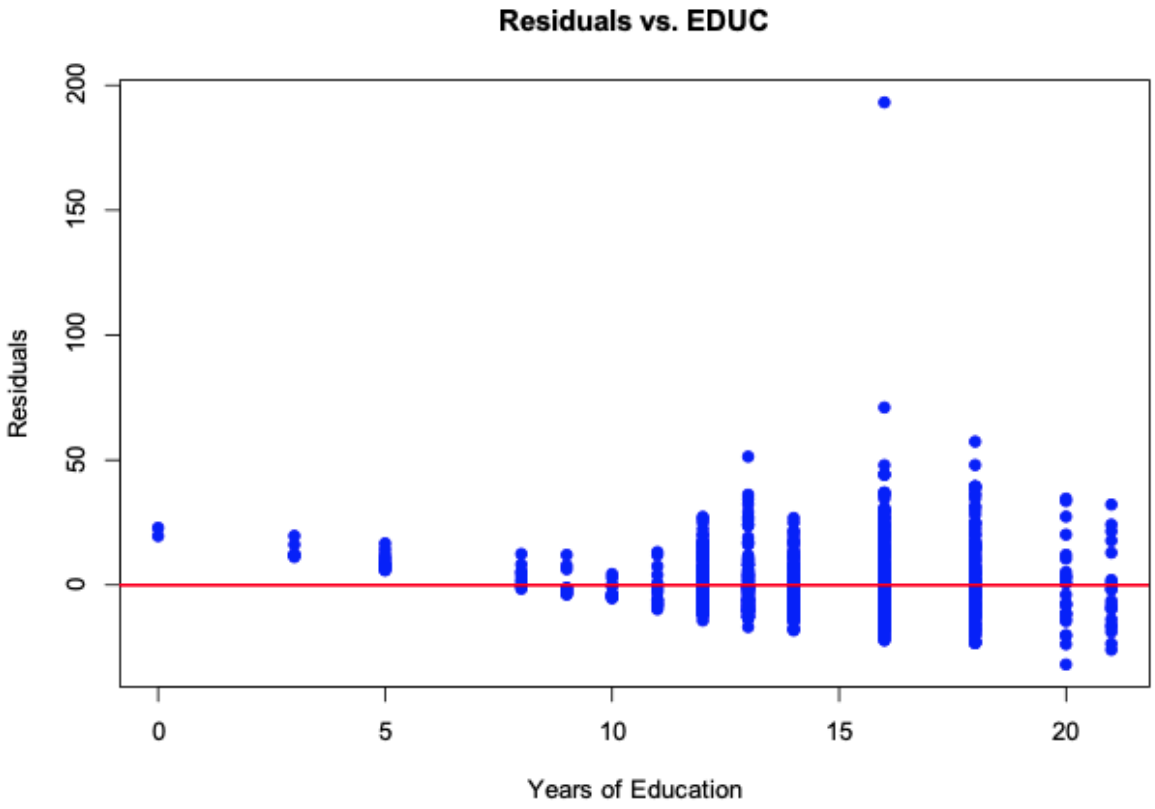
According to this model, each extra year of education corresponds to a $2.40 increase in hourly wages, assuming all other factors remain the same. The intercept of -10.40 literally indicates that, if someone had zero years of schooling (EDUC = 0), the model would predict an hourly wage of -$10.40.

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 residuals plotted against EDUC do not display a clear pattern, such as a U-shape, indicating that using a linear form in EDUC does not appear to be fundamentally

incorrect. The residuals mostly hover around zero, although a few high positive values—particularly near EDUC = 15 and 20—suggest certain individuals earn more than the model predicts. There is a slight uptick in residual spread at higher education levels, implying mild heteroskedasticity but not at a severe level.

Overall, the assumptions of strict exogeneity, linearity, and uncorrelated errors seem reasonably upheld, and EDUC shows enough variation. The residuals form a relatively random scatter around zero, which aligns with classical OLS assumptions (SR1–SR5) and suggests no major violations in adopting a linear model.



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

Separate Regressions by Gender and Race

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

When wage is regressed on education across various demographic groups, notable differences emerge in the returns to schooling. Women see the largest wage gain per additional year of education ($2.66), followed by men ($2.53) and whites ($2.42), while Black individuals receive the smallest return ($1.92).

Though the intercepts are negative—an unrealistic value for zero years of schooling—they simply represent the regression line extrapolated to that point. In terms of explanatory power, the female subsample has the highest R-squared (0.276), indicating that education alone explains a greater portion of wage variation for women. Conversely, Black individuals exhibit both the lowest return on education and the lowest R-squared (0.185), suggesting other factors heavily influence their wages.

Overall, these results illustrate that the linear education–wage relationship varies by demographic group, with women gaining the most from extra schooling, Black individuals the least, and men and whites in between.

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).

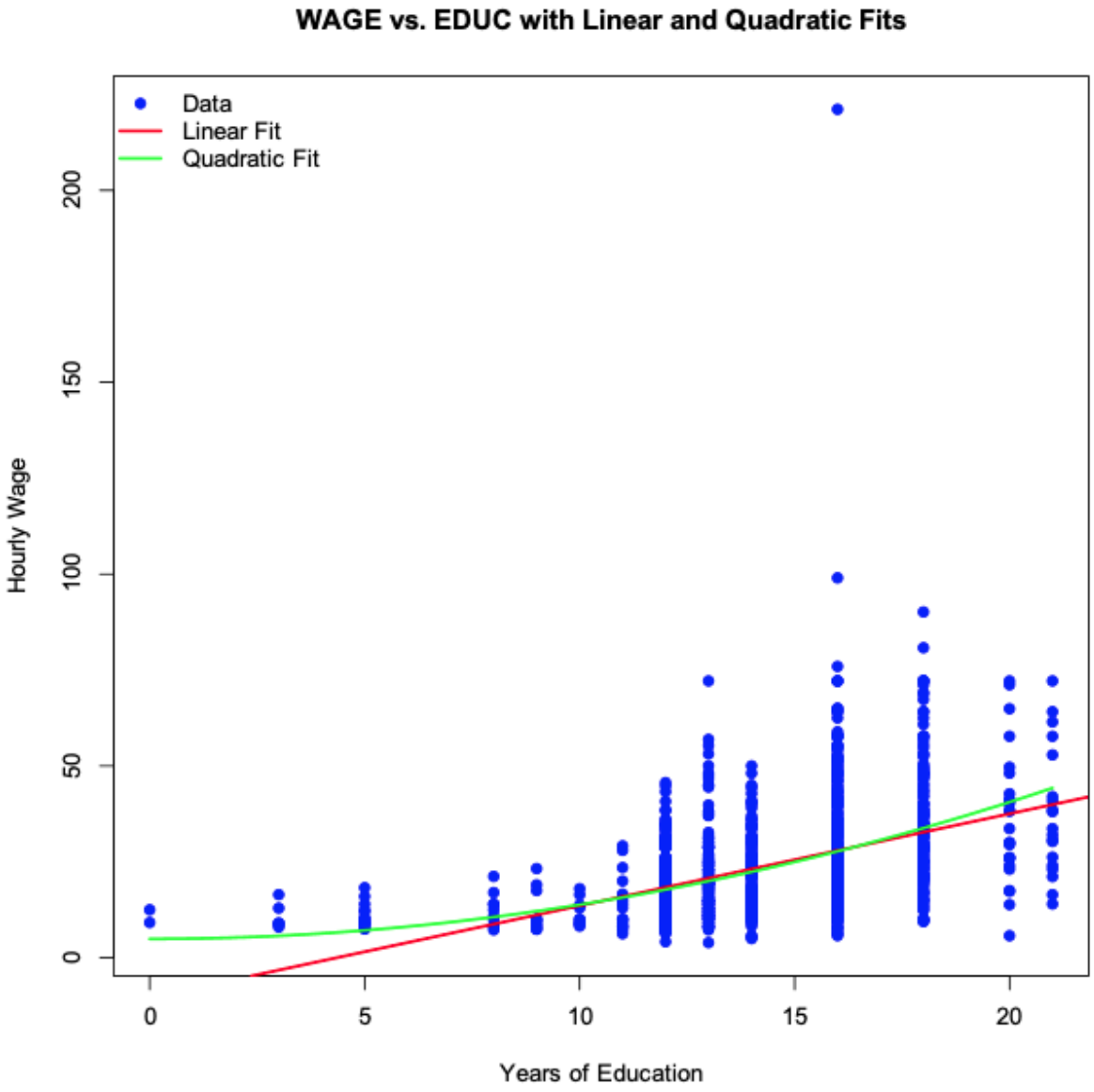
```
Call:
lm(formula = wage ~ I(educ^2), data = cps5_small)

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 ***
I(educ^2)   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
```

Marginal Effect at 12 years of education: 2.139216

Marginal Effect at 16 years of education: 2.852288

Under the quadratic model, each extra year of schooling adds more to one’s wage than the previous year—moving from 12 to 13 years of education raises wages by about $2.14 per hour, whereas going from 16 to 17 years yields around $2.85. By contrast, the linear model assumes a constant marginal effect of $2.40 per hour for every additional year of education, regardless of one’s starting point.

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?



With the linear model, the adjusted R-squared is 0.2067, whereas for the quadratic model it is 0.2187. This indicates that the quadratic specification offers a marginally better fit, as reflected in both the visual plots and the regression results:

Visual Fit: The quadratic model sidesteps the implausible negative wage predictions for individuals with very low schooling and better captures the more pronounced wage gains at higher education levels (15–20 years).

Statistical Indicators: The quadratic model achieves a slightly higher R-squared and a lower sum of squared errors (SSE), suggesting it explains a bit more variation in wages.

Flexibility: Although the improvement is modest, the quadratic approach more flexibly accommodates accelerating wage growth with additional education, instead of imposing a constant slope as in the linear model.