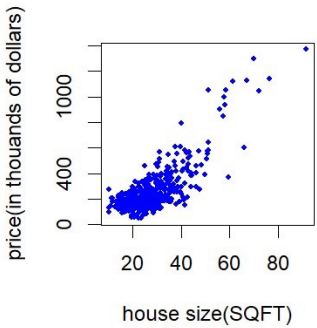
2.17

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.

a.



b.

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.

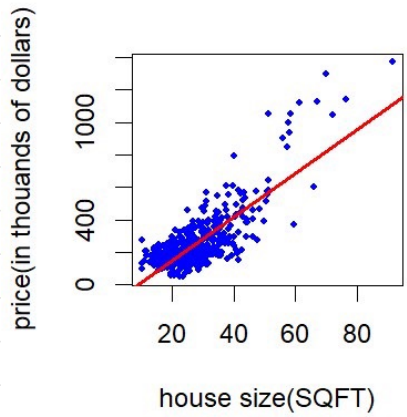
f. For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?

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?

$\beta_1 = -115.4236$, is an estimated value of y (price) when x (SQFT) = 0.

$\beta_2 = 13.4029$ means that for each unit increase in x (SQFT), y (price) is expected to increase by 13.4029 (in thounds of dollars), holding all other variable constant.

fitted line:



c. marginal effect:

$$\frac{\partial price}{\partial (SQFT)} = 2 \cdot \alpha_2 \cdot SQFT$$

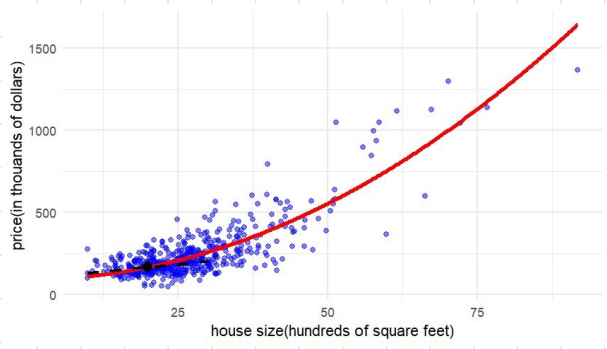
$\alpha_1 = 93.565854$

$\alpha_2 = 0.184519$

marginal effect $= 2 \cdot 0.184519 \cdot SQFT$

$= 0.369038 \cdot SQFT$ (in thousands of dollars)

d.



e.

$$\text{elasticity} = \frac{\partial price}{\partial SQFT} \times \frac{SQFT}{price}$$

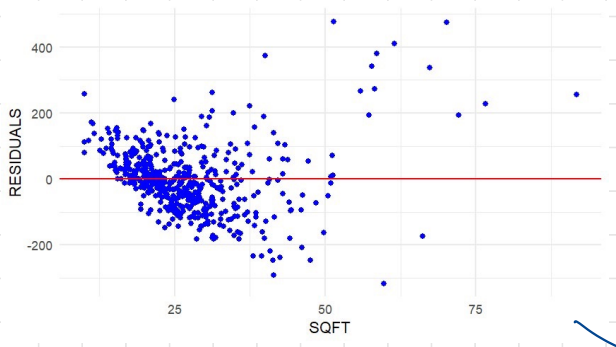
$$\frac{\partial price}{\partial SQFT} = 7.38076, \quad SQFT = 20,$$

$$price = 93.565854 + 0.184519 \times 20^2 = 165.565854$$

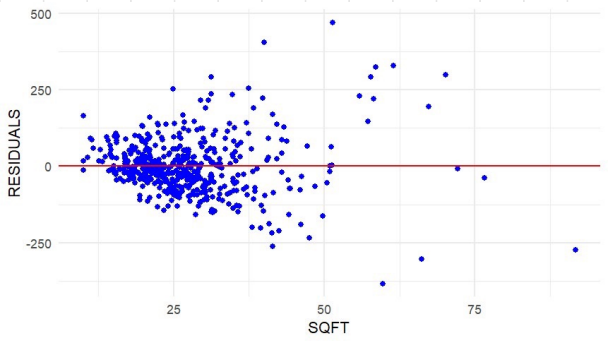
$$\text{elasticity} = 7.38076 \times \frac{20}{165.565854} \approx 0.8916$$

f.

Linear:



Quad:



the variance of redisuduals increases as SQFT increases, suggest heteroskedasticity. and it violate the OLS assumptions of homoskedasticity.

g

SSE linear = 5762847

SSE quad = 4222356

quad model have lower SSE, and a lower SSE model indicate a better-fitting model because it means that the model's predicted values are closer to the actual data, and it alse can better representation of data trends and reducing unexplained variability.

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?

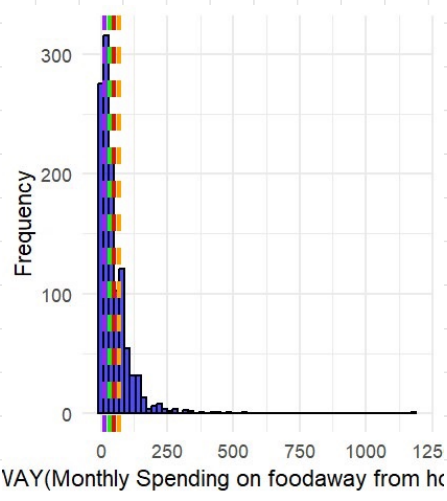
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

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

25th percentiles: 12.04

75th percentiles = 67.5025

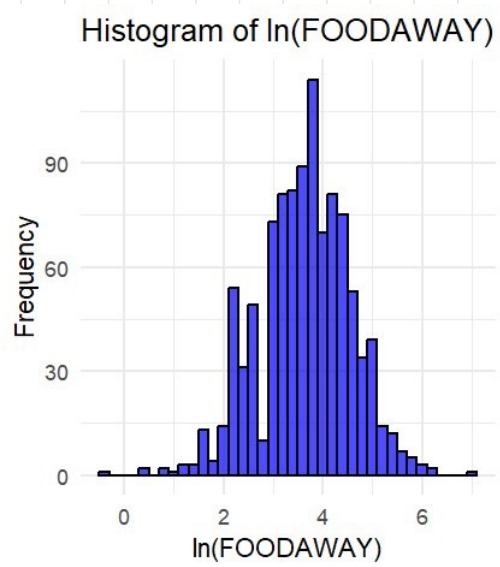
b.

advanced degree: mean = 73.15
median = 48.15

college degree: mean = 48.60
median = 36.11

no advanced or college degree: mean = 39.01
median = 26.05

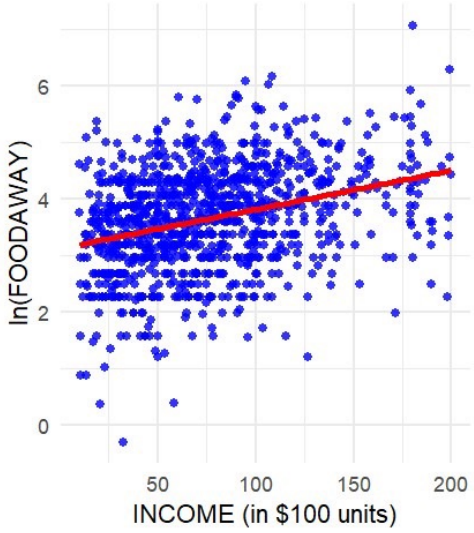
c.



FOODAWAY and ln(FOODAWAY) have different number of observation is because. natural logarithm can only be calculated for positive values, some household in the dataset have FOODAWAY=0, meaning that they did not spent on FOODAWAY from home.

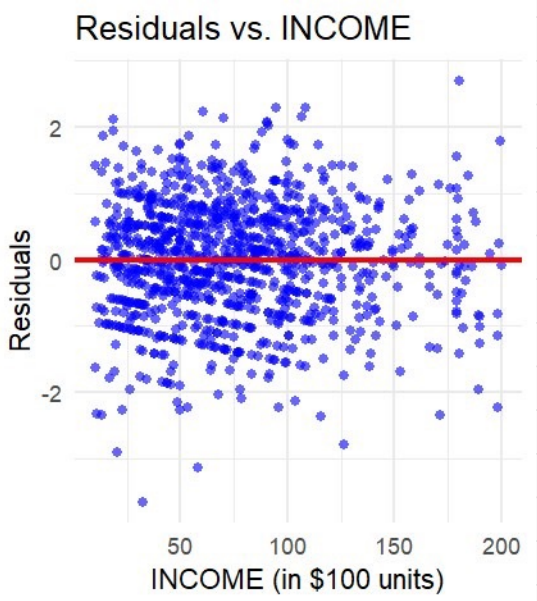
d. ln(FOODAWAY) = 3.1293 + 0.0069 INCOME.
When monthly income increase $100 units, food away from home expenditure will increase 0.69% per person, holding all other variable constant.

e.



ln(FOODAWAY) and Income have positive relation.

f.



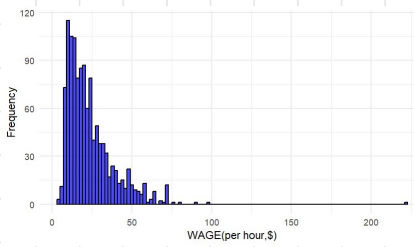
The model is generally well-fitted not violate OLS assumption.

2.28

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.



The majority of workers earn below $50 per hour,

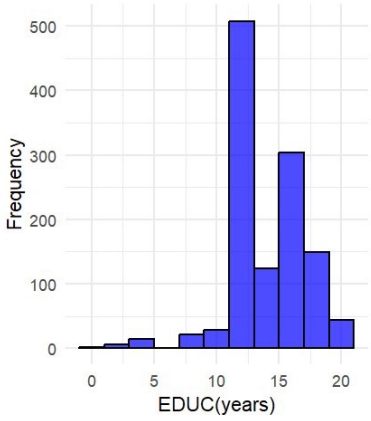
right - skewed

mean: 23.64 min = 3.94

median: 19.3 Max = 221.10

Q1 : 13

Q3 : 29.8



The distribution is clustered at 12 years - 16 years (high school - college degree)

mean: 14.20 min = 0

median: 14. Max = 21.

Q1 = 12

Q3 = 16

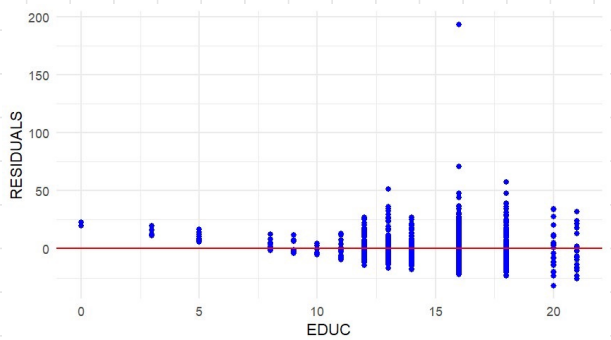
b.

$WAGE = -10.40 + 2.3968 \cdot EDUC + e.$

$\beta_1 = -10.40$ is an estimated value of y (WAGE) when x (EDUC) = 0

$\beta_2 = 2.3968$ means that for each unit increase in x (EDUC), y (WAGE) is expected to increase $2.3968 (per hour), holding all other variable constant.

c.



The spread of the residuals increases as EDUC increases, this suggest the presense of heteroskedasticity, violating homoskedasticity.

d.

male: WAGE = -8.2849 + 2.3785 EDUC + e

female: WAGE = -16.6028 + 2.6595 EDUC + e

White: WAGE = -10.495 + 2.418 EDUC + e

black: WAGE = -6.2541 + 1.9233 EDUC + e

The education coefficient represents the expected increase in wages (per hour) for each additional year of education.

Female (2.6595) > White (2.42) > Male (2.3785) > Black (1.9233)

Female > male means female experience high return on education than male, but female have lower baseline wage. means when education = 0, male wage > female.

White > black means white experience high return on education than black. but white have lower baseline wage. means when education = 0, black wage > white wage.

e.

WAGE = -4.916477 + 0.089134 $EDUC^2$ + e

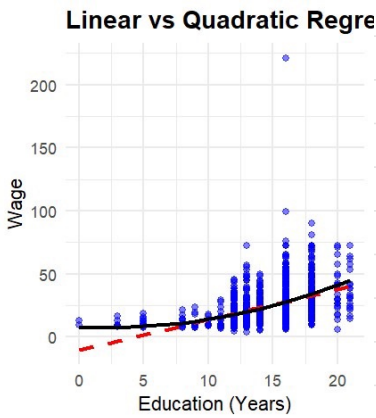
marginal effect for 12 years : 2 . 0.089134 × 12 = 2.139

16 years : 2 · 0.089134 × 16 = 2.852

in (b), marginal effect = β_2 = 2.3968.

in (e), marginal effect : 2. α_2. EDUC, it means in quadratic model, when EDUC increases, the marginal effect is expected to increase.

f.



the quadratic model appears to fit the data better than the linear model. because the linear model underestimates the wages increase for higher education levels and fails to capture the upward curvature.