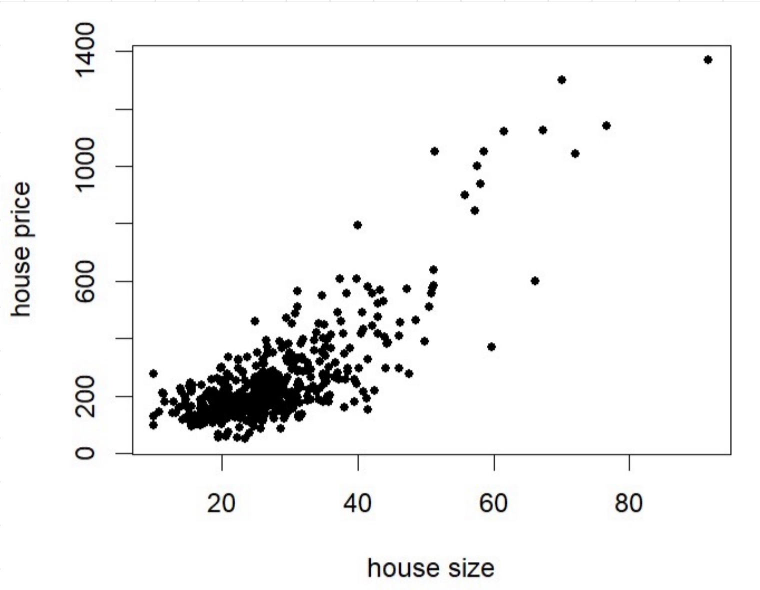
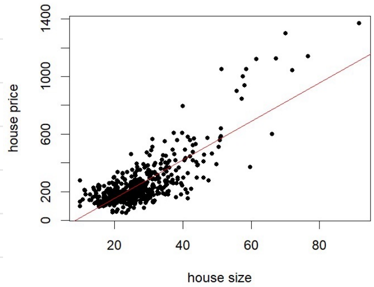
2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

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.

```
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 ***
```



$\widehat{PRICE} = -115.4236 + 13.4029 * SQFT$

Holding all else constant, in average, an additional 100 square feet of house size, house price will increase \$13402.
When a house with zero square feet, the expected price is \$-115423.6

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.

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

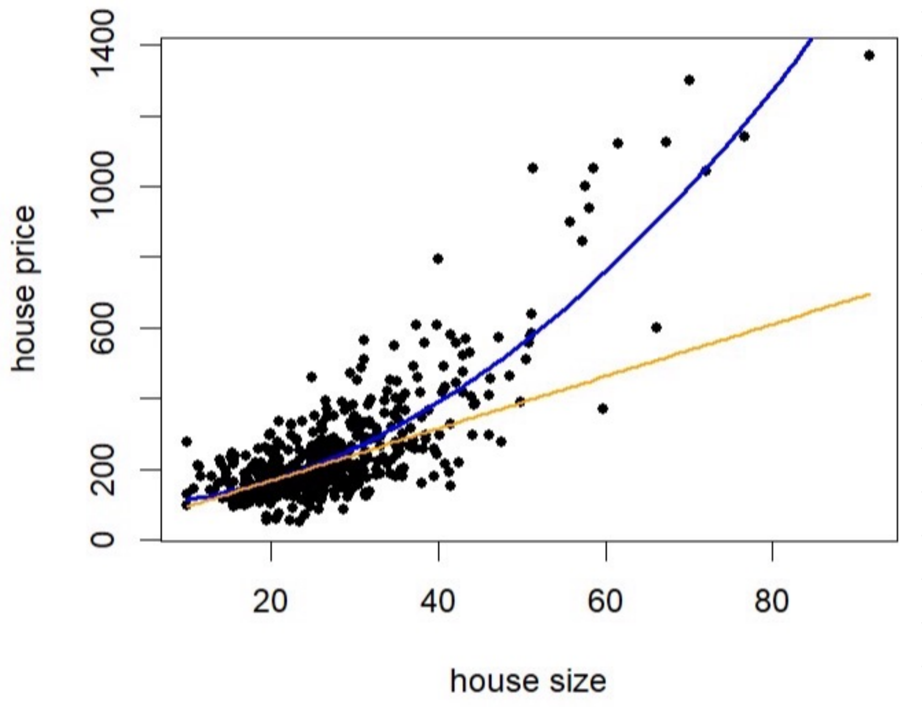
$\widehat{PRICE} = 93.5659 + 0.184519 * SQFT^2$

$$\frac{\partial PRICE}{\partial SQFT} = 2 \times 0.184519 \times SQFT$$

$$\left.\frac{\partial PRICE}{\partial SQFT}\right|_{SQFT=20} = 7.38076$$

marginal effect $= \$7380.76$

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.

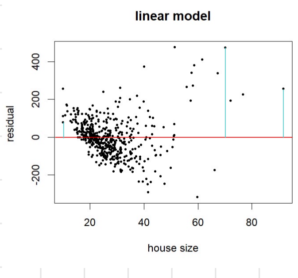


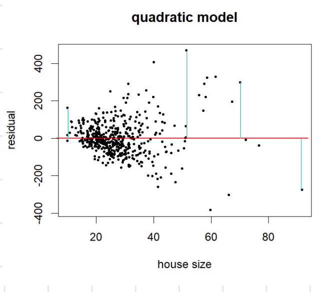
e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

$$\varepsilon = \frac{\partial PRICE}{\partial SQFT} \times \frac{SQFT}{PRICE} = 2 \times 0.184519 \times 20 \times \frac{20}{93.5659 + 0.184519 * 20^2}$$

$$\doteq 0.882$$

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?





SR 3: Conditional Homoskedasticity $Var(e_i | \mathbf{x}) = \sigma^2$

In both models, the residual patterns do not appear random. The variation in the residuals increases as SQFT increases, suggesting that the homoskedasticity assumption may be 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?

(b) SSE : 5262847
(c) SSE : 4222356

$SSE_b > SSE_c$

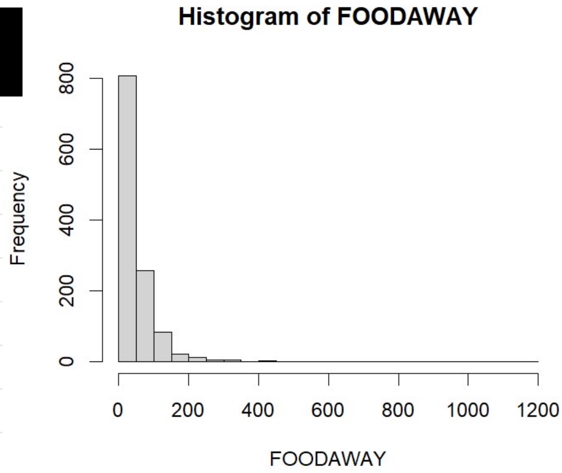
In this case the quadratic model has the lower SSE. The lower SSE means that the data values are closer to the fitted line for the quadratic model than for the linear .

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?

25th 75th

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00   12.04   32.55   49.27   67.50 1179.00
```



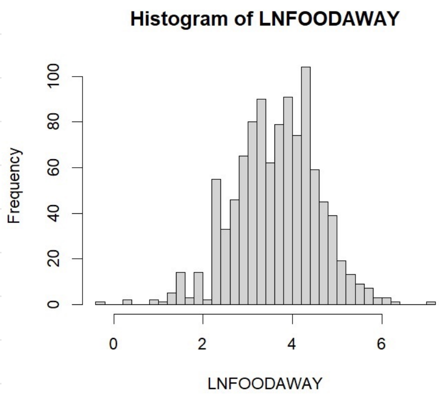
b. What are the mean and median values of *FOODAWAY* for households including a member with an advanced degree? With a college degree member? With no advanced or college degree member?

	Mean	Median
advanced	73.15	48.15
college	48.60	36.11
None	39.01	26.02

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

1200 − 1022 = 178

```
> summary(cex5_small$lnfoodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5539  3.1210  3.7113  3.6901  4.2935  7.0733
```



Because there are some value 0 in FOODAWAY. When take the logarithm of 0, the value will be $-\infty$, should be dropped. Therefore, the numbers of obs. will decrease.

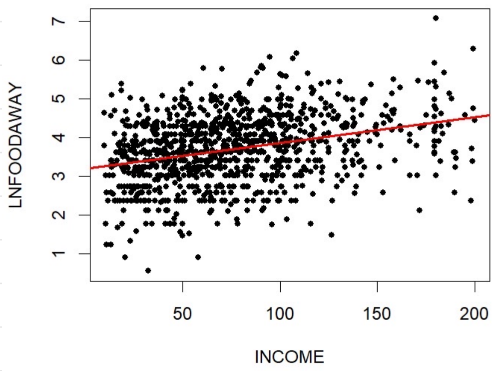
d. Estimate the linear regression $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

```
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 ***
```

Holding all else constant, in average, an addition \$100 of income, food away expenditure per person of about 0.69%

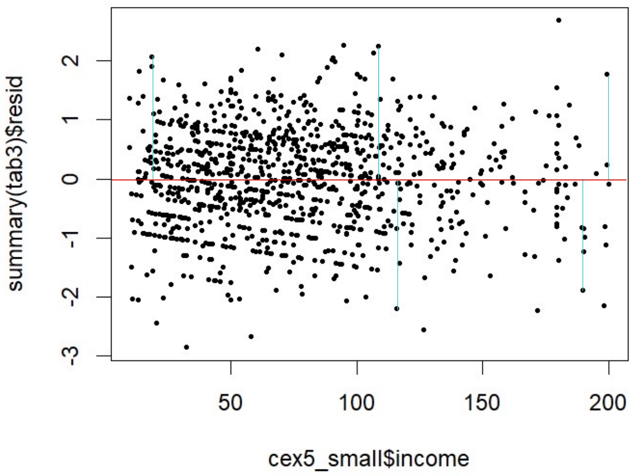
$$\widehat{\ln(FOODAWAY)} = 3.1293 + 0.0069 * INCOME$$

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?

Sum(e_i)
$=-1.135203e-14$

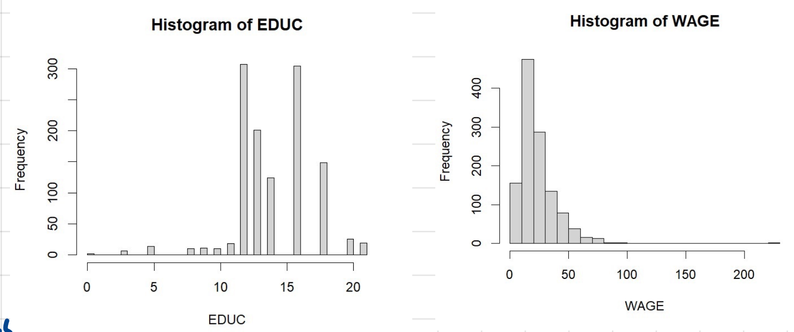


Appear ramdomly distributed with no obvious pattern.

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



Most of individuals' education year are between 10 and 18 years
Most of individuals' wage are less than 80

b. Estimate the linear regression $WAGE = \beta_1 + \beta_2 EDUC + e$ and discuss the results.

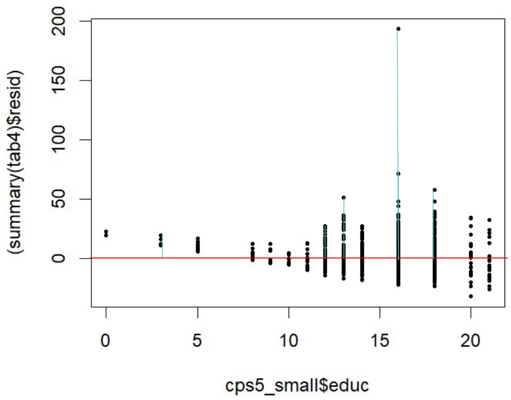
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.4000     1.9624    -5.3 1.38e-07 ***
educ          2.3968     0.1354    17.7  < 2e-16 ***
---
```

Holding all else constant, in average, an additional 1 year of education, income will increase 2.3968 units.
when an individual with 0 year eduction, the expected income is \$-10.4 units

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?

$sum(e_i) = 9.824919e-13$

In the model, the residual patterns do not appear random. The variation in the residuals increases as EDUC increases, suggesting that the homoskedasticity assumption may be violated.



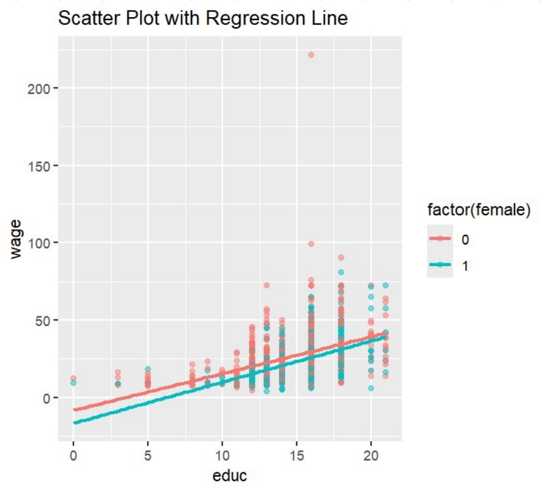
d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

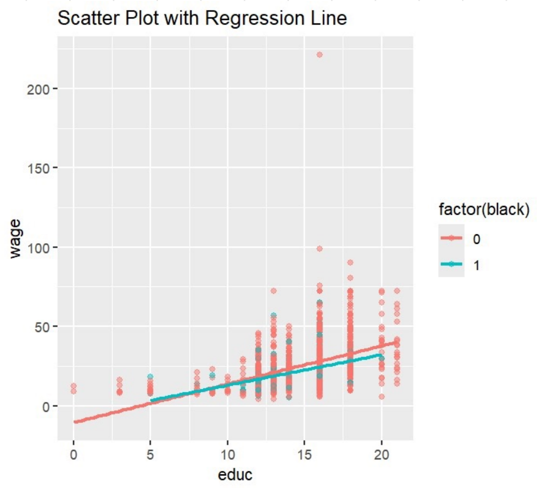
```
Call: male
lm(formula = wage ~ educ, data = cps5_small, subset = (female == 
    0))

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
```

```
Call: female
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
```

```
Call: black
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
```

```
Call: white
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
```





The growth speeds (slope) of wage are not much difference between female and male, but at any year of education, the average wage of male is higher than that of female

The growth speeds (slope) of wage in white is a little bit higher than black.

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).

$$\widehat{WAGE} = 4.9165 + 0.0891\, EDUC^2$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.916477   1.091864   4.503 7.36e-06 ***
educ_square 0.089134   0.004858  18.347  < 2e-16 ***
```

marginal effect : $\frac{\partial WAGE}{\partial EDUC} = 2\alpha_2 EDUC = 2 \times 0.0891 EDUC$

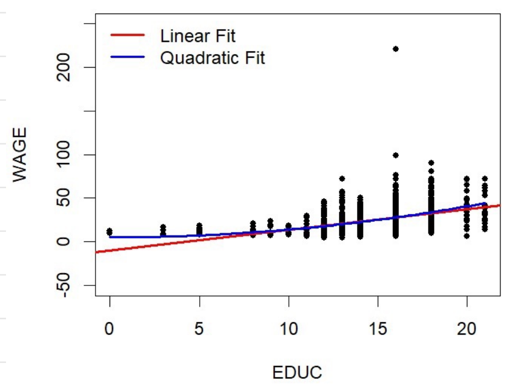
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.4000     1.9624    -5.3 1.38e-07 ***
educ          2.3968     0.1354    17.7  < 2e-16 ***
```

12y : 2.1392 ; 16y : 2.8523

marginal effect on linear : $\frac{\partial WAGE}{\partial EDUC} = \alpha_2 = 2.3968$

In a quadratic model, the marginal effect of another year of education on wage for a person with 12 years of education is smaller than in a linear model; for a person with 16 years of education, it is larger than in a linear model.

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?



Quadratic model fit the data better