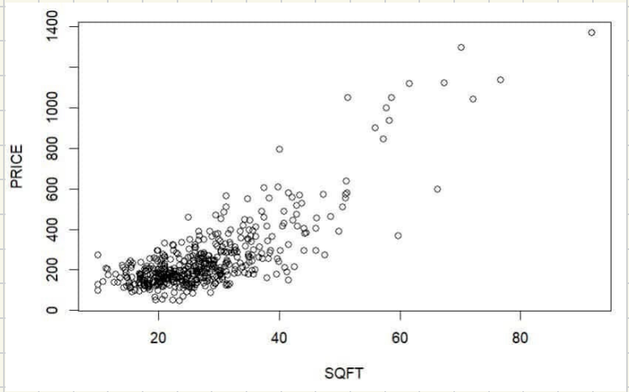
2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

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.

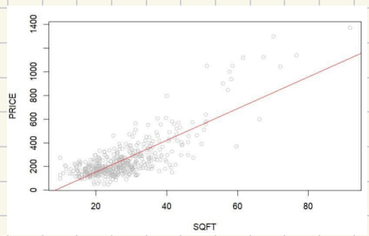
e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

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?

b.

```
> B1
[1] -115.42361   13.40294
```



Since $b_1 = -115.42361$, when the hundreds of house size equal to zero, the estimated price will be -115.42361.

Since $b_2 = 13.4$, it is the slop, that is, when the hundreds of house size increase 1 unit, the price will increase 13.4 thousands of dollars. Therefore, the price will equal to base price(b_1) add the slop (b_2) multiplied with the house size.

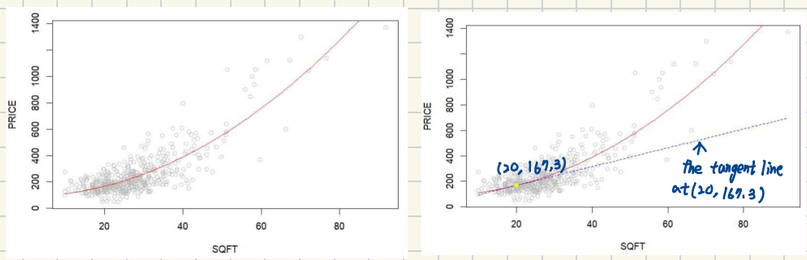
C.

```
> theMargin
[1] 7.38076
```

with R code:

```
#17c compute the margin of the quadratic model
mod_quadratic=lm(P~I(Q^2), data=PQ)
b1<-coef(mod_quadratic)[[1]]
b2<-coef(mod_quadratic)[[2]]
theQ=20
theP=b1+b2*(theQ^2)#The model
theMargin<-2*b2*theQ
theMargin
```

d.



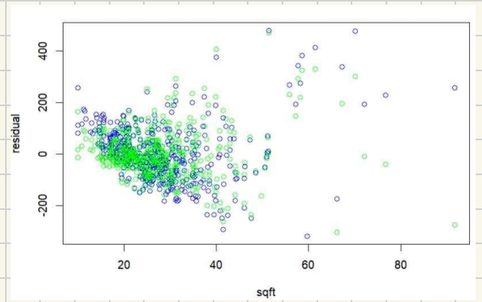
e.

```
> elasticity
[1] 0.8819511
```

with R code:

```
#The elasticity=margin*sqft/price, where price=b1+b2*sqft^2
elasticity=theMargin*theQ/(b1+b2*(theQ^2))
elasticity
```

f.



As the sqft become larger, the residual also become larger, which contradiction with homoskedasticity assumption

blue: linear model
green: quadratic model

g.

```
> sse
[1] 5262847 4222356
```

① According to the picture, the SSE of quadratic model is smaller

② since $|price_i - \hat{price}_i| = e_i$ and $SSE = \sum e_i^2$, is the model has lower SSE, that means the estimated price will be closer to the real price. Thus, the model with lower SSE is a better-fitting model.

with R code:

```
#17g e1 is the residual of linear model and e2 is residual of quadratic model
se1<-sum(e1)
se2<-sum(e2)
sse<-c(se1, se2)
sse
```

R code in Q 17

```r
#17a let P be the Price, Q be the sqft
P<-c(collegetown$price)
Q<-c(collegetown$sqft)
plot(Q, P, xlab="SQFT", ylab="PRICE", xlim=c(min(Q), max(Q)), ylim=c(min(P), max(P)), col="grey")


#17b Estimate the model coefficient and the red regression line
PQ=data.frame(Q, P)
mod_linear=lm(P~Q, data=PQ)
b11=coef(mod_linear)[[1]]
b21=coef(mod_linear)[[2]]
B1<-c(b11, b21)
B1
abline(B1, col="red")


#17c compute the margin of the quadratic model
mod_quadratic=lm(P~I(Q^2), data=PQ)
b1<-coef(mod_quadratic)[[1]]
b2<-coef(mod_quadratic)[[2]]
theQ=20
theP=b1+b2*(theQ^2)#The model
theMargin<-2*b2*theQ
theMargin

#17d
curve(b1+b2*x^2, col="red", add=TRUE)
#the tangent line y=mx+n
m<-theMargin
n<-theP-m*theQ
y<-m*theQ+n
curve(m*x+n, add=TRUE, col="blue", lty=2)
points(theQ, theP, col="yellow", pch=19)

#17e
#The elasticity=margin*sqft/price, where price=b1+b2*sqft^2
elasticity=theMargin*theQ/(b1+b2*(theQ^2))
elasticity

#17f P1 is the linear model and P2 is the quadratic model
P1=b11+b21*Q
P2=b1+b2*(Q^2)
e1<-resid(mod_linear)
e2<-resid(mod_quadratic)
plot(Q, e1, xlab="sqft", ylab="residual")
points(Q, e1, col="blue")
points(Q, e2, col="green")

#17g e1 is the residual of linear model and e2 is residual of quadratic model
se1<-sum(e1)
se2<-sum(e2)
sse<-c(se1, se2)
sse
```

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?

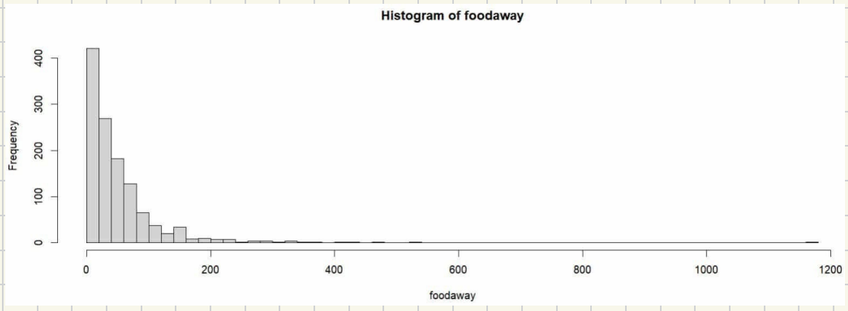
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

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.



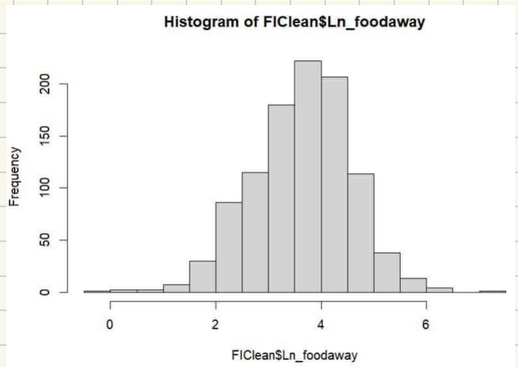

```
> summary(foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.04   32.55   49.27   67.50 1179.00
```

b.

```
> theMean
[1] 73.15494 48.59718 39.01017
> theMedian
[1] 48.15 36.11 26.02
```

From left to right: Advanced degree, college degree, neither

c.




```
> summary(FIClean$Ln_foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.3011  3.0759  3.6865  3.6508  4.2797  7.0724
```

```
> length(Ln_foodaway)
[1] 1200
> length(FIClean$Ln_foodaway)
[1] 1022
```

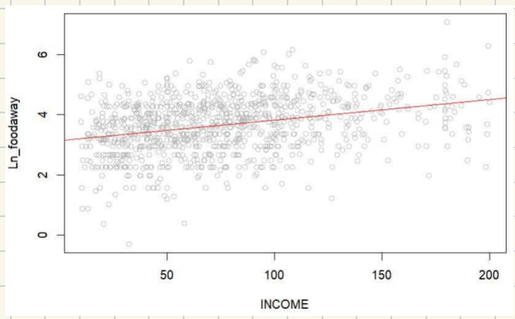
The number after transfer become smaller because some households spend 0 on foodaway, however, ln(0) is not desined, thus, those undesined data can not be used in regression model.

d.

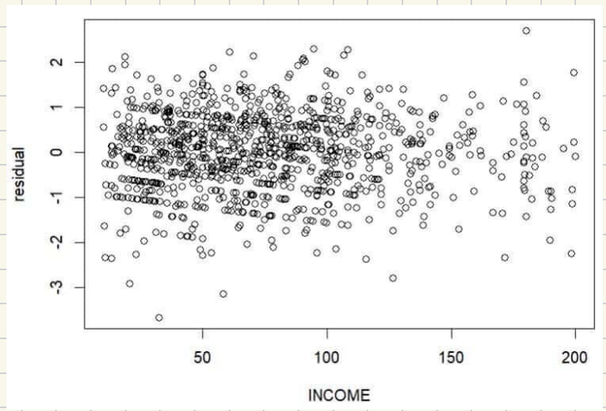
```
> b1
[1] 3.1293
> b2
[1] 0.006901748
```

Since $b_1 = 3.12$, when income equal to 0, the foodaway will be 312
Since the slope $b_2 = 0.0069$, as income increase 100 unit, the ln(foodaway) will increase 0.0069.

e.



f.



The picture seems like a random scatter, it does not have any unsual pattern.

R code in Q25

```r
#25a
foodaway<-c(cex5_small$foodaway)
advanced<-c(cex5_small$advanced)
colleged<-c(cex5_small$college)
hist(foodaway, breaks=50)
summary(foodaway)

#25b
f1<-c()
f2<-c()
f3<-c()
k<-1
#classify FOODAWAY by advanced degree and college degree
for(i in advanced){
  if(i==1){
    f1<-c(f1, foodaway[k])
  }
  else if(colleged[k]==0){
    f3<-c(f3, foodaway[k])
  }
  k=k+1
}
k<-1
for(i in colleged){
  if(i==1){
    f2<-c(f2, foodaway[k])
  }
  k=k+1
}
mean_advance<-mean(f1)
mean_college<-mean(f2)
mean_nocollege<-mean(f3)
theMean<-c(mean_advance, mean_college, mean_nocollege)
median_advance<-median(f1)
median_college<-median(f2)
median_nocollege<-median(f3)
theMedian<-c(median_advance, median_college, median_nocollege)
theMean
theMedian
```

```r
#25cd
Ln_foodaway<-log(foodaway)
length(Ln_foodaway)
income<-c(cex5_small$income)
FI=data.frame(income, Ln_foodaway)
#FIClean is the data without infinity
FIclean<-FI[!is.infinite(FI$Ln_foodaway), ]
length(FIclean$Ln_foodaway)
hist(FIclean$Ln_foodaway)
summary(FIclean$Ln_foodaway)
modFI<-lm(Ln_foodaway~income, data=FIclean)
b1<-coef(modFI)[[1]]
b2<-coef(modFI)[[2]]
b1
b2

#25e
plot(income, Ln_foodaway, xlab="INCOME", ylab="Ln_foodaway", col="grey")
abline(b1, b2, col="red")

#25f e is the residual of lnFoodaway
e<-resid(modFI)
plot(FIclean$income, e, xlab="INCOME", ylab="residual")
```

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.

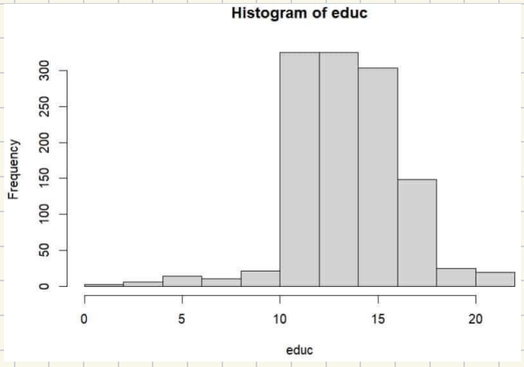
c. Calculate the least squares residuals and plot them against *EDUC*. Are any patterns evident? If assumptions SR1–SR5 hold, should any patterns be evident in the least squares residuals?

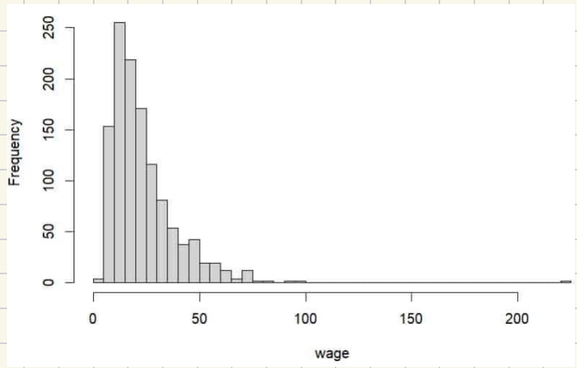
d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

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(educ)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    12.0    14.0    14.2    16.0    21.0
```

```
> summary(wage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.94   13.00   19.30   23.64   29.80  221.10
```

The histogram of wage is positive skewness, since it mean=23.64, median=19.3. More people get lower wage, few people get higher wage, that is, the distribution of Wage is not even.

The histogram of EDUC shows that most of people are educated with 10~15 years.

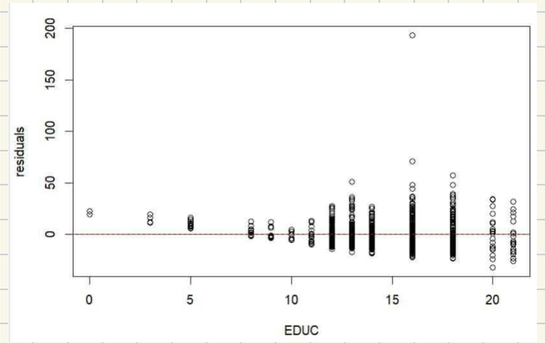
b.

```
> b11
[1] -10.39996
> b21
[1] 2.396761
```

Since $b_1 = -10.399$, when education equal to zero, the base wage is -10.399.

Since $b_2 = 2.396761$, when the education increase one unit, the wage increase 2.396

c.



The higher EDUC, the larger residual. which contradiction to homoskedasticity assumption.

If SR1-SR5 hold, it should be a random scatter. And in each EDUC, it will have zero mean, and be approximately normally distribution.

d. $\hat{wage} = b_1 + b_2 \times EDUC$

female male black white

```
> b
[1] -16.602785   2.659487  -8.284936   2.378471  -6.254114   1.923330 -10.474710   2.417769
```

The left side data is b_1. When the education is 0, the wage will equal to b_1, and b_2 is the slope, when education increase one unit, the wage will increase b_2. According to the picture, male has higher base wage(educ=0), when the education is low, male will probably have higher wage, however, female have higher slope, which means, as the education is high enough, female will have higher wage. About the race, black has higher base wage and lower slope. similarly, if the education is high enough, white has higher wage, otherwise, black has higher wage.

e.

```
> b1
[1] 4.916477
> b2
[1] 0.08913401
```

since $b_1 = 4.9$, when the education equal to zero, the wage is 4.9.

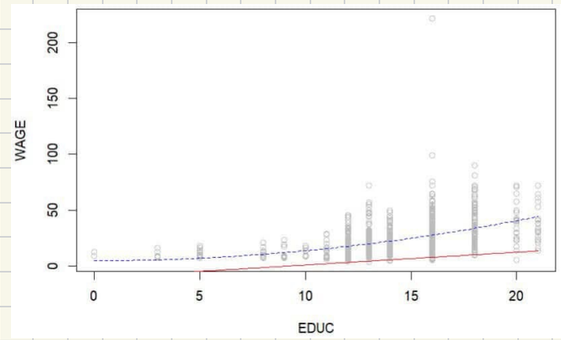
since $b_2 = 0.089$, it is a slop, As the education increase one unit, the wage will increase $2 \times 0.089 = 0.178$

```
> theMargin_qua
[1] 2.139216 2.852288
> theMargin_ln
[1] 2.396761
```

when EDUC=12, linear model effect more, since $2.13 < 2.396$

when EDUC=16, quadratic model effect more, since $2.85 > 2.396$

f.



According to the picture, the quadratic model fit the data better. The education between 10~20 has apparently more wage than those below 10 whose wage are close to zero, the quadratic model can better indicate that the education over 10 can have much higher wage

R code in Q28

```r
#28a
wage<-c(cps5_small$wage)
educ<-c(cps5_small$educ)
hist(wage, breaks=50)
summary(wage)
hist(educ)
summary(educ)

#28b
WE=data.frame(educ, wage)
modWE<-lm(wage~educ, data=WE)
b11<-coef(modWE)[[1]]
b21<-coef(modWE)[[2]]
b11
b21

#28c
e<-resid(modWE)
plot(educ, e, xlab="EDUC", ylab="residuals")
RE=data.frame(educ, e)
modRE<-lm(e~educ, data=RE)
a1<-coef(modRE)[[1]]
a2<-coef(modRE)[[2]]
abline(a1, a2, col="red", lty=2)

#28d
female<-c(cps5_small$female)
black<-c(cps5_small$black)
wage_f<-c()
educ_f<-c()
wage_m<-c()
educ_m<-c()
wage_b<-c()
educ_b<-c()
wage_w<-c()
educ_w<-c()
b<-c()
for(i in c(1:1200)){
  if(female[i]==1){
    wage_f<-c(wage_f, wage[i])
    educ_f<-c(educ_f, educ[i])
  }
  else{
    wage_m<-c(wage_m, wage[i])
    educ_m<-c(educ_m, educ[i])
  }
  if(black[i]==1){
    wage_b<-c(wage_b, wage[i])
    educ_b<-c(educ_b, educ[i])
  }
  else{
    wage_w<-c(wage_w, wage[i])
    educ_w<-c(educ_w, educ[i])
  }
}
WE1=data.frame(educ_f, wage_f)
modWE1<-lm(wage_f~educ_f, data=WE1)
b1<-coef(modWE1)[[1]]
b2<-coef(modWE1)[[2]]
b<-c(b, b1, b2)
WE2=data.frame(educ_m, wage_m)
modWE2<-lm(wage_m~educ_m, data=WE2)
b1<-coef(modWE2)[[1]]
b2<-coef(modWE2)[[2]]
b<-c(b, b1, b2)
WE3=data.frame(educ_b, wage_b)
modWE3<-lm(wage_b~educ_b, data=WE3)
b1<-coef(modWE3)[[1]]
b2<-coef(modWE3)[[2]]
b<-c(b, b1, b2)
WE4=data.frame(educ_w, wage_w)
modWE4<-lm(wage_w~educ_w, data=WE4)
b1<-coef(modWE4)[[1]]
b2<-coef(modWE4)[[2]]
b<-c(b, b1, b2)
b
#28e
mod_quatratic<-lm(wage~I(educ^2), data=WE)
b1<-coef(mod_quatratic)[[1]]
b2<-coef(mod_quatratic)[[2]]
b1
b2
theEduc<-c(12, 16)
theWage<-b1+b2*(theEduc^2)
theMargin_qua<-2*b2*(theEduc)
theMargin_ln<-b21
theMargin_qua
theMargin_ln

#28f
plot(educ, wage, xlab="EDUC", ylab="WAGE", col="grey")
curve(b11+b12*x, col="red", add=TRUE)
curve(b1+b2*(x^2), col="blue", add=TRUE, lty=2)
```