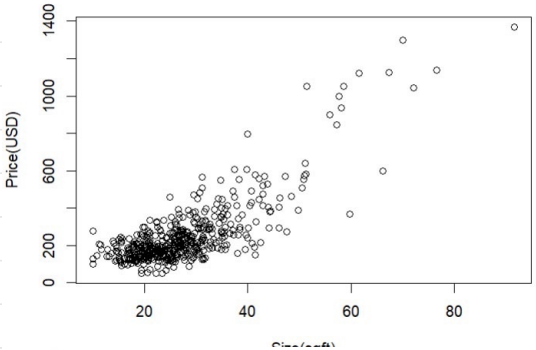
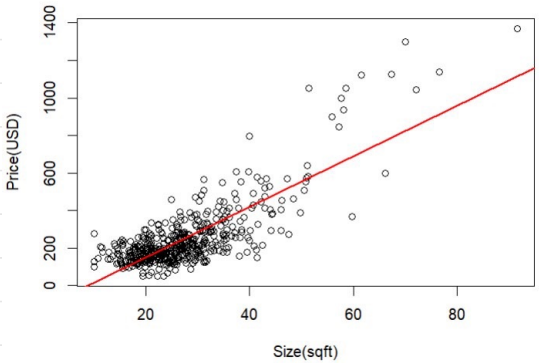
Q2.17.a





Q2.17.b

$$PRICE = \beta_1 + \beta_2 SQFT + e$$

$$\beta_2 = \frac{\partial E(PRICE \mid SQFT)}{\partial SQFT} = 13.4$$

$$\beta_1 = E(PRICE \mid SQFT = 0) = -115$$

$$\widehat{PRICE} = -115 + 13.4\, SQFT$$

Each 100 square-foot increased in size will bring up house price for $13.4 thousands on average.

Q2.17.c

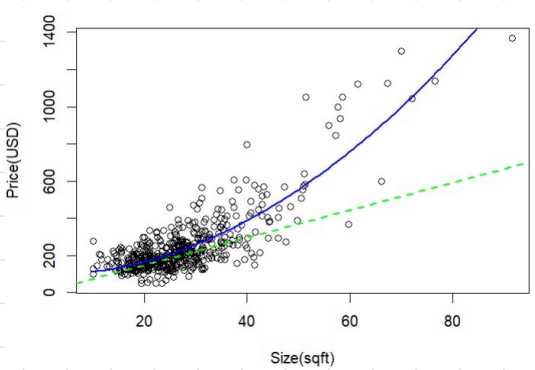
$$PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$$

$$\widehat{PRICE} = 93.6 + 0.185\, SQFT^2$$

Marginal eff: $\frac{\partial E(PRICE \mid SQFT = 20)}{\partial SQFT}$

$= 2 \times 0.185 \times 20 = 7.4$ #

Q2.17.d



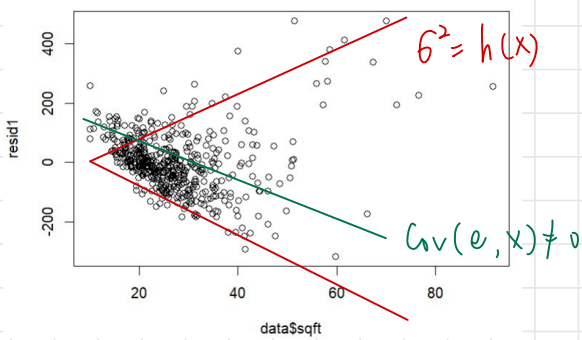
Q.2.17.e

$$\frac{\partial \widehat{PRICE}/\widehat{PRICE}}{\partial SQFT/SQFT} = 0.37 \times SQFT \cdot \frac{SQFT}{\widehat{PRICE}}$$

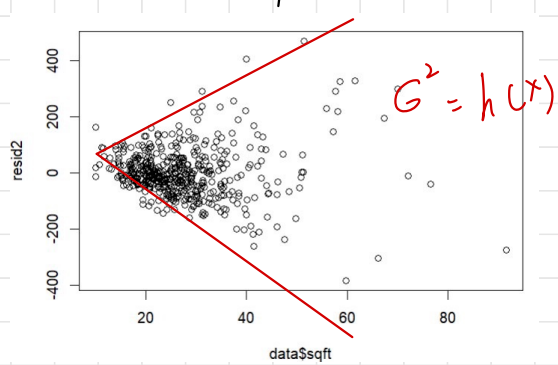
$$= 0.37 \times 20 \times \frac{20}{93.6 + 0.185 \times 20^2} = 0.883$$

Q.2.17.f

For Linear model, the residual plot seems to present a heteroskedasticity and endogenity problem $\sigma^2 = h(x)$, $Cov(e, x) \neq 0$



For Quadratic model, the residual of seems to present a heteroskedasticity problem. As for endogenity, we still need futhur method to testify.



Q. 2. 17. g

$$SSE_{LM} = \sum_{i=1}^{500} [PRICE_i - (-115 + 13.4\,SQFT_i)]^2 = \sum_{i=1}^{500} \hat{e}_i^2$$

$$= 5726846.97$$

$$SSE_{QM} = \sum_{i=1}^{500} [PRICE_i - (93.6 + 0.185\,SQFT_i^2)]^2 = \sum_{i=1}^{500} \hat{e}_i^2$$

$$= 4222356.35$$

$SSE_{QM} < SSE_{LM}$ $\quad R^2 = 1 - \frac{SSE}{SST}$, SSE↑, R^2↓

A quadratic model has lower SSE and thus bigger R^2, indicating it's a better-fitting model.

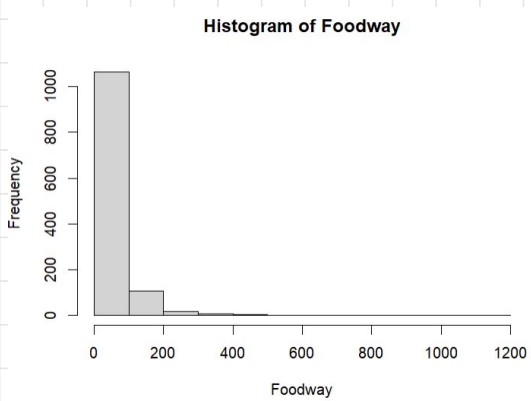
Q 2. 25. a

$M = 49.27$

$Me = 32.55$

$Q_1 = P_{25} = 12.04$

$Q_3 = P_{75} = 67.5$




```
View(data2)
hist(data2$foodaway)
hist(data2$foodaway,
     xlab = 'Foodway')
summary(data2$foodaway)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   12.04   32.55   49.27   67.50 1179.00
```

Q 2.25.b

```
> summary(data2_adv$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   21.67   48.15   73.15   90.00 1179.00
> summary(data2_col$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   14.44   36.11   48.60   68.67  416.11
> summary(data2_non$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.63   26.02   39.01   52.65  437.78
```

For household with advenced degree,

$M_{adv} = 73.15$ $Me_{adv} = 48.15$

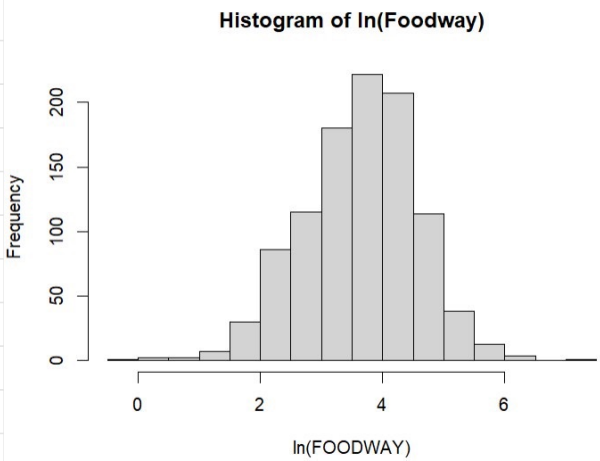
For households with college degree,

$M_{col} = 48.6$ $Me_{col} = 36.11$

For households with no degree,

$M_{non} = 39.01$ $Me_{non} = 26.02$

Q 2.25.c



Foodway is skewed to the right and takes a way bigger value ln(Foodway) scaled the values to smaller number and make the distribution closer to normality.

Q 2.25.d $\widehat{\ln(FOODWAY)} = 3.1293 + 0.0069\ INCOME$

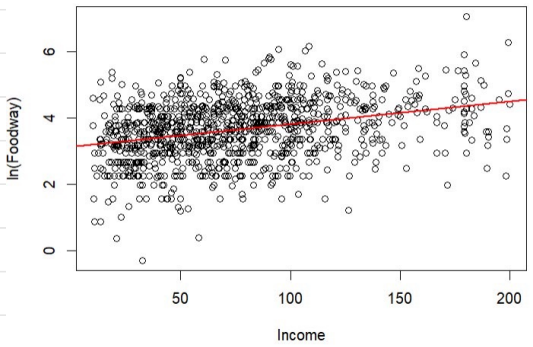
```
Call:
lm(formula = data2$lnfoodway ~ data2$income)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6547 -0.5777  0.0530  0.5937  2.7000

Coefficients:
             Estimate Std. Error t value
(Intercept)  3.1293004  0.0565503   55.34
data2$income 0.0069017  0.0006546   10.54
             Pr(>|t|)
(Intercept)    <2e-16 ***
data2$income   <2e-16 ***
```

$$\frac{\partial \ln(FOODWAY)}{\partial\ INCOME} = \frac{\partial Foodway / Foodway}{\partial INCOME}$$

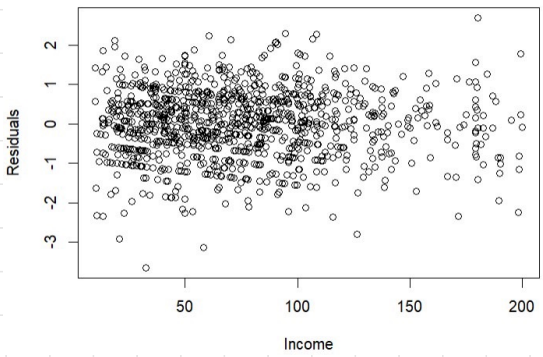
$= 0.0069$

$100 unit change in INCOME will bring 0.69% change in FOODWAY

Q2.25.e



Q2.25.f



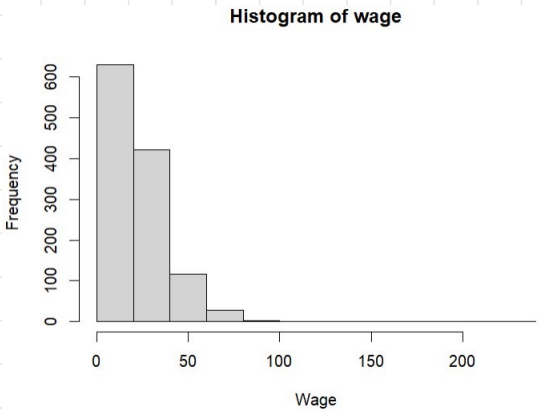
It seems the Gauss-Markov assumptions are not violated from the residual plot.

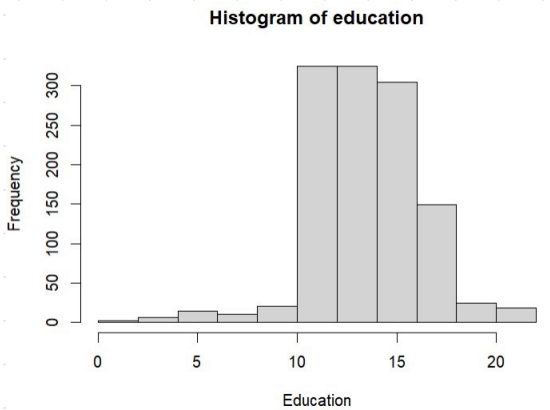
Q2.28.a

```
> summary(data3$wage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.94   13.00   19.30   23.64   29.80  221.10
> summary(data3$educ)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    12.0    14.0    14.2    16.0    21.0
```

Wage skewed to the right.

Education skewed to the left.





Q 2.28.b

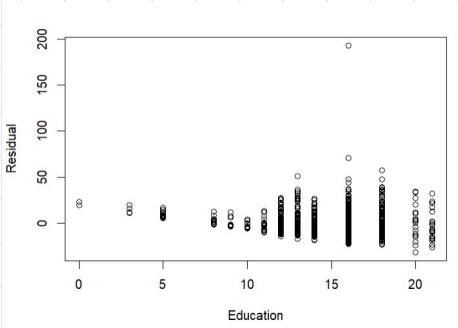
```
Call:
lm(formula = data3$wage ~ data3$educ)

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
data3$educ    2.3968     0.1354    17.7  < 2e-16

(Intercept) ***
data3$educ  ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\widehat{Wage} = -10.4 + 2.3968\ Educ$

The year of education will postively affect hourly wage rate.

Each 1 year increase will bring up $2.3968 / hr.

Q 2.28.c



The variation of residuals gets bigger with education year, showing a Heteroskedasticity problem occured in the regression setting.

If SR1-5 holds $\begin{cases} e_i \overset{iid}{\sim} N(0, \sigma^2) \\ Cov(educ, e) = 0 \\ \sigma^2 \neq h(x) \text{ but fixed} \end{cases}$

→ residual plot should be fully random

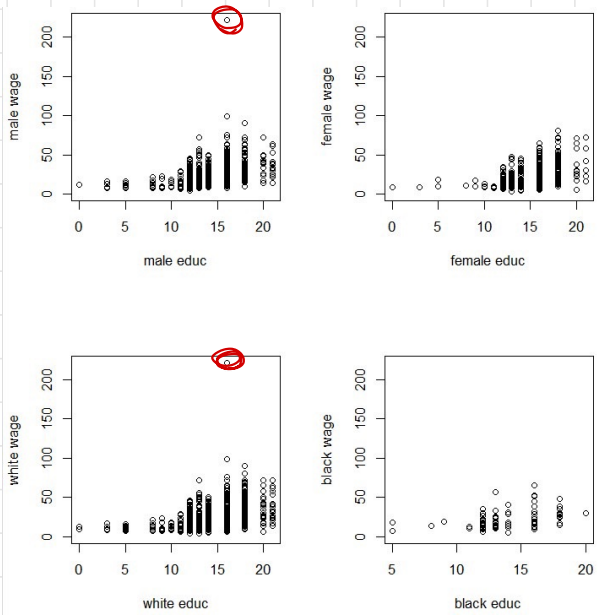
Q 2.28.d

```
$male

Call:
lm(formula = wage ~ educ, data = df)

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
```

```
$female

Call:
lm(formula = wage ~ educ, data = df)

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
```

```
$white

Call:
lm(formula = wage ~ educ, data = df)

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
```

```
$black

Call:
lm(formula = wage ~ educ, data = df)

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
```



Note that there's a white male earn extremely more than other observations and it thus affect the regression result.

We should therefore winsorize it to do further analysis and inference.

Q2.28.e

```
Call:
lm(formula = wage ~ I(educ^2), data = data3)

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
```

For quadratic model,

$\widehat{Wage} = 4.9165 + 0.0891\ Educ^2$

Marginal eff.:

$$\frac{\partial E(Wage \mid Educ)}{\partial Educ} = 2 \times 0.0891 \times Educ$$

$$\frac{\partial E(Wage \mid Educ=12)}{\partial Educ} = 2 \times 0.0891 \times 12 = 2.1384$$

$$\frac{\partial E(Wage \mid Educ=16)}{\partial Educ} = 2 \times 0.0891 \times 16 = 2.8512$$

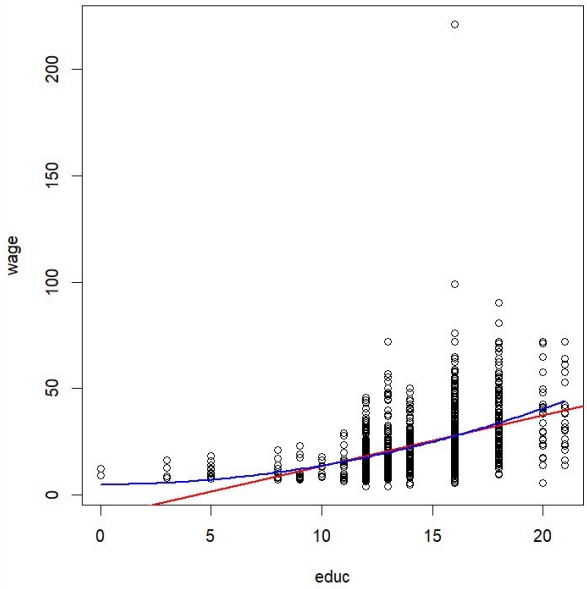
For linear model

$\widehat{Wage} = -10.4 + 2.3968\ Educ$

Marginal eff.

$$\frac{\partial \widehat{Wage}}{\partial Educ} = 2.3968$$

Q2.28. f




```
> summary(mod4)$r.squared
[1] 0.2073273
> summary(mod5)$r.squared
[1] 0.2193561
```

Mod4 is a linear model with 0.2073 R^2.

Mod5 is a quadratic model with 0.2194 R^2.

The quadratic model performs better than the linear model.