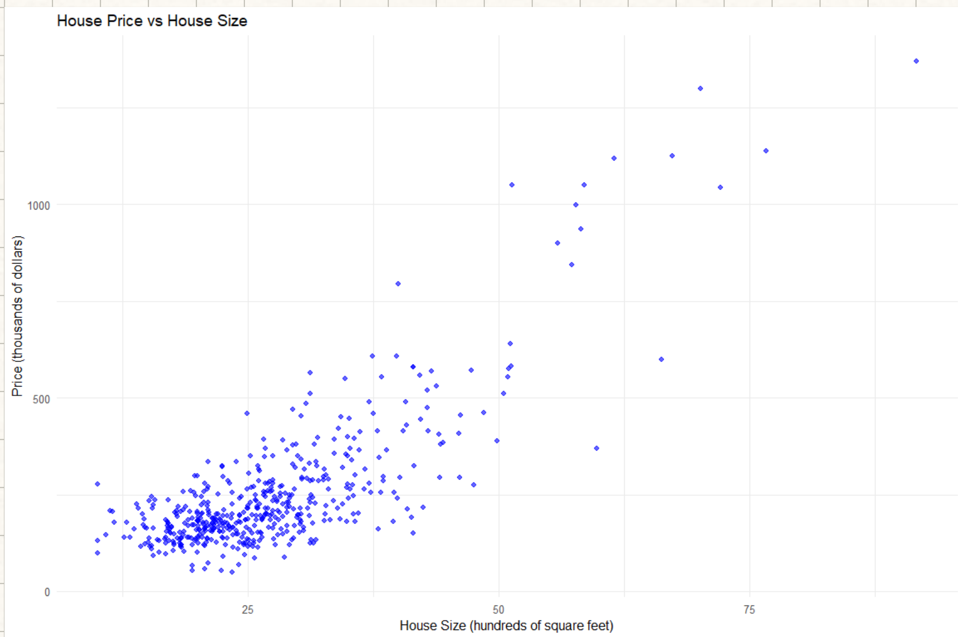
2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

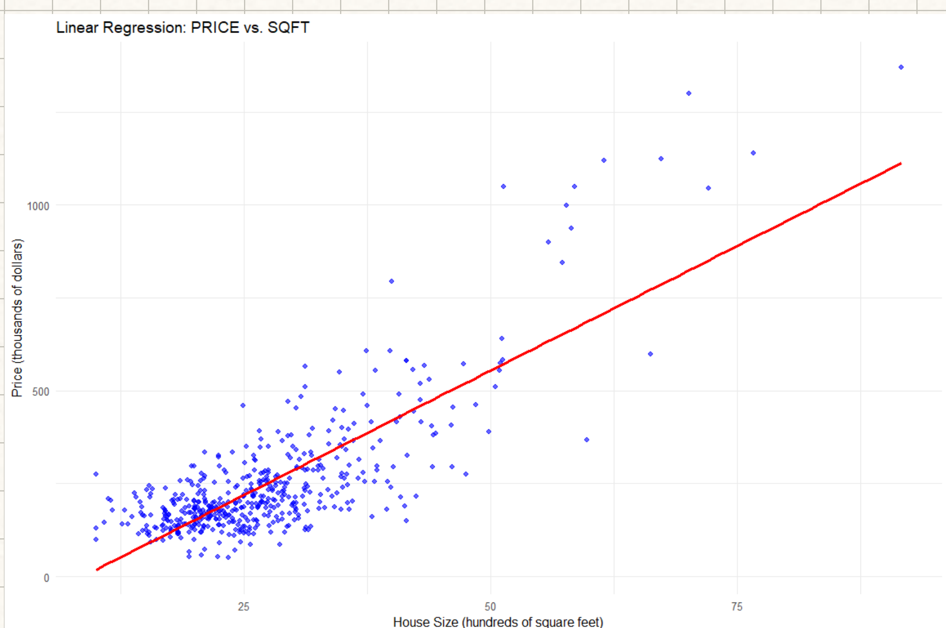
a. Plot house price against house size in a scatter diagram.



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 ***
```

$\widehat{Price} = -115.4236 + 13.4029 \cdot SQFT$ ☆



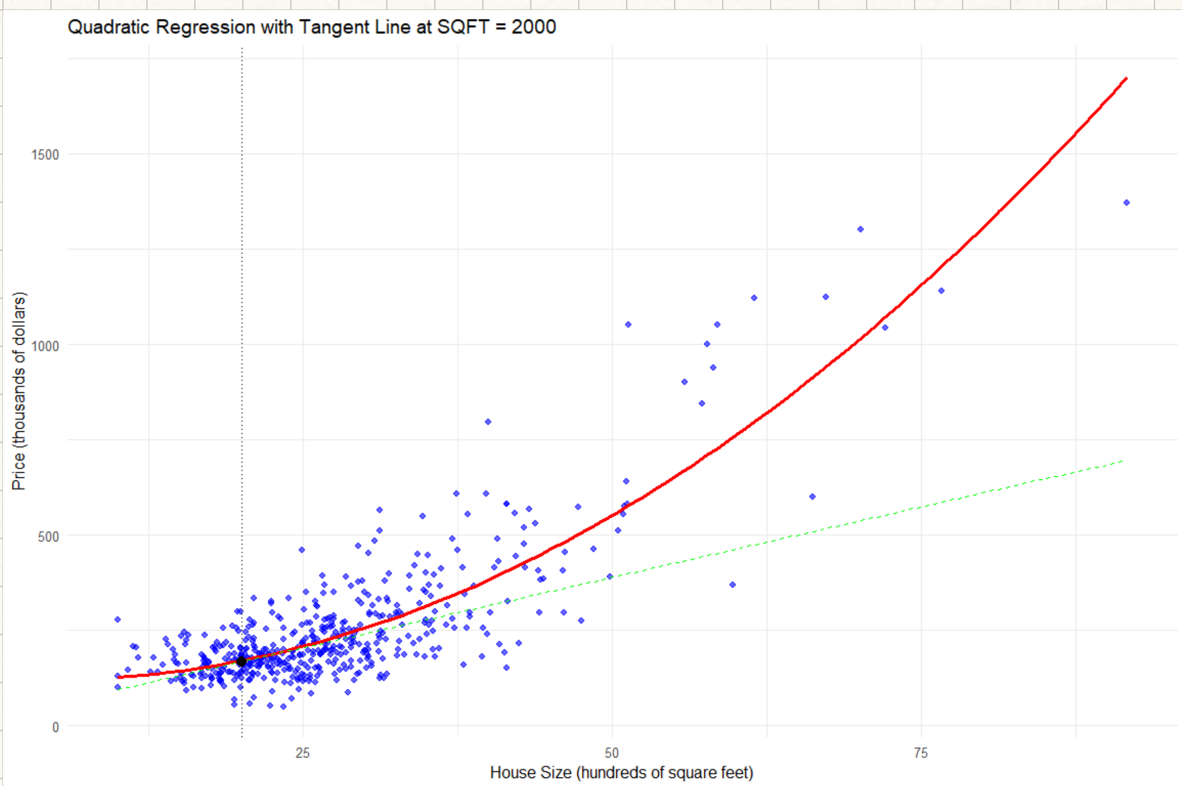
c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.565854   6.072226   15.41   <2e-16 ***
sqft2        0.184519   0.005256   35.11   <2e-16 ***
```

$$Price = 93.565854 + 0.184519 \cdot SQFT^2$$

```
> alpha2 <- coef(model_quad_pure)["sqft2"]
> sqft_value <- 20  # 2000 平方英尺 (單位是 "hundreds of square feet")
> marginal_effect <- 2 * alpha2 * sqft_value
> # 顯示邊際影響
> cat("當房屋面積為 2000 平方英尺時，增加 100 平方英尺對價格的影響為：", marginal_effect, "千美元\n")
當房屋面積為 2000 平方英尺時，增加 100 平方英尺對價格的影響為： 7.38076 千美元
```

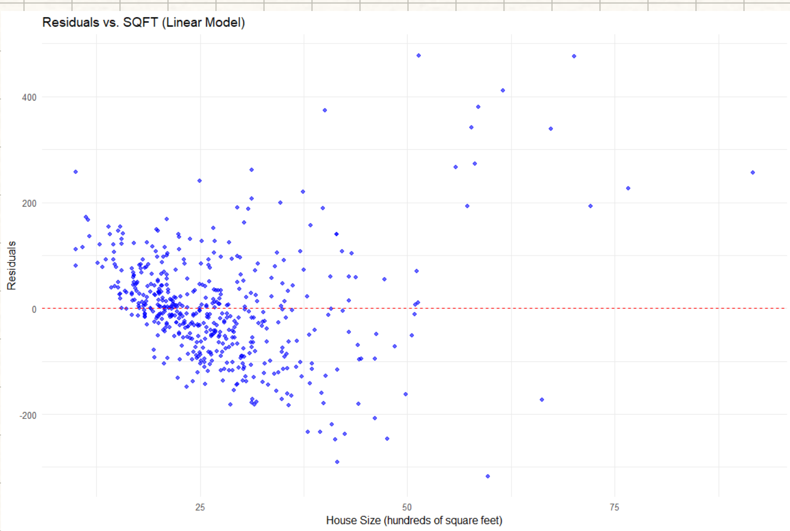
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

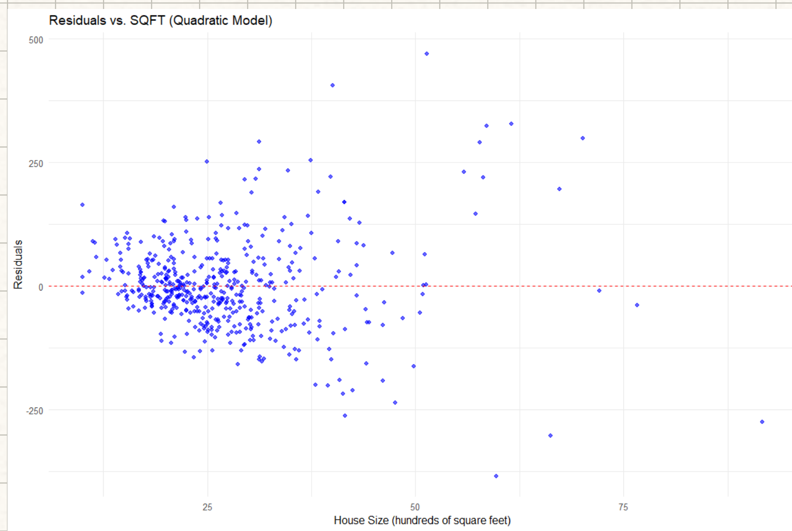


e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

```
> # 提取回歸係數
> alpha0 <- coef(model_quad_pure)["(Intercept)"]
> alpha2 <- coef(model_quad_pure)["sqft2"]
> # 設定 SQFT = 20 ( 2000 平方英尺 )
> sqft_value <- 20
>
> # 計算邊際影響
> marginal_effect <- 2 * alpha2 * sqft_value
>
> # 計算 PRICE 預測值
> price_predicted <- alpha0 + alpha2 * sqft_value^2
>
> # 計算彈性
> elasticity <- (marginal_effect * sqft_value) / price_predicted
>
> # 顯示結果
> cat("當房屋面積為 2000 平方英尺時，PRICE 對 SQFT 的彈性為：", elasticity, "\n")
當房屋面積為 2000 平方英尺時，PRICE 對 SQFT 的彈性為： 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?





兩個模型中，殘差的變異數隨 SQFT 增加而變大，

表示可能不滿足 homoskedasticity 假設

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?

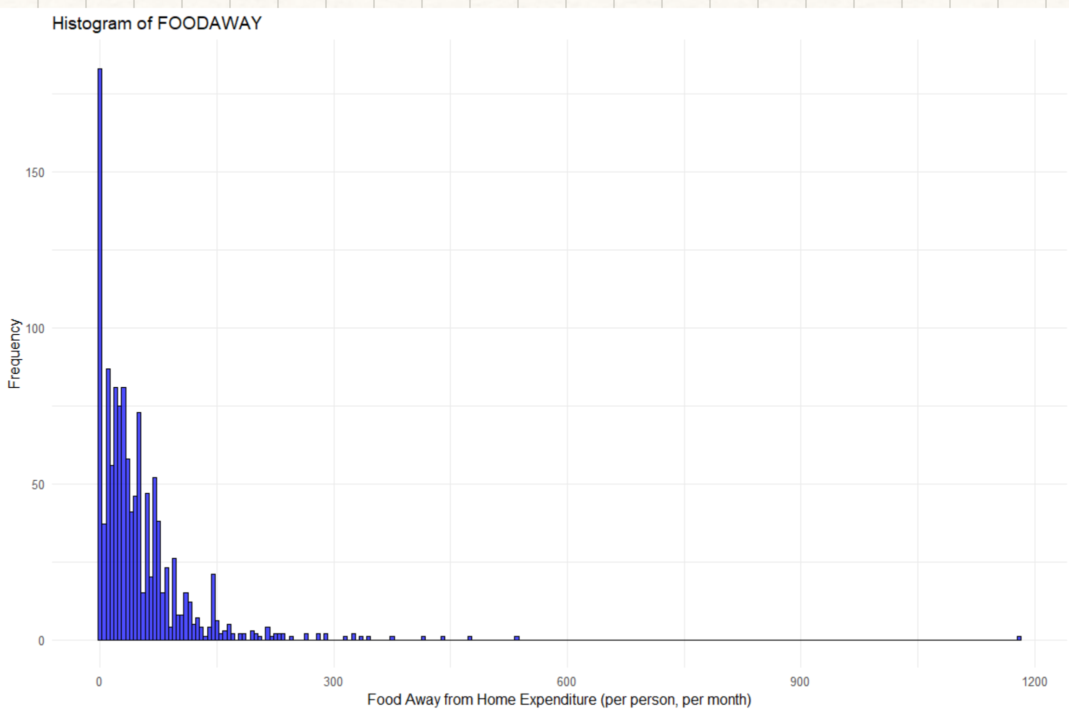
```
> # 計算 SSE (Sum of Squared Residuals)
> SSE_linear <- sum(residuals(model1)^2)          # 線性回歸
> SSE_quadratic <- sum(residuals(model_quad_pure)^2)  # 二次回歸
> # 輸出 SSE 結果
> cat("SSE for Linear Model:", SSE_linear, "\n")
SSE for Linear Model: 5262847
> cat("SSE for Quadratic Model:", SSE_quadratic, "\n")
SSE for Quadratic Model: 4222356
```

Quadratic model 的 SSE 比 Linear model 的 SSE 小

SSE 衡量模型的殘差平方合，表示模型預測值與實際值的誤差，所以 SSE 越小，表示模型能更準確擬合數據。

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?




```
> # 計算 foodaway 的統計數據
> foodaway_mean <- mean(cex5_small$foodaway, na.rm = TRUE)  # 計算平均值
> foodaway_median <- median(cex5_small$foodaway, na.rm = TRUE)  # 計算中位數
> foodaway_q1 <- quantile(cex5_small$foodaway, 0.25, na.rm = TRUE)  # 25th 百分位數
> foodaway_q3 <- quantile(cex5_small$foodaway, 0.75, na.rm = TRUE)  # 75th 百分位數
> # 顯示結果
> cat("平均值:", foodaway_mean, "\n")
平均值: 49.27085
> cat("中位數:", foodaway_median, "\n")
中位數: 32.555
> cat("25th 百分位數:", foodaway_q1, "\n")
25th 百分位數: 12.04
> cat("75th 百分位數:", foodaway_q3, "\n")
75th 百分位數: 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?

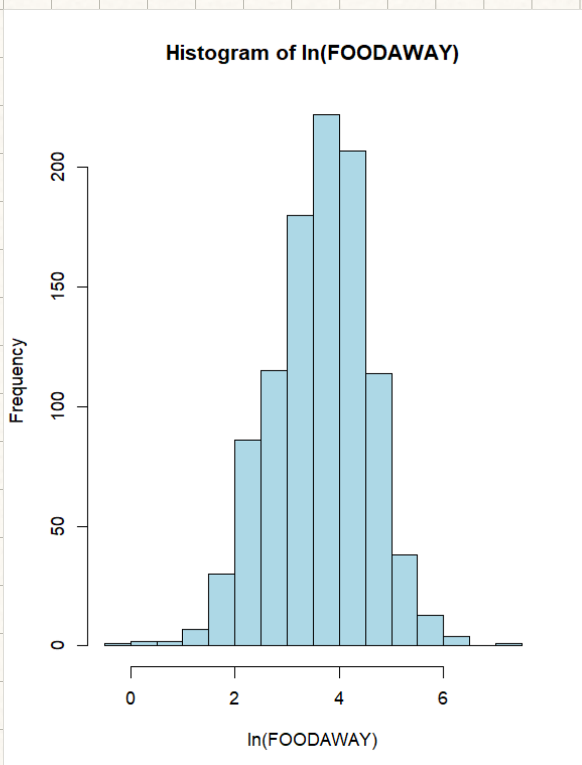
```
> # 輸出結果
> cat("\nHousehold with Advanced Degree:\n")

Household with Advanced Degree:
> cat("Mean FOODAWAY:", mean_advanced, "\n")
Mean FOODAWAY: 73.15494
> cat("Median FOODAWAY:", median_advanced, "\n")
Median FOODAWAY: 48.15
> cat("\nHousehold with College Degree:\n")

Household with College Degree:
> cat("Mean FOODAWAY:", mean_college, "\n")
Mean FOODAWAY: 48.59718
> cat("Median FOODAWAY:", median_college, "\n")
Median FOODAWAY: 36.11
>
> cat("\nHousehold with No College or Advanced Degree:\n")

Household with No College or Advanced Degree:
> cat("Mean FOODAWAY:", mean_no_degree, "\n")
Mean FOODAWAY: 39.01017
> cat("Median FOODAWAY:", median_no_degree, "\n")
Median FOODAWAY: 26.02
```

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




```
> cat("Mean ln(FOODAWAY):", mean_log_foodaway, "\n")
Mean ln(FOODAWAY): 3.650804
> cat("Median ln(FOODAWAY):", median_log_foodaway, "\n")
Median ln(FOODAWAY): 3.686499
> cat("25th Percentile ln(FOODAWAY):", percentile_25_log, "\n")
25th Percentile ln(FOODAWAY): 3.075929
> cat("75th Percentile ln(FOODAWAY):", percentile_75_log, "\n")
75th Percentile ln(FOODAWAY): 4.279717
```

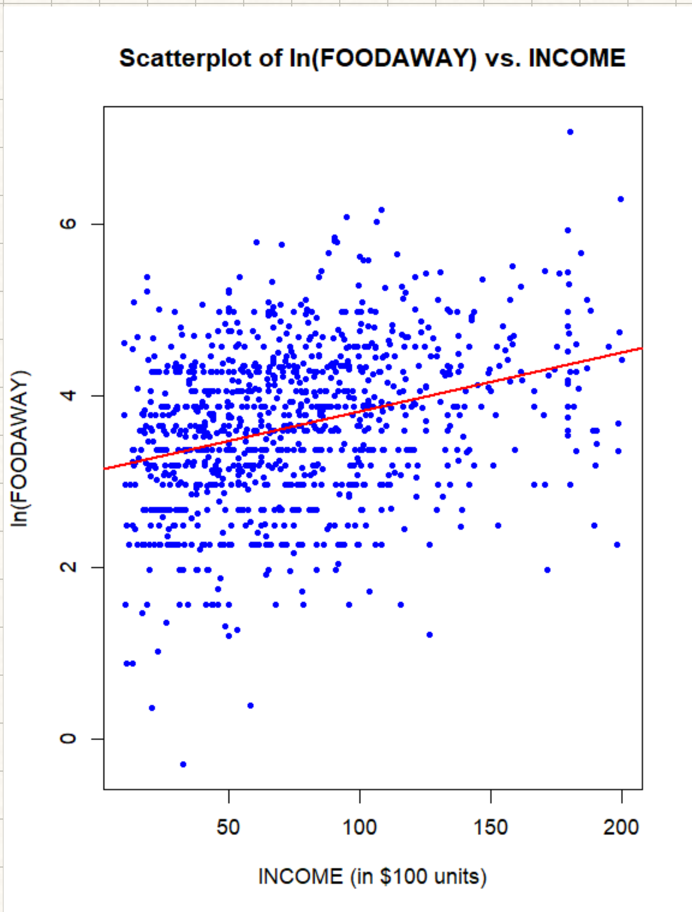
因為 ln(0) 並不存在 所以這些值需要被移除，造成觀測數減少

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

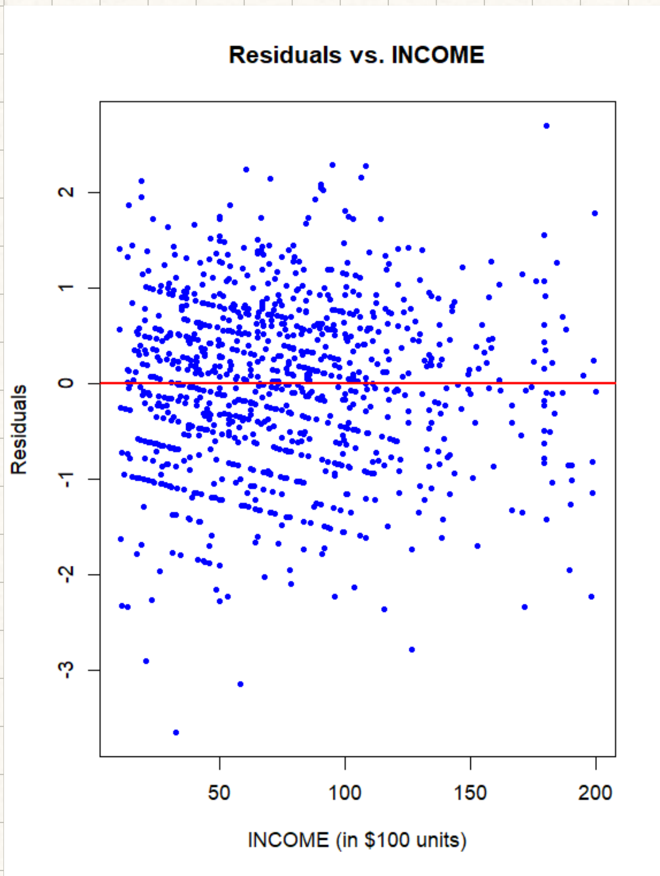
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.1293004  0.0565503   55.34   <2e-16 ***
income      0.0069017  0.0006546   10.54   <2e-16 ***
```

$\ln(FOODAWAY) = 3.1293004 + 0.0069017 \times INCOME$

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?



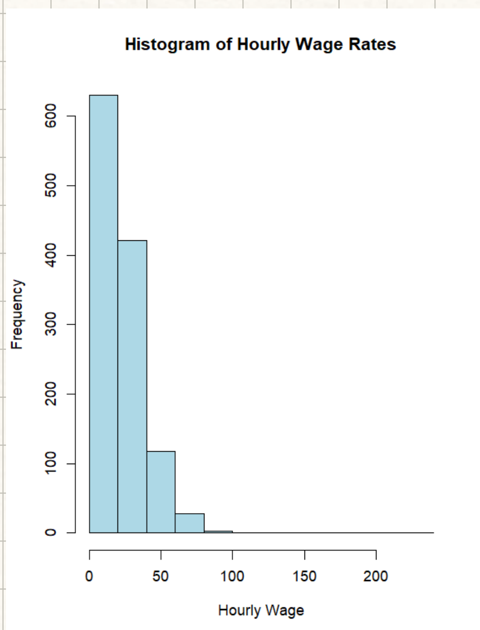
They are mostly random.

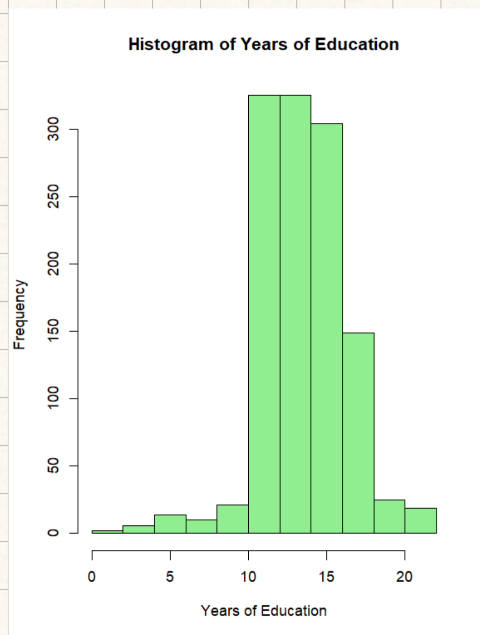
But with minor concerns about heteroscedasticity.

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





WAGE：明顯右偏，因為最大值 221.10 遠超過 Q_3 的 29.8，

說明少數高收入者拉高平均值

EDUC：受教育年數集中在 12~16 年之間，代表大多數人至少

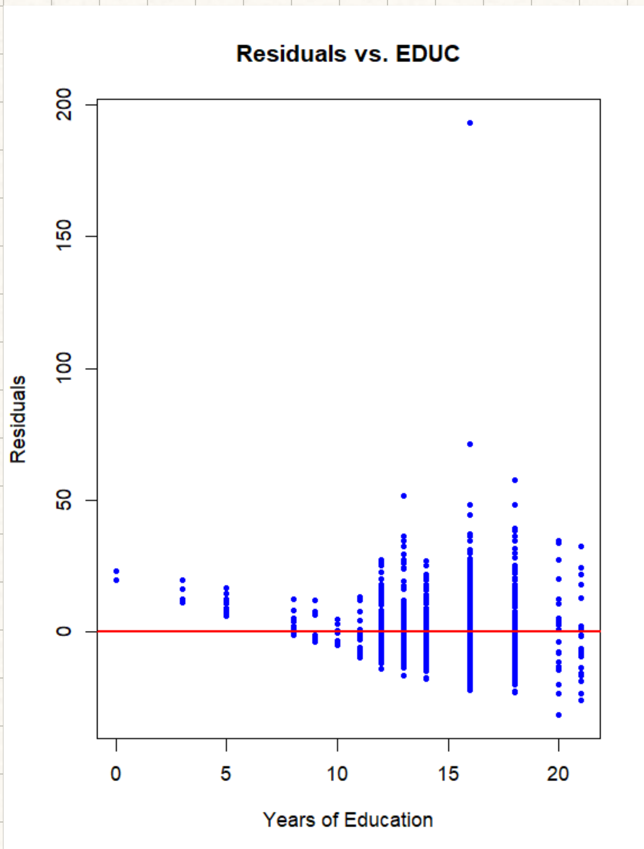
完成高中，並有部分人繼續接受大學教育

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 ***
```

$\widehat{WAGE} = -10.4000 + 2.3968 \cdot EDUC$

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?



殘差圖顯示明顯的異質變異性與可能的非線性關係，違反SR1和SR5

如果SR1–SR5成立，殘差圖應該是隨機分佈的，沒有任何模式

※

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

Males :

```
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 ***
```

Females :

```
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 ***
```

Blacks

```
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 ***
```

Whites

```
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 ***
```

⇒ 女人的教育回報最高，但起薪較低

黑人的教育回報最低，且 R^2 也最低，可能表示教育不是黑人薪資的主要決定因素。※

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).

```
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 ***
```

$WAGE = 4.9165 + 0.0891 \cdot EDUC^2$ nonlinear

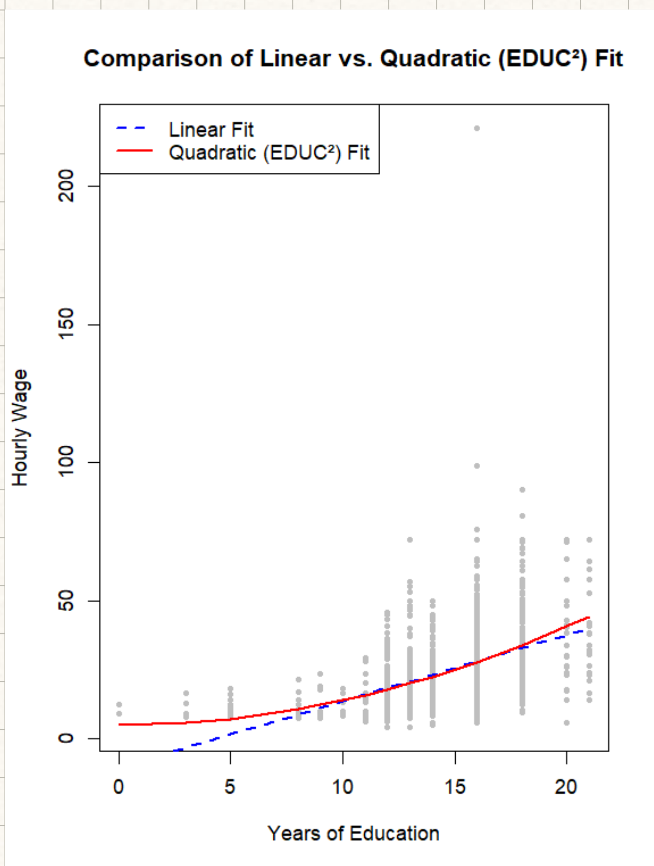
if $EDUC = 12$ $d(WAGE)/d(EDUC) = 2 \times 0.0891 \times 12 = 2.1392$

if $EDUC = 16$ $d(WAGE)/d(EDUC) = 2 \times 0.0891 \times 12 = 2.8523$

⇒ 邊際效果隨 EDUC 增加而增大

二次回歸比起線性回歸更強調教育對薪資的遞增影響 ※

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 更能捕捉到高學歷者的薪資成長，特別是在 EDUC>15 時表現更好.

※