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.

price = -115.42 + 13.4 sqft

其他條件相同每增加 100 單位的 sqft 會增加 13.4 thousands 的 home price

β_1 = -115.42 代表在 0 sqft 時，我們預估的 home price 為 -115.42 thousands

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.

$\widehat{price} = 93.57 + 0.18452 * sqft^2$

增加 2000 sqft，price 增加

$0.18452 \times 2 \times 20 = 7.38$ thousands

curve for a 2000 square foot house.

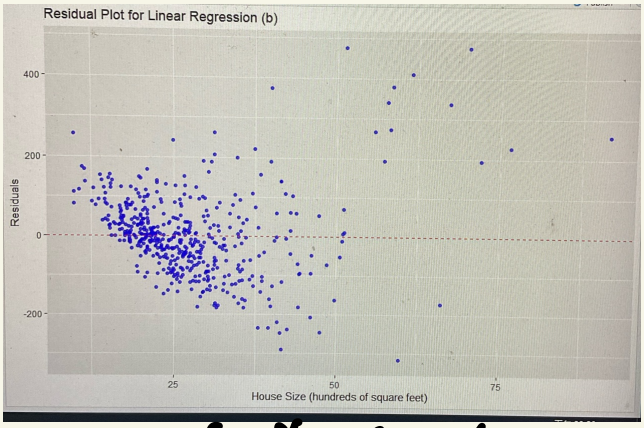
e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

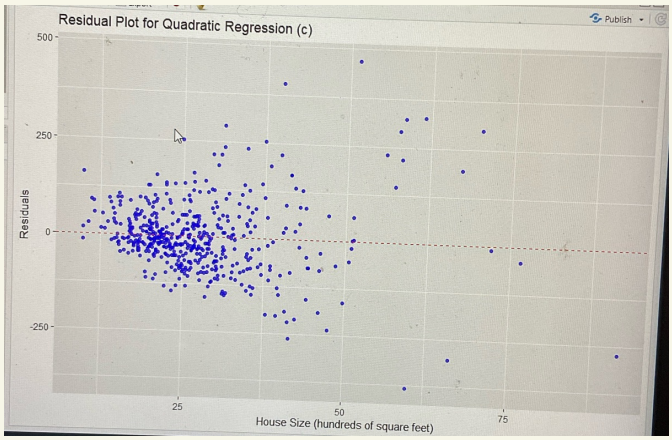
$$\varepsilon = \frac{d\,price}{d\,sqft} \times \frac{sqft}{price} = 2\beta_2 \cdot 20 \cdot \frac{20}{167.37}$$

$$= 0.8819$$

$sqft = 20 \Rightarrow \widehat{price} = 93.57 + 0.18452 \times 20^2 = 167.37$

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?





殘差值隨sqft個而變大，違反同質變異數假設

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?

SSE model(b) = 5262847

SSE model(c) = 4222356

∴ quadratic model has lower SSE

$SSE = \sum(y_i - \hat{y}_i)^2$，較小的SSE代表資料更貼近預估模型

2.17 (a)(b).R

Untitled1* | Untitled2* | collegetown | model | 2.17 a b.R | .Rhistory*

Source on Save | Run

```r
model <- lm(price ~ sqft, data = collegetown)
coefficients <- summary(model)$coefficients
coef_intercept <- coefficients[1, 1]  # 截距
coef_slope <- coefficients[2, 1]  # 斜率
intercept_se <- coefficients[1, 2]  # 截距標準誤
slope_se <- coefficients[2, 2]  # 斜率標準誤
regression_eq <- paste0("price = ", round(coef_intercept, 2), " + ", round(coef_slope, 2), " * sqft")
regression_se <- paste0("SE(Intercept) = ", round(intercept_se, 2), ", SE(Slope) = ", round(slope_se, 2))
ggplot(collegetown, aes(x = sqft, y = price)) +
  geom_point(color = "blue", alpha = 0.6) +  # 散點圖
  geom_smooth(method = "lm", se = FALSE, color = "red", size = 1) +  # 回歸線
  labs(
    title = "Scatter Plot of Price and Size with Regression Line",
    x = "House Size (hundreds of square feet)",
    y = "Price (thousands of dollars)"
  ) +
  annotate("text", x = max(collegetown$sqft) * 0.5, y = max(collegetown$price) * 0.8,
           label = paste(regression_eq, "\n", regression_se), size = 5, color = "black", hjust = 0)  # 調整大小
```

2.17 (c)(d).R

Run

```r
# 創建平方項
collegetown$sqft2 <- collegetown$sqft^2

# 估計線性回規模型
quad_model <- lm(price ~ sqft2, data = collegetown)


coef_intercept <- coef(quad_model)[1]
coef_sqft2 <- coef(quad_model)[2]

# 計算殘差
se_intercept <- summary(quad_model)$coefficients[1, 2]
se_sqft2 <- summary(quad_model)$coefficients[2, 2]


x0 <- 20
y0 <- coef_intercept + coef_sqft2 * x0^2
slope_tangent <- 2 * coef_sqft2 * x0  #x=20切線斜率

#回歸公式包含標準差
regression_eq <- paste0("price = ",
                        round(coef_intercept, 2), " + ",
                        round(coef_sqft2, 5), " * sqft*2\n",
                        "Intercept SE = ", round(se_intercept, 3), "\n",
                        "sqft? SE = ", round(se_sqft2, 3))


ggplot(collegetown, aes(x = sqft, y = price)) +
  geom_point(color = "blue", alpha = 0.6) +  # 散點圖
  geom_smooth(method = "lm", formula = y ~ I(x^2), se = TRUE, color = "red", size = 1) +  # 二次回歸線
  geom_abline(intercept = y0 - slope_tangent * x0, slope = slope_tangent, color = "blue", linetype = "dashed", size = 1) +  # 切線
  geom_point(aes(x = x0, y = y0), color = "black", size = 3) +
  labs(
    title = "Quadratic Regression: Price vs. Size (with Tangent at X=20)",
    x = "House Size (hundreds of square feet)",
    y = "Price (thousands of dollars)"
  ) +
  annotate("text", x = max(collegetown$sqft) * 0.6, y = max(collegetown$price) * 0.9,
           label = regression_eq, size = 5, color = "black", hjust = 0)
```

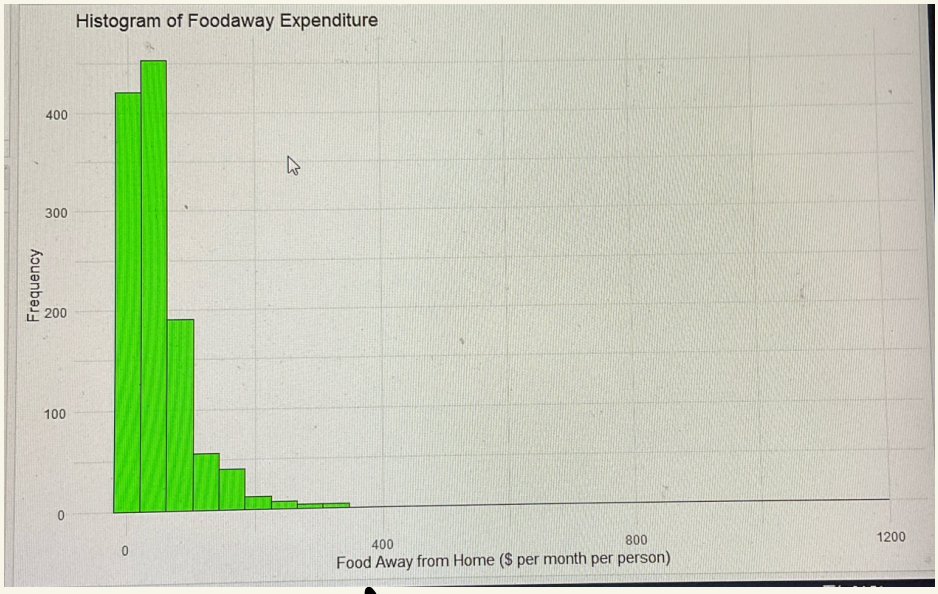
32:65 (Top Level)

2.17 fig1.R

```r
# 創建平方項
collegetown$sqft2 <- collegetown$sqft^2


linear_model <- lm(price ~ sqft, data = collegetown)


quad_model <- lm(price ~ sqft2, data = collegetown)

# 計算殘差並存入資料框
collegetown$linear_resid <- collegetown$price - predict(linear_model)
collegetown$quad_resid <- collegetown$price - predict(quad_model)


p1 <- ggplot(collegetown, aes(x = sqft, y = linear_resid)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(
    title = "Residual Plot for Linear Regression (b)",
    x = "House Size (hundreds of square feet)",
    y = "Residuals"
  )


p2 <- ggplot(collegetown, aes(x = sqft, y = quad_resid)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(
    title = "Residual Plot for Quadratic Regression (c)",
    x = "House Size (hundreds of square feet)",
    y = "Residuals"
  )


print(p1)
print(p2)
linear_resid <- collegetown$price - predict(linear_model)
quad_resid <- collegetown$price - predict(quad_model)
SSE_linear <- sum(linear_resid^2)
SSE_quad <- sum(quad_resid^2)
cat("Linear  SSE:", SSE_linear, "\n") #顯示結果
cat("Quadratic SSE:", SSE_quad, "\n")
```

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?



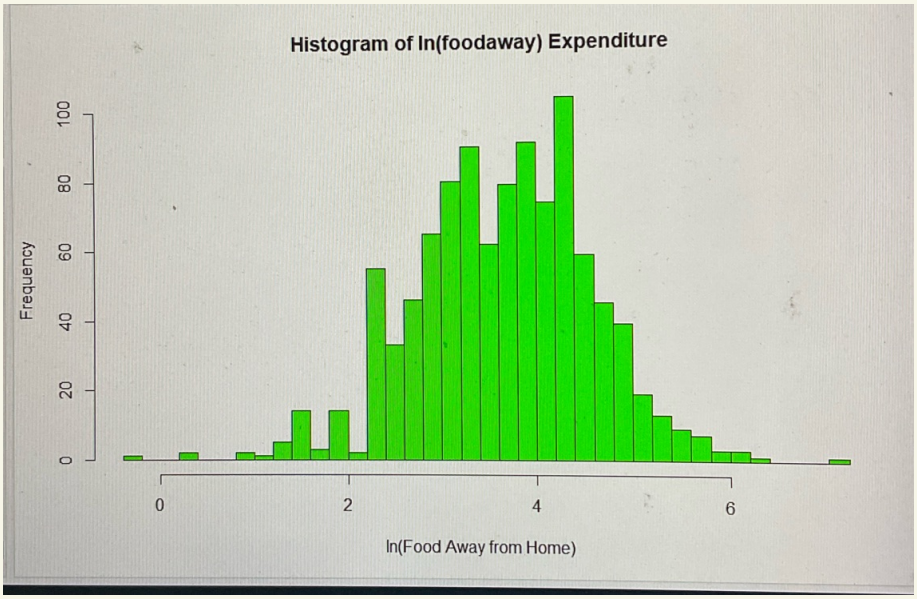
Mean 49.29085 Median 32.555 25th 12.04 75th 69.525

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?

	category	foodaway.Mean	foodaway.Median	N
	Advanced Degree	73.15494	48.15000	257
2	College Degree	48.59718	36.11000	369
3	No College/Advanced Degree	39.01017	26.02000	574

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

```
> print(summary_log)
        Mean   Median      Q25      Q75 Count_Log Count_Original
25% 3.650804 3.686499 3.075929 4.279717      1022           1200
```



Count_Original = 1200

Count_Log = 1022

1200 - 1022 = 178

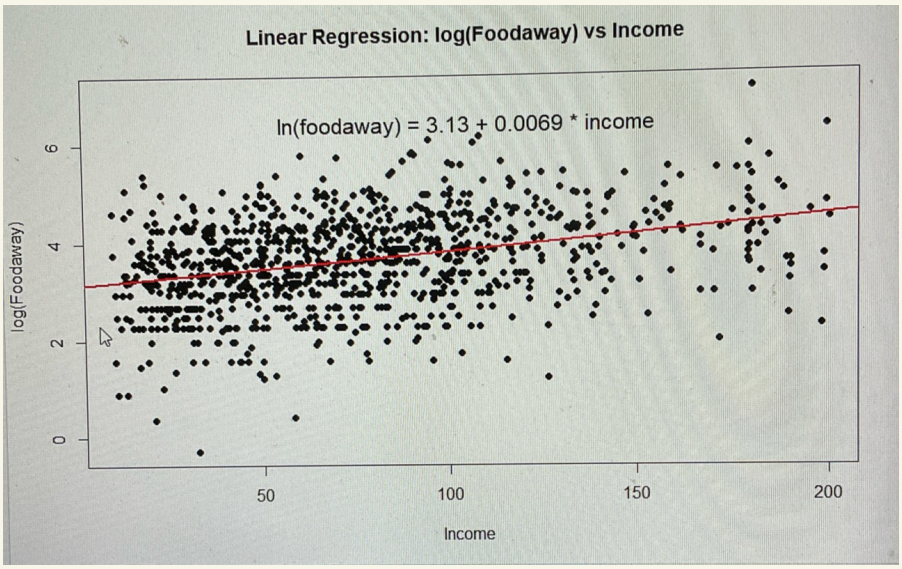
有178個家庭花費0元在Foodaway上

∵ ln(0) is not defined

∴這178個值無法用在迴歸上

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

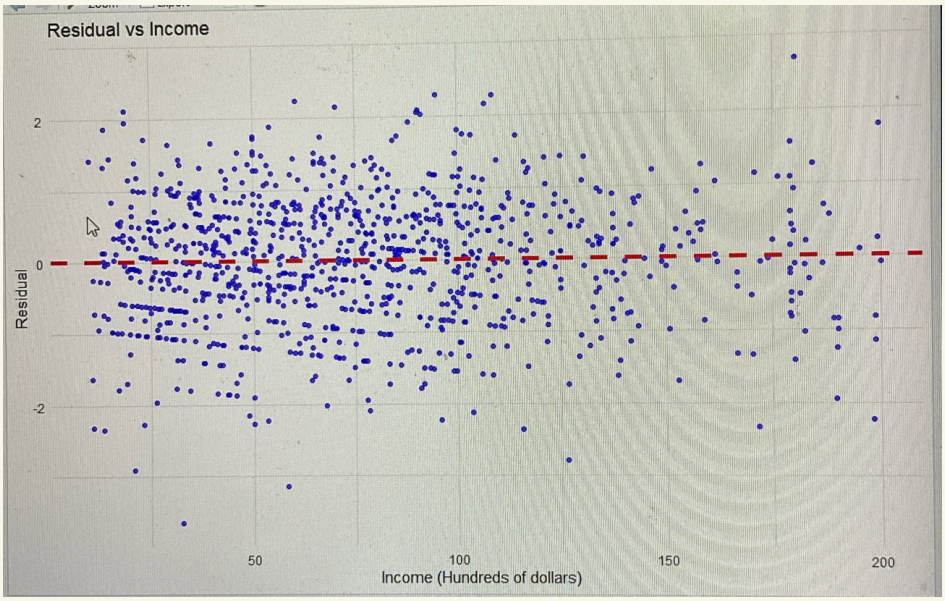
e. Plot $\ln(FOODAWAY)$ against *INCOME*, and include the fitted line from part (d).



model: $\ln(\text{foodaway}) = 3.13 + 0.0069\ \text{income}$

其他條件相同時，每增加100 income，會增加0.69%的 ln(foodaway) 的花費

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?



殘差圖看起來是隨機沒有特定模式的

2.25 R

```r
summary_stats <- data.frame(
  Mean = mean(cex5_small$foodaway, na.rm = TRUE),
  Median = median(cex5_small$foodaway, na.rm = TRUE),
  Q25 = quantile(cex5_small$foodaway, 0.25, na.rm = TRUE),
  Q75 = quantile(cex5_small$foodaway, 0.75, na.rm = TRUE)
)
print(summary_stats)
ggplot(cex5_small, aes(x = foodaway)) +
  geom_histogram(color = "black", fill = "green", bins = 30, alpha = 0.7) +
  labs(
    title = "Histogram of Foodaway Expenditure",
    x = "Food Away from Home ($ per month per person)",
    y = "Frequency"
  ) +
  theme_minimal()


# b.創建category
cex5_small$category <- ifelse(cex5_small$advanced == 1, "Advanced Degree",
                              ifelse(cex5_small$college == 1 & cex5_small$advanced == 0, "College Degree",
                                     "No College/Advanced Degree"))

summary_stats <- aggregate(foodaway ~ category, data = cex5_small,
                           FUN = function(x) c(Mean = mean(x, na.rm = TRUE), Median = median(x, na.rm = TRUE)))

# 計算每個category的數量N
category_counts <- as.data.frame(table(cex5_small$category))
colnames(category_counts) <- c("category", "N")

# 合併數量N到summary_states
summary_stats <- merge(summary_stats, category_counts, by = "category")
print(summary_stats)

#c. 計算log(foodaway) 對於foodaway>0的值進行轉換；其他為NA
cex5_small$log_foodaway <- ifelse(cex5_small$foodaway > 0, log(cex5_small$foodaway), NA)
Mean_log <- mean(cex5_small$log_foodaway, na.rm = TRUE)
Median_log <- median(cex5_small$log_foodaway, na.rm = TRUE)
Q25_log <- quantile(cex5_small$log_foodaway, 0.25, na.rm = TRUE)
Q75_log <- quantile(cex5_small$log_foodaway, 0.75, na.rm = TRUE)
Count_Log <- sum(!is.na(cex5_small$log_foodaway))#計算有效值的數量
Count_Original <- sum(!is.na(cex5_small$foodaway))
summary_log <- data.frame(
  Mean = Mean_log,
  Median = Median_log,
  Q25 = Q25_log,
  Q75 = Q75_log,
  Count_Log = Count_Log,
  Count_Original = Count_Original
)
print(summary_log)
hist(cex5_small$log_foodaway, breaks = 30, col = "GREEN", border = "black",
     xlab = "ln(Food Away from Home)", ylab = "Frequency",
     main = "Histogram of ln(foodaway) Expenditure")

#d. e.
cex5_clean <- cex5_small[!is.na(cex5_small$log_foodaway) & !is.na(cex5_small$income), ]#清理數據 去除NA
model <- lm(log_foodaway ~ income, data = cex5_clean)
summary(model)
intercept <- coef(model)[1]
```

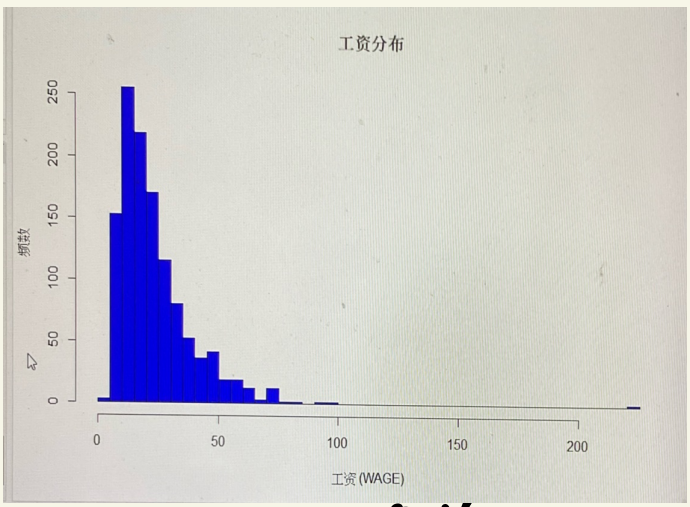
```r
slope <- coef(model)[2]
regression_eq <- paste0("ln(foodaway) = ", round(intercept, 2),
                        " + ", round(slope, 4), " * income")
x_pos <- max(cex5_clean$income) - 0.5 * (max(cex5_clean$income) - min(cex5_clean$income))  #調整位置
y_pos <- max(cex5_clean$log_foodaway) * 0.9
print(regression_eq)
plot(cex5_clean$income, cex5_clean$log_foodaway,
     main = "Linear Regression: log(Foodaway) vs Income",
     xlab = "Income", ylab = "log(Foodaway)",
     pch = 19, col = "black")
abline(model, col = "red", lwd = 2)
text(x_pos, y_pos, labels = regression_eq, col = "black", cex = 1.3)

#f,
cex5_clean$residual <- resid(model)
ggplot(cex5_clean, aes(x = income, y = residual)) +
  geom_point(colour = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed", size = 1.5 ) + #添加紅色虛線表示0殘差的基準線
  labs(
    title = "Residual vs Income",
    x = "Income (Hundreds of dollars)",
    y = "Residual"
  ) +
  theme_minimal()
```

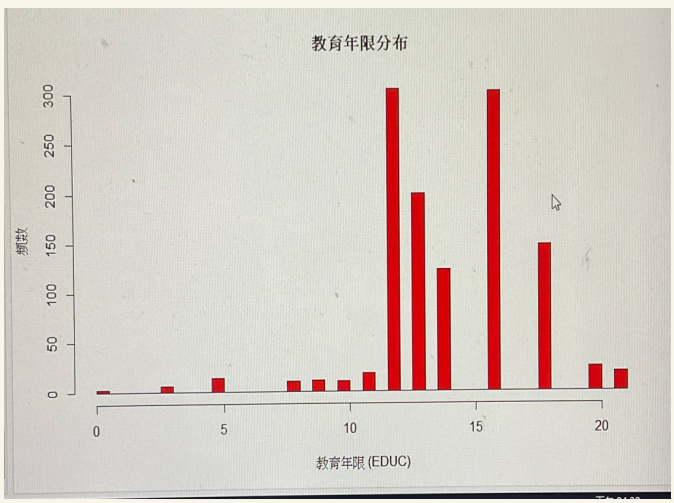
rate which repairs were being made?

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.

```
print(summary_stats)
Mean_wage Median_wage  SD_wage Min_wage Max_wage Mean_educ Median_educ  SD_educ Min_educ Max_educ
 23.64004        19.3 15.21655     3.94    221.1   14.2025          14 2.890811        0       21
```





平均工資為23.6但中位數僅為19.3
工資分佈呈現右偏，並且標準差較大

直方圖顯示大多數人工資較低，少部分工資很高，有工資不均衡的情況

平均教育年限為14 中位數也是14，分佈較為對稱，直方圖顯示大部分人受教程度集中在12~16年之間

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

```
> print(regression_eq)  # 输出回归方程
[1] "Wage = -10.40 + 2.40 * Educ"
>
```

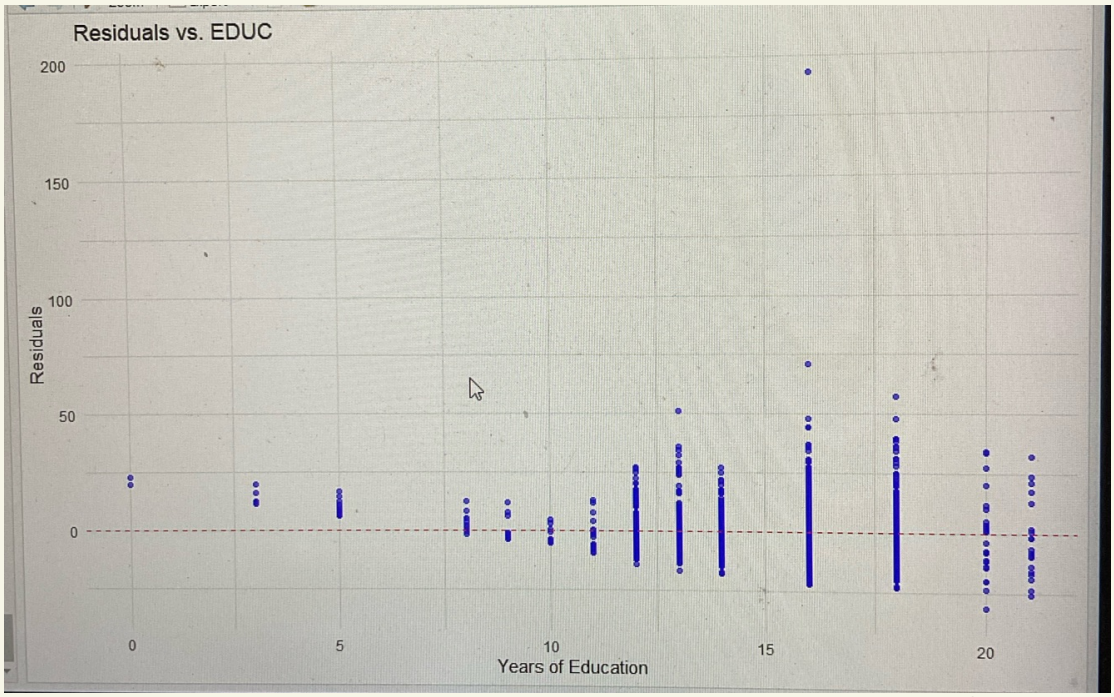
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.4000     1.9624    -5.3 1.38e-07 ***
educ          2.3968     0.1354    17.7  < 2e-16 ***
---
```

model: $\widehat{Wage} = -10.4 + 2.4\,Educ$

$\beta_0 = 2.4$：每多一年教育，工資平均增加2.4

p值(< 2e-16)：educ對wage具有顯著統計關係

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?



EDUC↑，殘差也變大，不滿足同質變異數(SR5)。

如果滿足SR1~SR5，殘差應該在各個EDUC水平下均勻分佈

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

males: $\widehat{WAGE} = -8.2849 + 2.3785\,EDUC$

females: $\widehat{WAGE} = -16.6028 + 2.6595\,EDUC$

whites: $\widehat{WAGE} = -10.4715 + 2.418\ EDUC$

blacks: $\widehat{WAGE} = -6.2541 + 1.9233\ EDUC$

女性的教育回報(2.6595)>男性的教育回報(2.3785).說明每多一年教育,女性能獲得更多的薪資增幅,女性截距低於男性,說明在低教育水準下,女性薪資較男性低

白人的教育回報(2.418)>黑人的教育回報(1.9233).說明每多一年教育,白人能獲得更多的薪資增幅,白人截距低於黑人,說明在低教育水準下,黑人薪資較白人高

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 ***
```

$\widehat{WAGE} = 4.916477 + 0.089134\ EDUC^2$

marginal effect $= 2 \times 0.089134 \times EDUC$

$EDUC = 12$, margin $= 2 \times 0.089134 \times 12 = 2.1392$

$EDUC = 16$, margin $= 2 \times 0.089134 \times 16 = 2.8523$

from (b) $\widehat{WAGE} = -10.4 + 2.4\ EDUC$

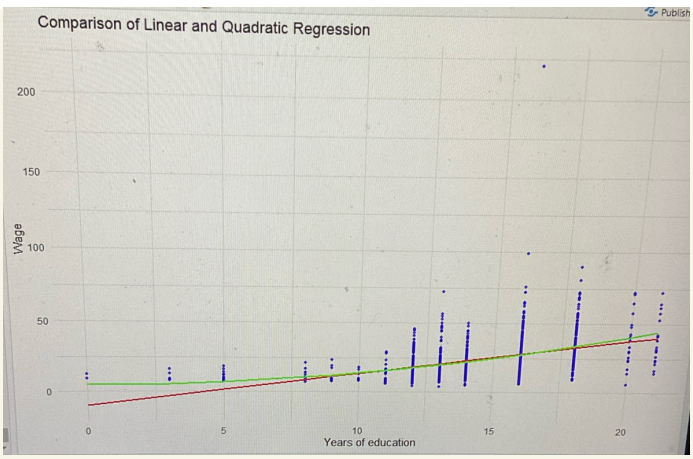
marginal effect $= 2.4$

在EDUC=12時，線性迴歸模型對工資影響更大

在EDUC=16時，二次迴歸模型對工資影響更大

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?



二次迴歸模型更貼近散點分佈，更合適

2.28 R

```r
#a.
head(cps5_small)
summary_stats <- data.frame(
  Mean_wage = mean(cps5_small$wage, na.rm = TRUE),
  Median_wage = median(cps5_small$wage, na.rm = TRUE),
  SD_wage = sd(cps5_small$wage, na.rm = TRUE),
  Min_wage = min(cps5_small$wage, na.rm = TRUE),
  Max_wage = max(cps5_small$wage, na.rm = TRUE),
  Mean_educ = mean(cps5_small$educ, na.rm = TRUE),
  Median_educ = median(cps5_small$educ, na.rm = TRUE),
  SD_educ = sd(cps5_small$educ, na.rm = TRUE),
  Min_educ = min(cps5_small$educ, na.rm = TRUE),
  Max_educ = max(cps5_small$educ, na.rm = TRUE)
)

print(summary_stats)
# 工资 (WAGE) 直方图
hist(cps5_small$wage, breaks = 50, col = "blue", main = "工资分布", xlab = "工资 (WAGE)", ylab = "频数")

# 教育年限 (EDUC) 直方图
hist(cps5_small$educ, breaks = 50, col = "red", main = "教育年限分布", xlab = "教育年限 (EDUC)", ylab = "频数")

#b.
model <- lm(wage ~ educ, data = cps5_small)
summary(model)
intercept <- coef(model)[1]
slope <- coef(model)[2]
regression_eq <- sprintf("Wage = %.2f + %.2f * Educ", intercept, slope)
print(regression_eq)  # 输出回归方程

#c.
cps5_small$residuals <- residuals(model)
ggplot(cps5_small, aes(x = educ, y = residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. EDUC",
       x = "Years of Education",
       y = "Residuals") +
  theme_minimal()
#d.
male_model <- lm(wage ~ educ, data = cps5_small[cps5_small$female == 0, ])
female_model <- lm(wage ~ educ, data = cps5_small[cps5_small$female == 1, ])
white_model <- lm(wage ~ educ, data = cps5_small[cps5_small$black == 0, ])
black_model <- lm(wage ~ educ, data = cps5_small[cps5_small$black == 1, ])
summary(male_model)
summary(female_model)
summary(white_model)
summary(black_model)

#e.
quad_model <- lm(wage ~ I(educ^2), data = cps5_small)
cps5_small$cps5_small_prdct <- predict(quad_model)
summary(quad_model)

#f.
linear_model = lm(wage ~ educ, data = cps5_small)
quadratic_model = lm(wage ~ I(educ^2), data = cps5_small)
cps5_small$cps5_linear <- predict(linear_model)
cps5_small$cps5_quadratic <- predict(quadratic_model)

ggplot(cps5_small, aes(x = educ, y = wage)) +
  geom_point(color = "blue", size = 1, alpha = 0.6) +
  geom_line(aes(y = cps5_linear), color = "red",size = 1 ) +
  geom_line(aes(y = cps5_quadratic), color = "green",size = 1 ) +
  labs(title = "Comparison of Linear and Quadratic Regression",
       x = "Years of education",
       y = "Wage"
  ) +
  theme_minimal()
```