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.

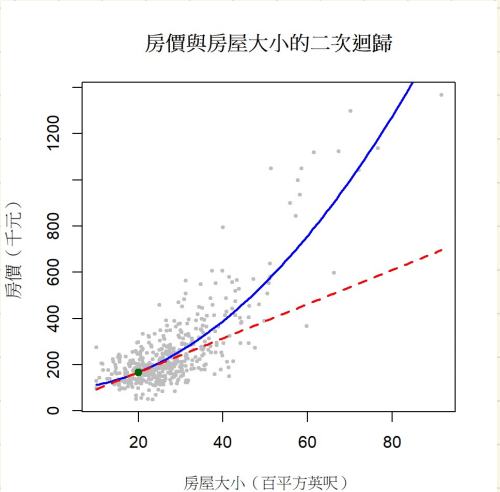
e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

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?

a. b. price = -155.42 + 13.4 SQFT

c. d.



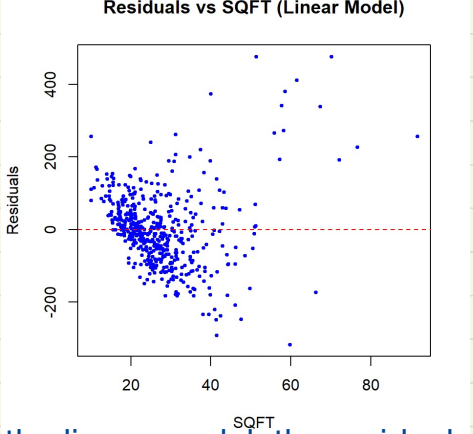
price = 93.57 + 0.185 SQFT²

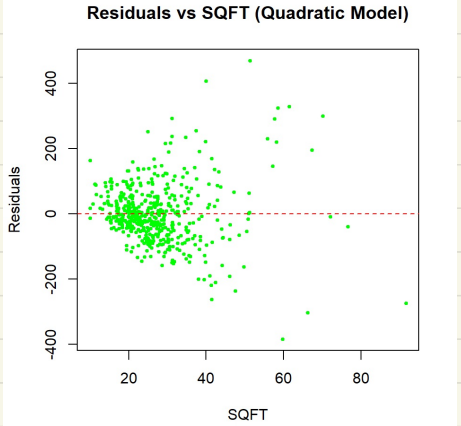
marginal effect ⇒ $\frac{d\,PRICE}{d\,SQFT} = 2\alpha_2 SQFT$

$= 2 \times 0.185 \times 20 = 7.38$

e. $\varepsilon = \frac{d\,PRICE}{d\,SQFT} \times \frac{SQFT}{PRICE} = 7.38 \times \frac{20}{167.17} = 0.882$

f.






```
> SSE_linear
[1] 5262847
> SSE_quad
[1] 4222356
```

In the linear model, the residuals exhibit heteroskedasticity, with the variance of residuals increasing as SQFT increases. There also appears to be a systematic pattern in the residuals, suggesting that the linear specification may be inappropriate.

Even in the quadratic model, the residuals appear to fan out as SQFT increases, particularly beyond 40 hundred square feet (4000 sqft). This suggests that the issue of heteroskedasticity is not fully resolved, although it is less severe compared to the linear model.

g

In comparing the two models, we find that the quadratic model in (c) has a lower sum of squared residuals (SSE) than the linear model in (b). A lower SSE indicates that the predicted values from the model are, on average, closer to the observed data points. Therefore, the quadratic model provides a better in-sample fit to the data. However, we should also be cautious of overfitting, as adding complexity to a model will often reduce SSE at the cost of generalizability.

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.

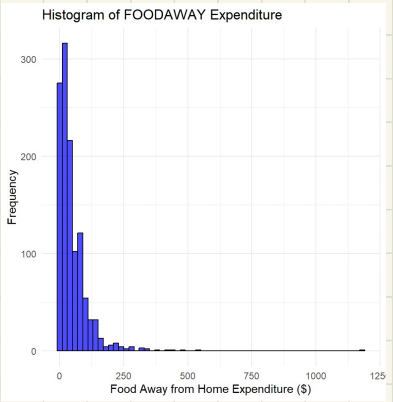
d. Estimate the linear regression $\ln(FOODAWAY) = \beta_1 + \beta_2 INCOME + e$. Interpret the estimated slope.

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.

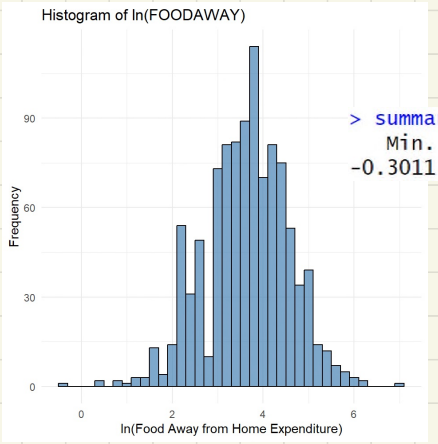
```
> summary(cex5_small$foodaway)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.04   32.55   49.27   67.50 1179.00
```



b.

```
> mean(adv_degree$foodaway)
[1] 73.15494
> median(adv_degree$foodaway)
[1] 48.15
>
> mean(college_degree$foodaway)
[1] 48.59718
> median(college_degree$foodaway)
[1] 36.11
>
> mean(no_degree$foodaway)
[1] 39.01017
> median(no_degree$foodaway)
[1] 26.02
```

c.




```
> summary(cex5_small_nonzero$ln_foodaway)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.3011   3.0759   3.6865   3.6508   4.2797   7.0724
```

ln(FOODAWAY) 值减少 178 个，
∵ 有 178 个家庭在 foodaway = 0
ln(0) 未定义 → 缺失值

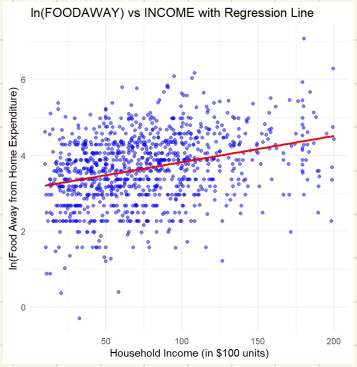
d.

$\ln(FOODAWAY) = 3.1293 + 0.0069\ INCOME$

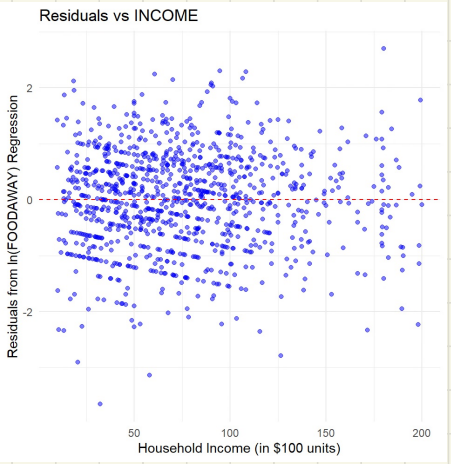
For a $100 increase in income, the log of food away increases by 0.0069

0.69%

e.



f.



They seem completely random

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?

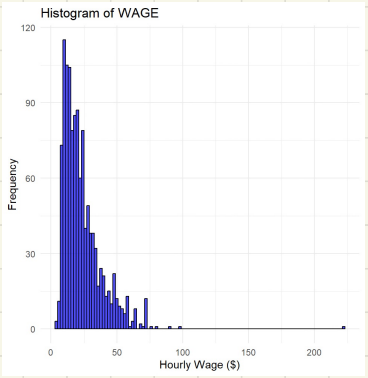
d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

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).

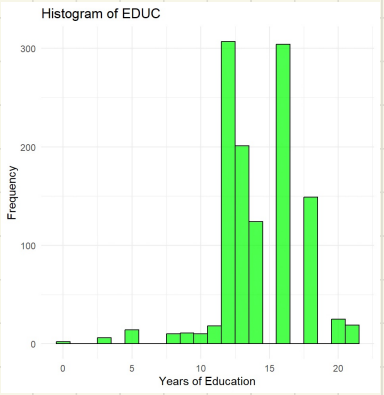
f. Plot the fitted linear model from part (b) and the fitted values from the quadratic model from part (e) in the same graph with the data on *WAGE* and *EDUC*. Which model appears to fit the data better?

a.

```
> summary(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 右偏分佈

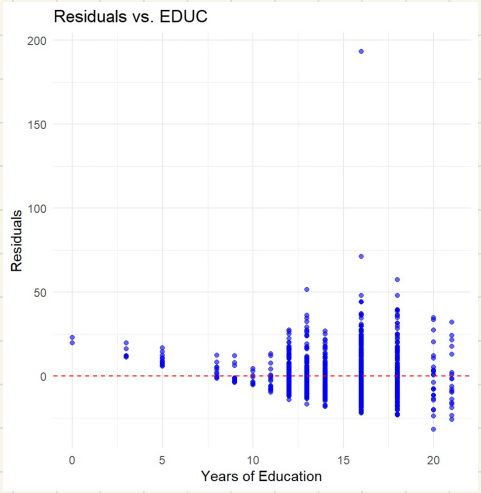


EDUC 大部分为 12~16年

b. $\widehat{WAGE} = -10.4 + 2.3968\, EDUC$

EDUC 增加1年 → WAGE 增加 2.3968 $

c.



EDUC ↑ residual ↑

⇒ 不满足 Homoskedasticity (SR5)

若 SR1-SR5 成立，残差应在各个 EDUC 下均匀分布

d.

男：wage = -8.2849 + 2.3785 educ

女：wage = -16.6028 + 2.6455 educ

] 女性 教育对薪资影响较大，但都跑低

黑：wage = -6.2541 + 1.9233 educ

白：wage = -10.4747 + 2.4178 educ

] 黑人 教育对薪资影响较大，但都跑低

e. $WAGE = 4.9165 + 0.08913\ EDUC^2$

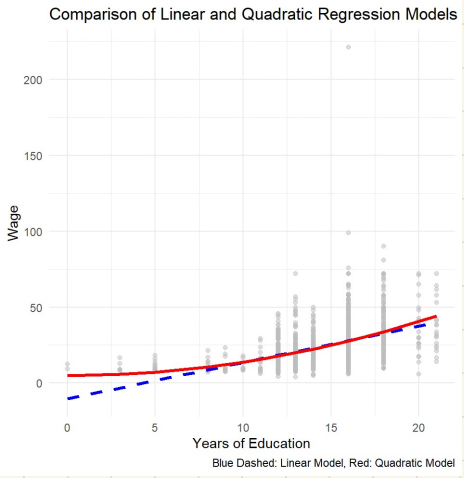
marginal effect $= 2 \times 0.08913 \times EDUC$

$EDUC = 12 \Rightarrow 2 \times 0.08913 \times 12 = 2.139$

$EDUC = 16 \Rightarrow 2 \times 0.08913 \times 16 = 2.852$

二次回歸的 marginal effect 會隨 EDUC 變大而變大

f.



quadratic regression fit the data better