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.

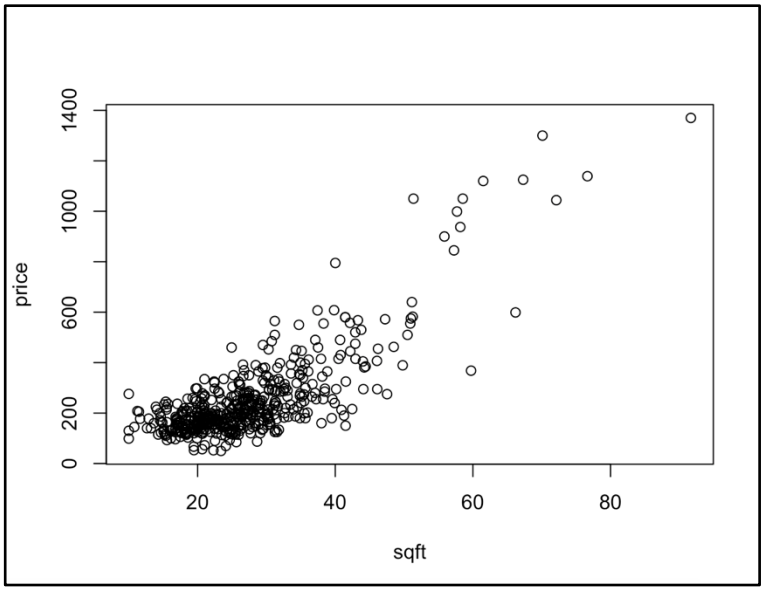
d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

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. Scatter plot (sqft, price)

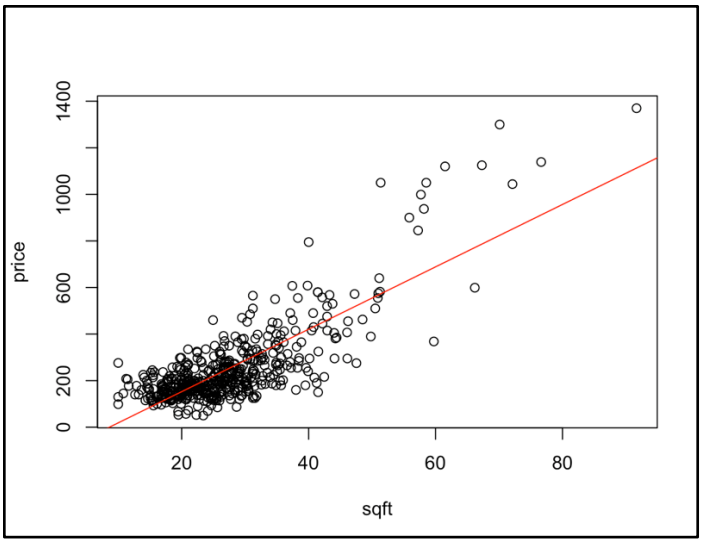


b. Estimate linear model (PRICE = β1 + β2SQFT + e)

```
> coef(ln_mod) #intercept:-115.42 x:13.40
(Intercept)           x
 -115.42361    13.40294
```

β1 = -115.4236 當房子大小為 0 時，預測的價格是-115.4236（千元）

β2 = 13.40294　房子大小每增加一單位（百 square feet），預測價格增加 13.40294（千元）



c. Quadratic model (PRICE = α1 + α2 SQFT^2 + e)

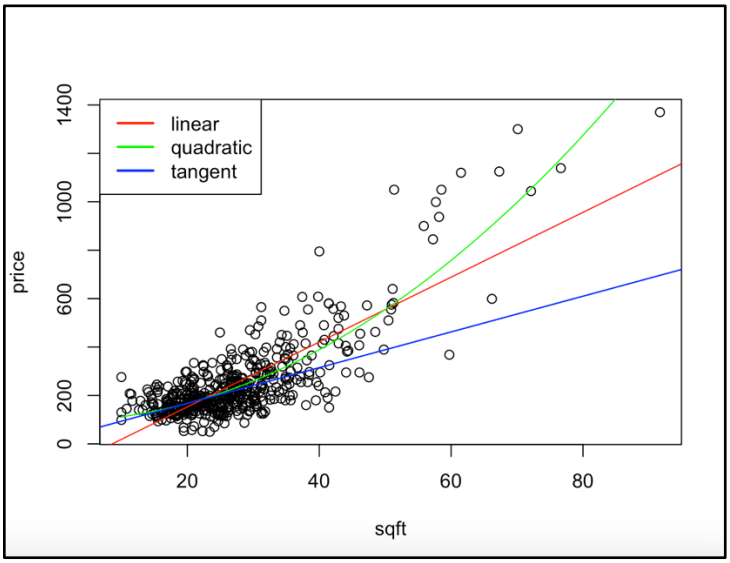
```
> coef(qd_mod) #intercept:93.57 I(x^2):0.18
(Intercept)      I(x^2)
  93.565854    0.184519
```

α1 = 93.565843

α2 = 0.184519

marginal effect = (2*α2*SQFT) = 7.38076 (20 單位 sqft)

d. Fitted curve and tangent to the curve for a 2000-square-foot house.



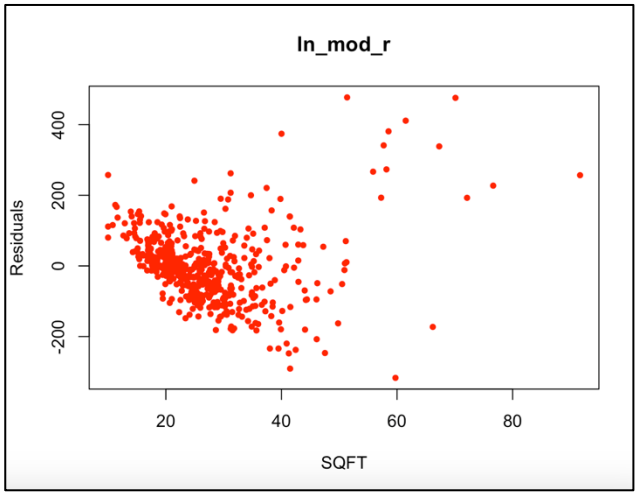
e. Elasticity 2000-square-foot house (2*α2*SQFT*SQFT/price)

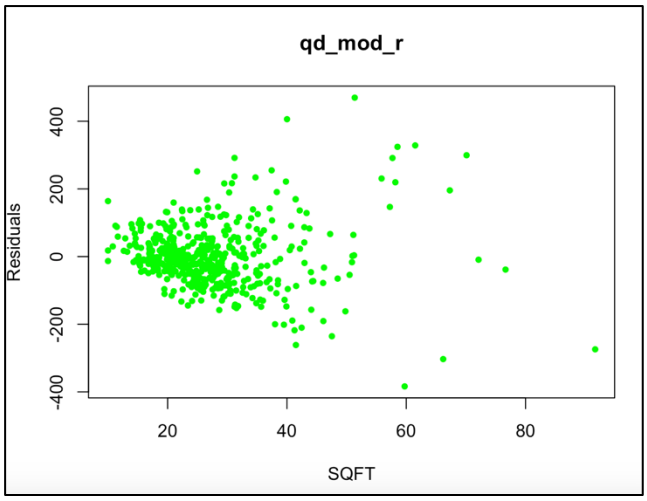
sqft = 20

price = α1 + α2*(sqft^2)

elasticity = mg_ef_20sqft*sqft/price = 0.8819511

f. Residuals under different model against sqft





觀察兩種模型的殘差對應房子大小，可以發現在線性模型下隨著房子變大，殘差的期望值越來越偏離 0，殘差變異數並非不變而有上升的趨勢。在二次模型下，即使房子變大，殘差的期望值仍接近 0，但如同線性模型，殘差變異數在 sqft 30 到 60 間有上升的趨勢。

g. SSE

```
> ln_r_sse
[1] 5262847
> qd_r_sse
[1] 4222356
```

Linear_SSE = 5262847, Quadratic_SSE = 4222356.

Quadratic model 有較小的 SSE, 代表用 Quadratic model 預測的值與實際值的誤差平方和較小，較為準確。

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?

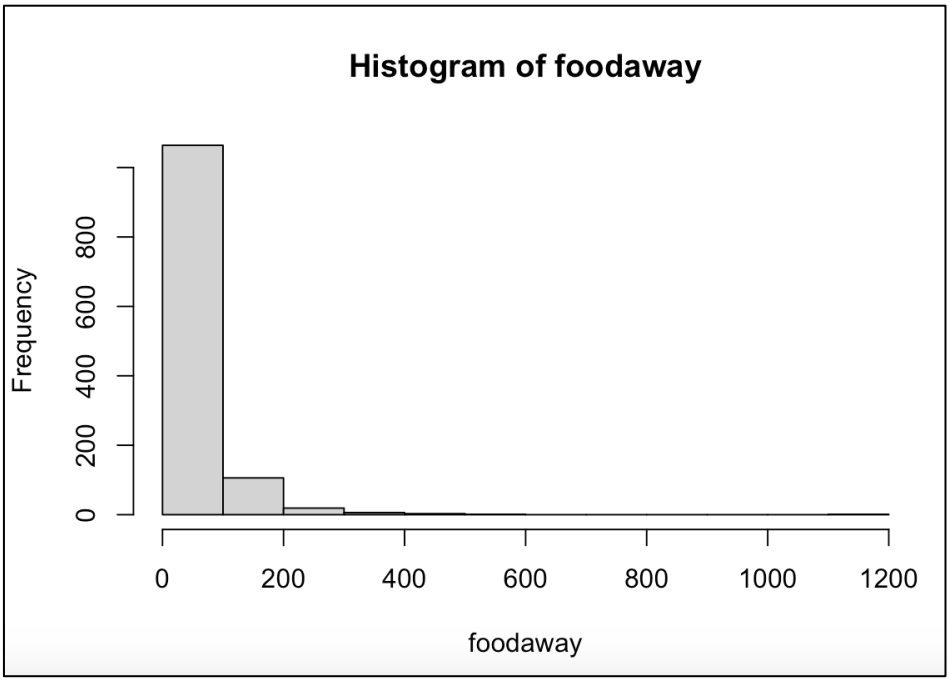
c. Construct a histogram of ln(*FOODAWAY*) and its summary statistics. Explain why *FOODAWAY* and ln(*FOODAWAY*) have different numbers of observations.

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. histogram 資料呈右偏情形, summary




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

25th percentiles: 12.04, Mean: 49.27, 75th percentiles: 67.5, Median: 32.55

b. Foodaway under advanced degree, college degree, without both degrees.

Advanced

```
> mean(ad) #73.15494
[1] 73.15494
> median(ad) #48.15
[1] 48.15
```

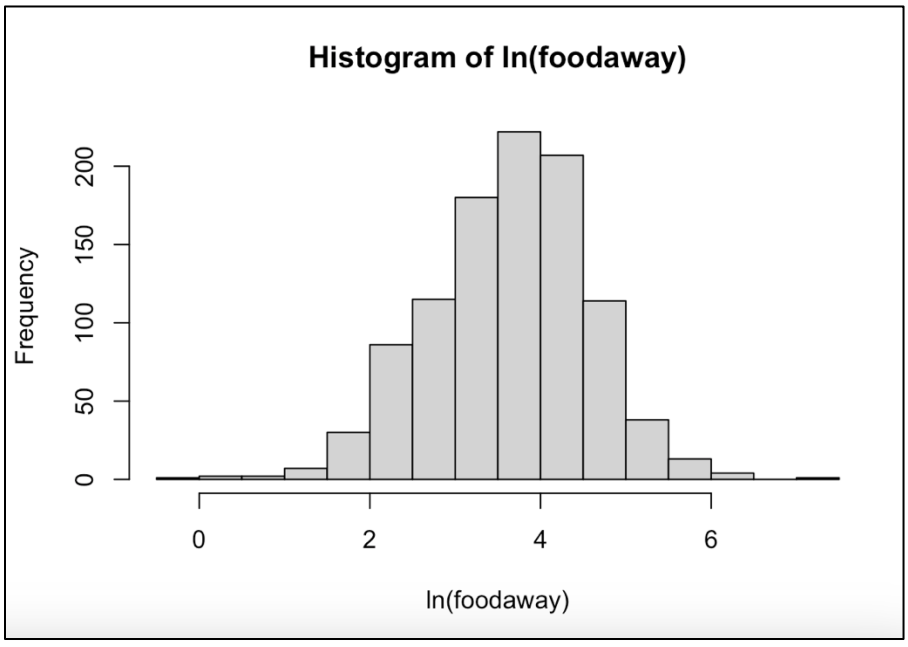
College

```
> mean(clg) #48.59718
[1] 48.59718
> median(clg) #36.11
[1] 36.11
```

Without both

```
> mean(both_no) #39.01017
[1] 39.01017
> median(both_no) #26.02
[1] 26.02
```

c. Ln(foodaway) histogram，因為有些原始 foodaway 數值是 0，會造成 ln(0)未定義，所以需要將 foodaway = 0 的直先拿掉。




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

25th percentiles: 3.0759, Mean: 3.6508, 75th percentiles: 4.2797, Median: 3.6865

d. Linear regression ln(FOODAWAY) = β1 + β2INCOME + e

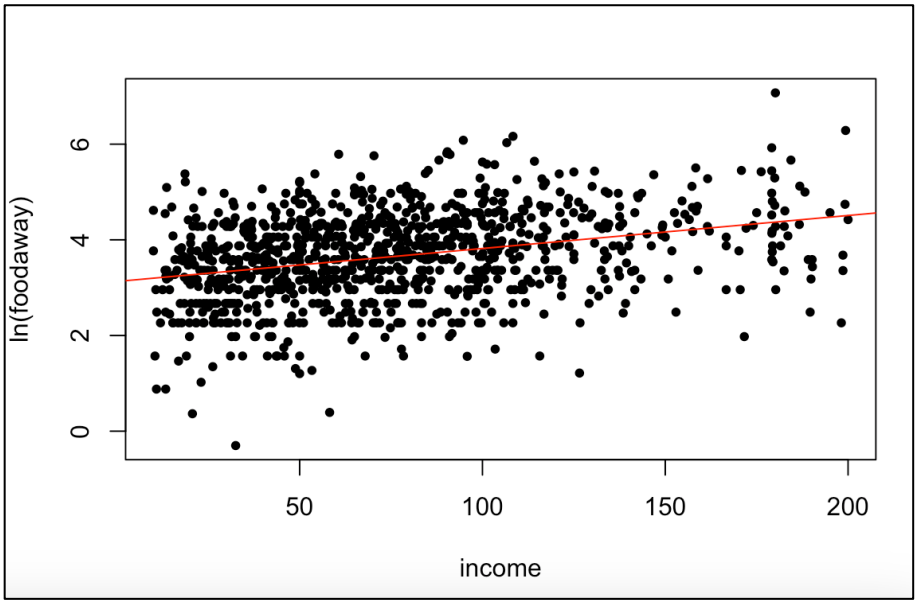
```
(Intercept)           x
   3.129300    0.006902
```

Estimated slope = 0.006902

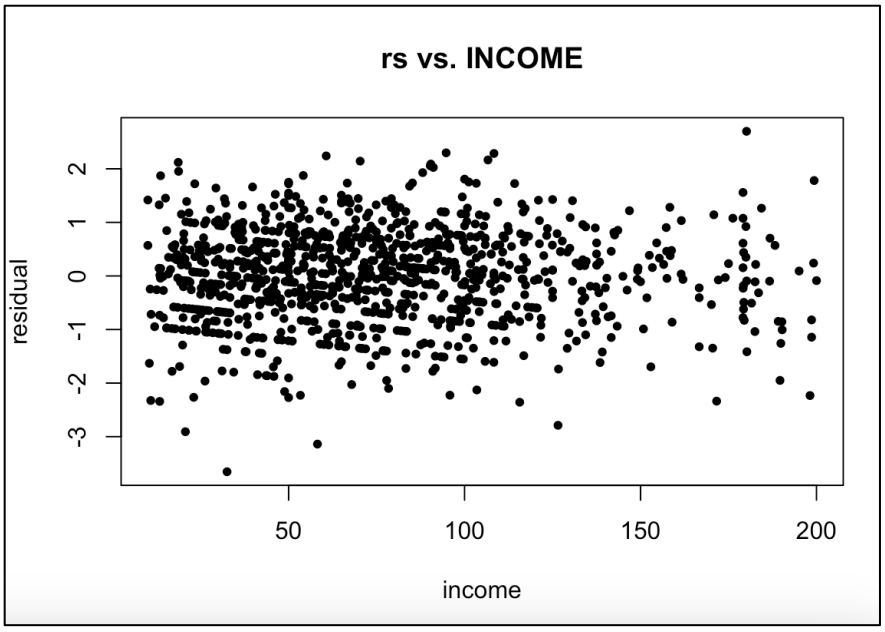
每增加一單位（100）income, ln(foodaway) 增加 0.0069 元（每月每人）

也就是 fooraway 增加 exp(0.0069) = 1.006924

e. Plot ln(FOODAWAY) = β1 + β2INCOME + e



f. Residuals v.s. Income 看起來隨機



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?

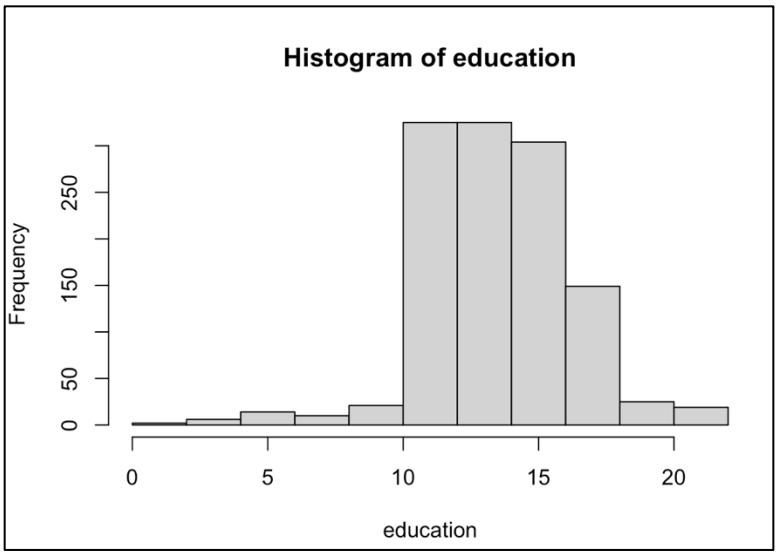
d. Estimate separate regressions for males, females, blacks, and whites. Compare the results.

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. Education

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



Wage

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

Wage 跟 Education 資料皆有左偏的情形。

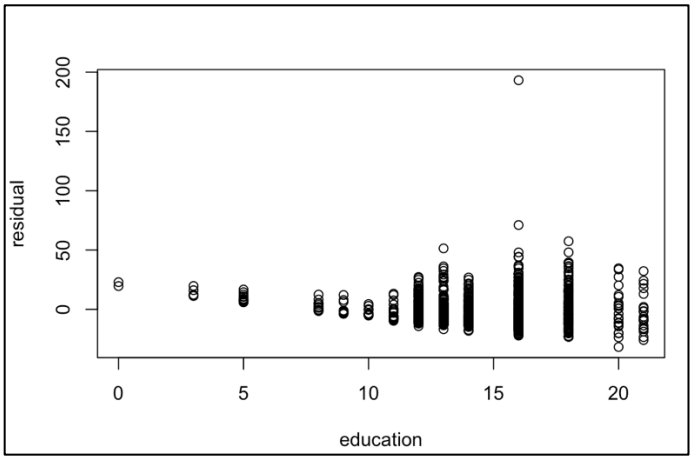
b. Linear model (WAGE = β1 + β2EDUC + e)

```
> coef(ln_model)
(Intercept)           x
 -10.399959    2.396761
```

β1 = -10.40 若沒受過教育，預測薪資為 -10.40

β2 = 2.40 教育程度增加一級，薪資提升 2.4

c. Residual v.s. Education 殘差的變異數在教育程度 10~16 區間有上升趨勢，且在許多教育程度下，殘差平均不是 0。If SR1~SR5 hold, there should not be patterns in the residuals.



d. 在未受教育時，男生相較女生有較好的薪資，但是女生接受教育對於薪資提升的幫助比男生大。

Male

```
(Intercept)        male_x
     -8.285         2.378
```

Female

```
(Intercept)      female_x
    -16.603         2.659
```

在未受教育時，黑人相較白人有較好的薪資，但是白人接受教育對於薪資提升的幫助比黑人大。

Black

```
(Intercept)        black_x
     -6.254         1.923
```

White

```
(Intercept)        white_x
    -10.475         2.418
```

e. quadratic (WAGE = α1 + α2EDUC^2 + e)

marginal effect = 2* α2*EDUC

```
> marginal_effect
[1] 2.139216 2.852288
```

使用二次模型時，12 級與 16 級教育的 marginal effect 分別是 2.14 與 2.85，表示在 16 級教育對於薪資成長的幅度較 12 級大。而 b 小題 linear model 的 marginal effect 固定是 β2 = 2.40。

f. model comparison. Quadratic model appears to fit better.