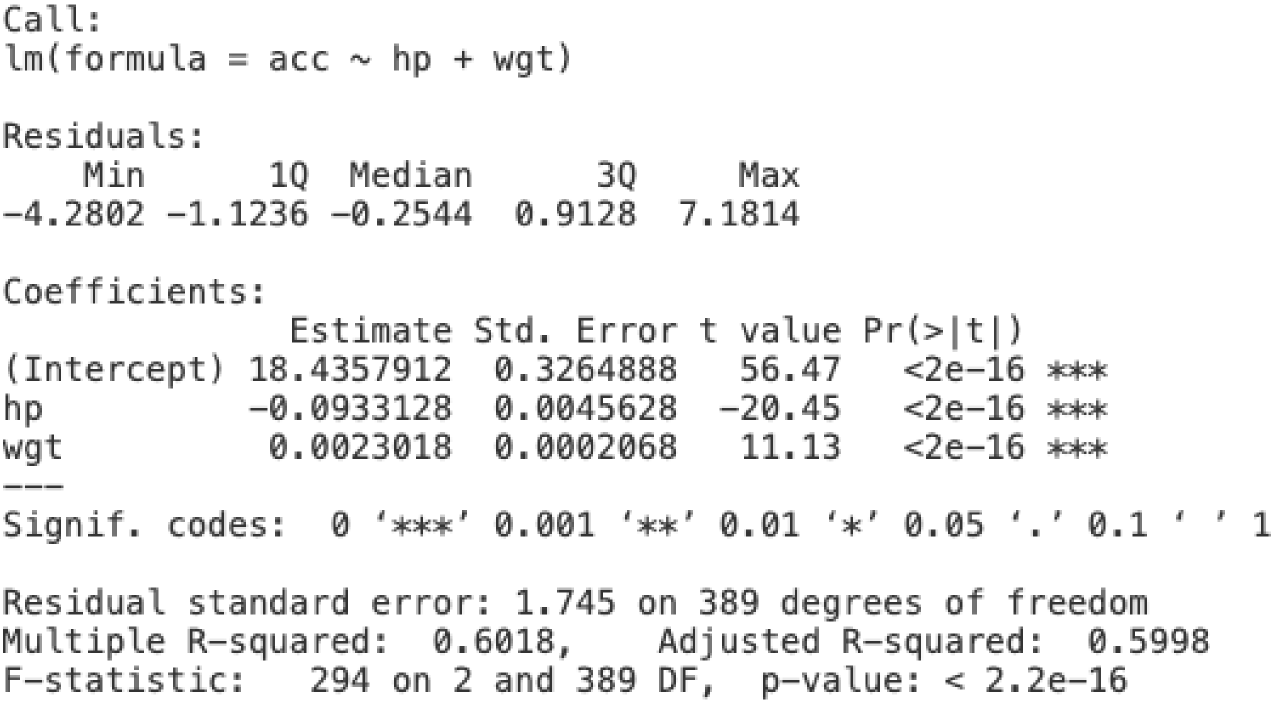
Filippo Riva
MT4420

Assignment 5

In Section 3.3.2, the authors introduce us to extensions of the linear model. The two main approaches here are to include interaction terms and to consider non-linear relationships between the response and the predictors.
1. In Assignment 4, we studied the Auto data set and found a model for acceleration in terms of horsepower and weight.
a. Find a new model that includes the interaction term (horsepower)×(weight). Include the output from the summary command, and complete the table:

CODE OLD MODEL:

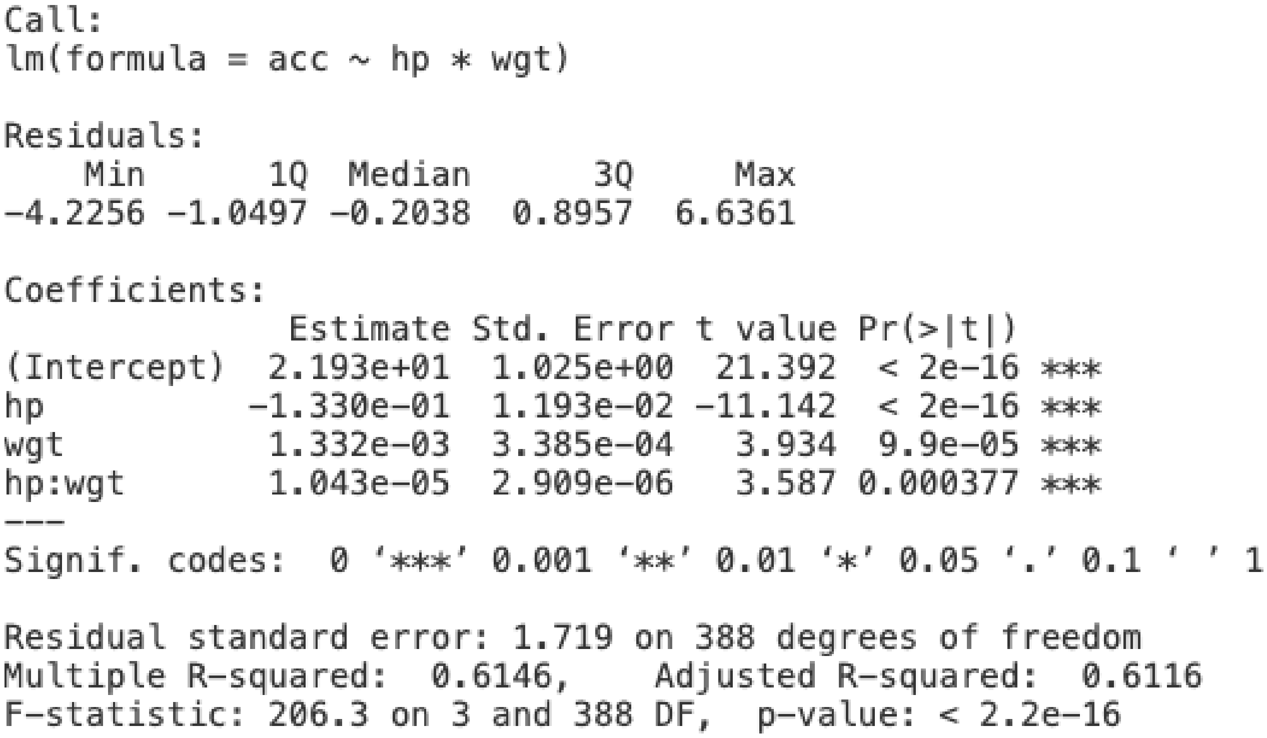
```
Auto = read.csv(file="Auto.csv",head=TRUE,sep=",",stringsAsFactors=FALSE, na.strings="?")
Auto = na.omit(Auto)
acc=Auto$acceleration
hp=Auto$horsepower
wgt=Auto$weight
Model1=lm(acc~hp+wgt)
summary(Model1)
```



```
Call:
lm(formula = acc ~ hp + wgt)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2802 -1.1236 -0.2544  0.9128  7.1814 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.4357912  0.3264888   56.47   <2e-16 ***
hp          -0.0933128  0.0045628  -20.45   <2e-16 ***
wgt          0.0023018  0.0002068   11.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.745 on 389 degrees of freedom
Multiple R-squared:  0.6018,    Adjusted R-squared:  0.5998 
F-statistic:   294 on 2 and 389 DF,  p-value: < 2.2e-16
```

CODE NEW MODEL:

```
Model2=lm(acc ~ hp * wgt)
summary(Model2)
```



```
Call:
lm(formula = acc ~ hp * wgt)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2256 -1.0497 -0.2038  0.8957  6.6361 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.193e+01  1.025e+00  21.392  < 2e-16 ***
hp          -1.330e-01  1.193e-02 -11.142  < 2e-16 ***
wgt          1.332e-03  3.385e-04   3.934  9.9e-05 ***
hp:wgt       1.043e-05  2.909e-06   3.587 0.000377 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.719 on 388 degrees of freedom
Multiple R-squared:  0.6146,	Adjusted R-squared:  0.6116 
F-statistic: 206.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

	RSE	Adjusted R^2
Old model from Assignment 4	1.745	0.5998
New model with interaction term	1.719	0.6116

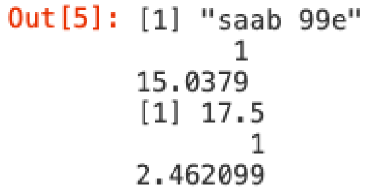
b. Based on what we know about RSE, and adjusted R^2, is our new model an improvement over the old? Explain.

Based on RSE and adjusted R^2 values, it is possible to notice an improvement in the new model since the values of RSE decreased showing a smaller average error and the value of R^2 increased showing a more linear model, conformed to the actual model.

c. Now use your old model and new model to find acceleration for the Saab 99e. Complete the table.

CODE Model 1:

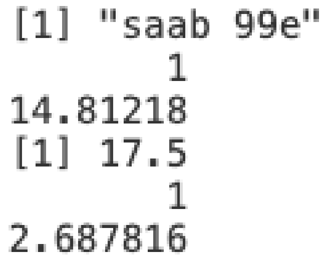
```
nam=Auto$name
print(nam[23])
p=predict(Model1data.frame(hp=horsepower[23],wgt=weight[23]))
print(p)
print(acc23])
print(abs(acc23]-p))
```



```
Out[5]: [1] "saab 99e"
               1 
        15.0379 
        [1] 17.5
               1 
        2.462099 
```

CODE Model 2:

```
nam=Auto$name
print(nam[23])
p=predict(Model2,data.frame(hp=horsepower[23],wgt=weight[23]))
print(p)
print(acc[23])
print(abs(acc[23]-p))
```



```
[1] "saab 99e"
       1
14.81218
[1] 17.5
       1
2.687816
```

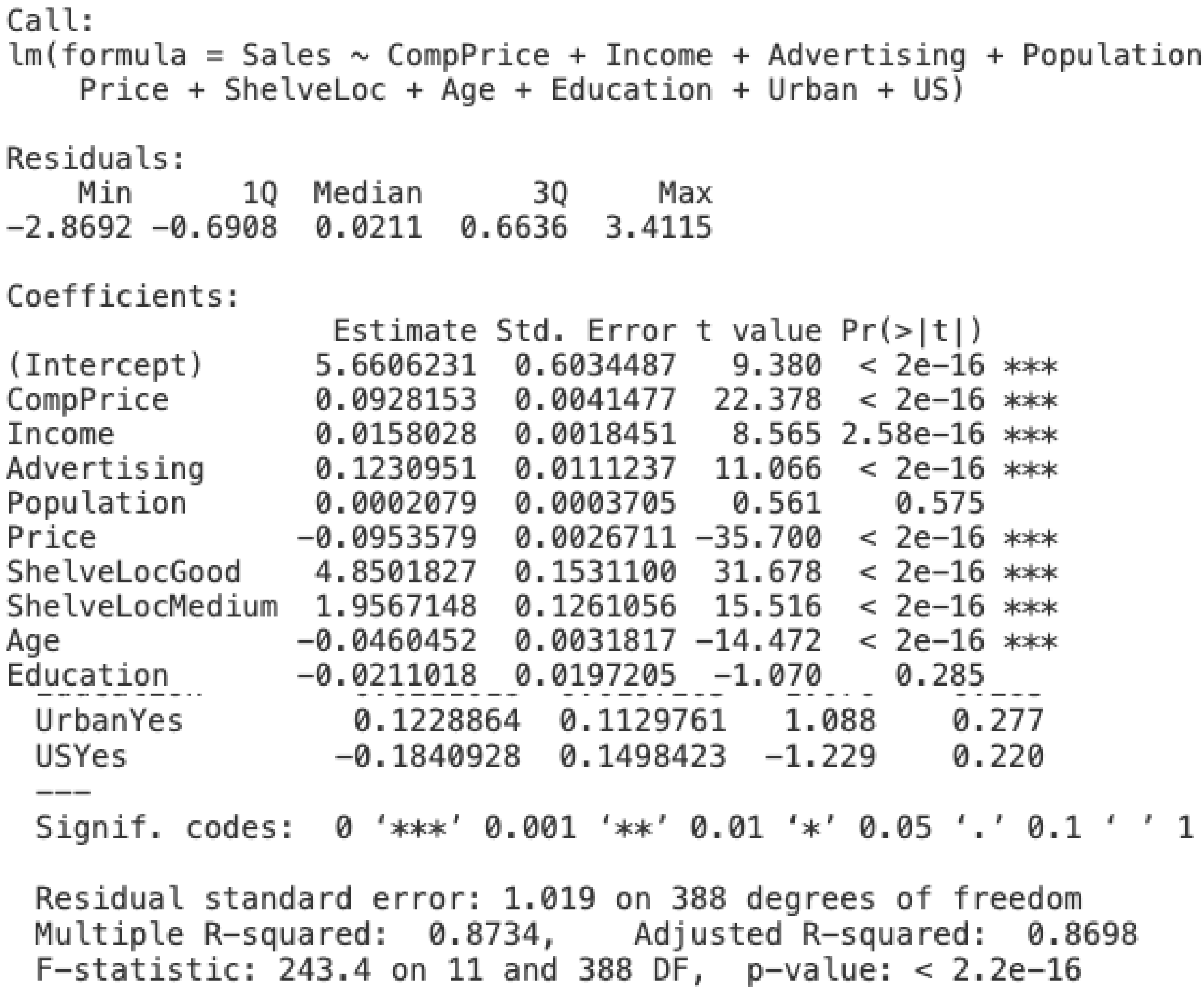
Saab 99e	Predicted Acceleration	Actual Acceleration	Absolute Error	95% Confidence interval
Using old model	15.0379	17.5	2.462099	[14.790356, 15.285446]
Using new model	14.81218	17.5	2.687816	[14.538742, 15.085625]

2. Your boss has asked you for a model that predicts sales of car seats (it is one of our data sets). Your boss wants your model to include shelf location and have at most 6 predictors. Use the variable section (your choice of approach) to find your best model. You may consider interaction terms, but don't include them unless they improve your model's RSE and adjusted R2 .

In your report, describe how you found your final model and state which variables are included. Give commands and output for your final model only.

CODE MODEL 1:

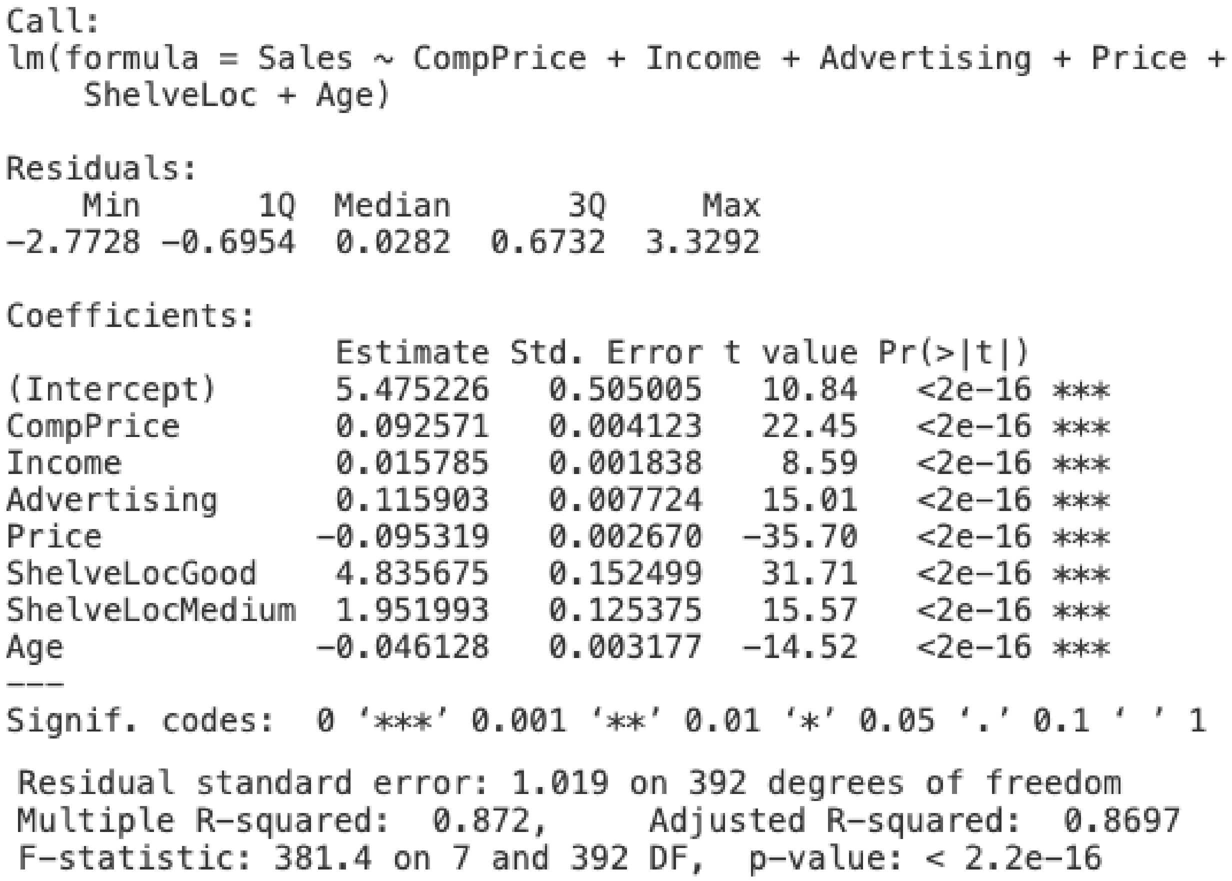
```
Carseats = read.csv(file="Carseats.csv", head=TRUE,sep=",",stringsAsFactors = FALSE)
Sales=Carseats$Sales
CompPrice=Carseats$CompPrice
Income=Carseats$Income
Advertising=Carseats$Advertising
Population=Carseats$Population
Price=Carseats$Price
ShelveLoc=Carseats$ShelveLoc
Age=Carseats$Age
Education=Carseats$Education
Urban=Carseats$Urban
US=Carseats$US
Model1=lm(Sales~CompPrice+Income+Advertising+Population+Price+ShelveLoc+Age+Education+U
rban+US)
summary(Model1)
```



```
Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Population +
    Price + ShelveLoc + Age + Education + Urban + US)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8692 -0.6908  0.0211  0.6636  3.4115 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
Income           0.0158028  0.0018451   8.565 2.58e-16 ***
Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
Population       0.0002079  0.0003705   0.561    0.575    
Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
Education       -0.0211018  0.0197205  -1.070    0.285    
UrbanYes         0.1228864  0.1129761   1.088    0.277    
USYes           -0.1840928  0.1498423  -1.229    0.220    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8698 
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

CODE MODEL 2:

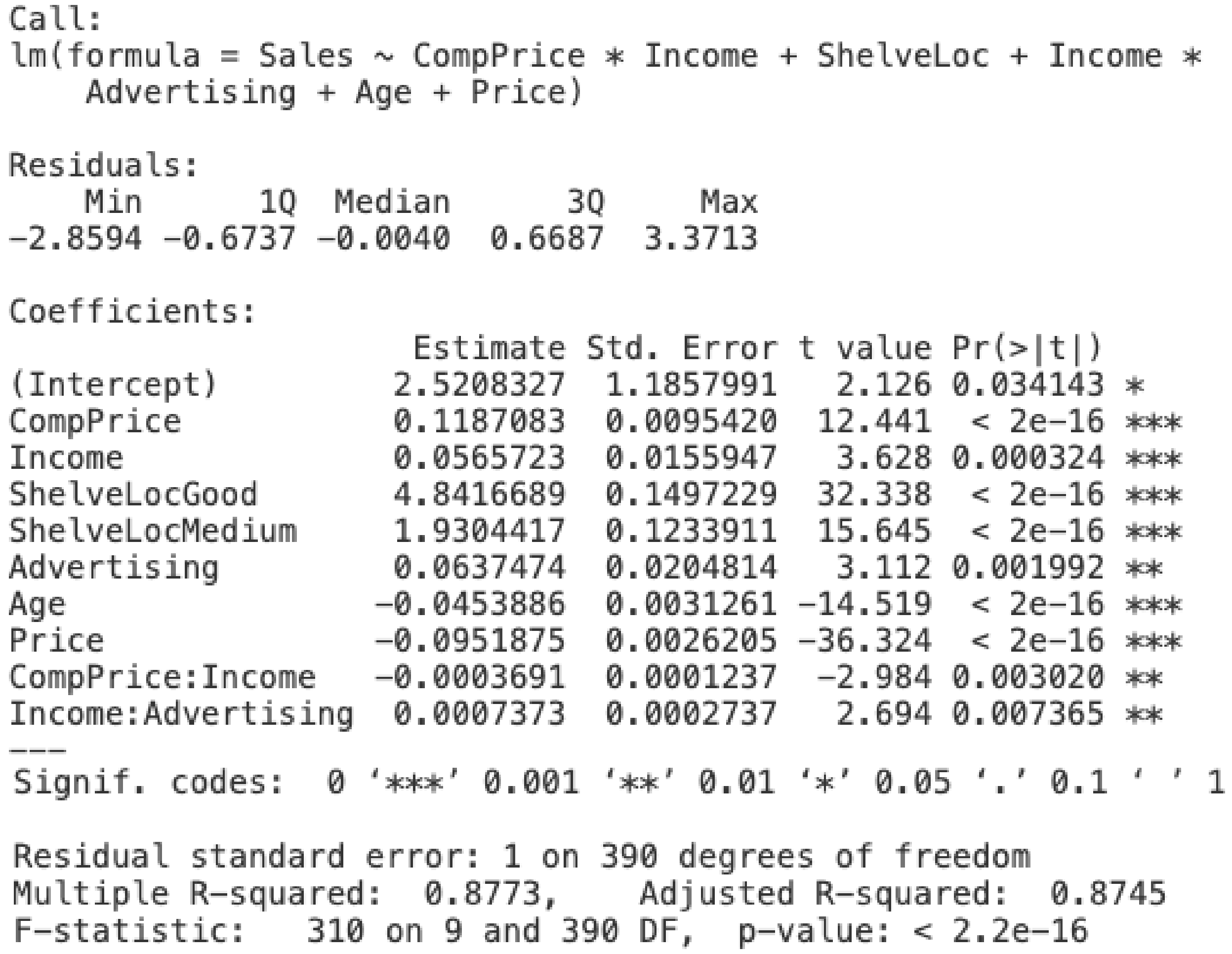
```
Model2=lm(Sales~CompPrice+Income+Advertising+Price+ShelveLoc+Age)
summary(Model2)
```



```
Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
    ShelveLoc + Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7728 -0.6954  0.0282  0.6732  3.3292 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.475226   0.505005   10.84   <2e-16 ***
CompPrice        0.092571   0.004123   22.45   <2e-16 ***
Income           0.015785   0.001838    8.59   <2e-16 ***
Advertising      0.115903   0.007724   15.01   <2e-16 ***
Price           -0.095319   0.002670  -35.70   <2e-16 ***
ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
Age             -0.046128   0.003177  -14.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 1.019 on 392 degrees of freedom
 Multiple R-squared:  0.872,	Adjusted R-squared:  0.8697 
 F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

CODE MODEL 3:

```
Model3=lm(Sales~CompPrice*Income+ShelveLoc+Income*Advertising+Age+Price)
summary(Model3)
```



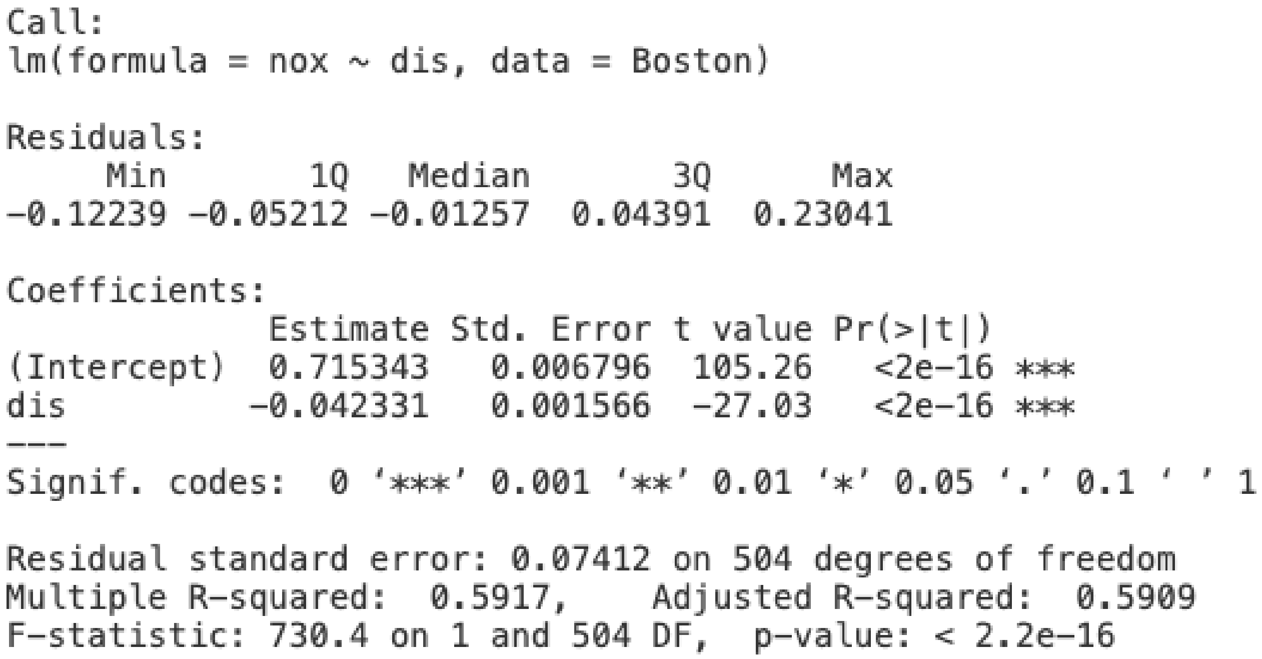
```
Call:
lm(formula = Sales ~ CompPrice * Income + ShelveLoc + Income *
    Advertising + Age + Price)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8594 -0.6737 -0.0040  0.6687  3.3713

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.5208327  1.1857991   2.126 0.034143 *
CompPrice            0.1187083  0.0095420  12.441  < 2e-16 ***
Income               0.0565723  0.0155947   3.628 0.000324 ***
ShelveLocGood        4.8416689  0.1497229  32.338  < 2e-16 ***
ShelveLocMedium      1.9304417  0.1233911  15.645  < 2e-16 ***
Advertising          0.0637474  0.0204814   3.112 0.001992 **
Age                 -0.0453886  0.0031261 -14.519  < 2e-16 ***
Price               -0.0951875  0.0026205 -36.324  < 2e-16 ***
CompPrice:Income    -0.0003691  0.0001237  -2.984 0.003020 **
Income:Advertising   0.0007373  0.0002737   2.694 0.007365 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 390 degrees of freedom
Multiple R-squared:  0.8773,    Adjusted R-squared:  0.8745
F-statistic:   310 on 9 and 390 DF,  p-value: < 2.2e-16
```

EXPLANATION OF THE PROCESS:
In the process of finding the best model that was able to predict Sales based on 6 different predictors including ShelveLoc, I decided to start by creating a Model1 where I included all the variables present in the Carseats Dataset. After creating the first model I worked backward and I decided to delete from Model1 all the predictors with a small P-value (Education, Population). After running Model2 I did not notice a big improvement, the RSE stayed pretty much the same as well as the R^2.
Once again I tried to use interaction terms between all the predictors left in Model 2 until I was finally able to find the best Model 3 which had a lower RSE and a higher R^2. (variable in model3: Income, Compprice, ShelveLoc, Advertising, Age, Price)

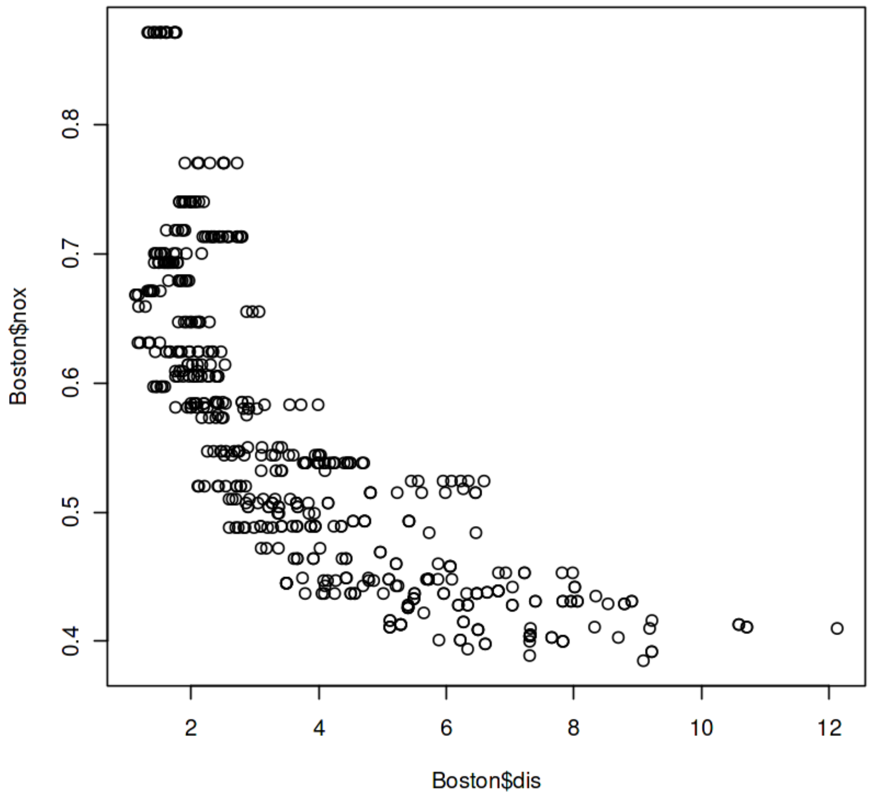
3. In the Boston data set, the variable nox describes concentrations of nitric oxides (pollutants), and the variable dis is a weighted distance to Boston employment centers (industrial areas). As you might guess, the further a community is from industrial areas, the lower the concentrations of nitric oxides. Our goal is to find models for nox as a function of dis. For each model, give the formula, include output from the summary command, and plot the model with the data. For parts a, b, and c, explain what is wrong with the model when dis is large.

a. Fit a linear model.
CODE:

```
Boston=read.csv("Boston.csv",header=T,na.strings="?",stringsAsFactors=FALSE)
Module1 = lm(nox ~ dis, data = Boston)
summary(Module1)
plot(Boston$dis, Boston$nox
```



```
Call:
lm(formula = nox ~ dis, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.12239 -0.05212 -0.01257  0.04391  0.23041 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.715343   0.006796  105.26   <2e-16 ***
dis         -0.042331   0.001566  -27.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07412 on 504 degrees of freedom
Multiple R-squared:  0.5917,	Adjusted R-squared:  0.5909 
F-statistic: 730.4 on 1 and 504 DF,  p-value: < 2.2e-16
```

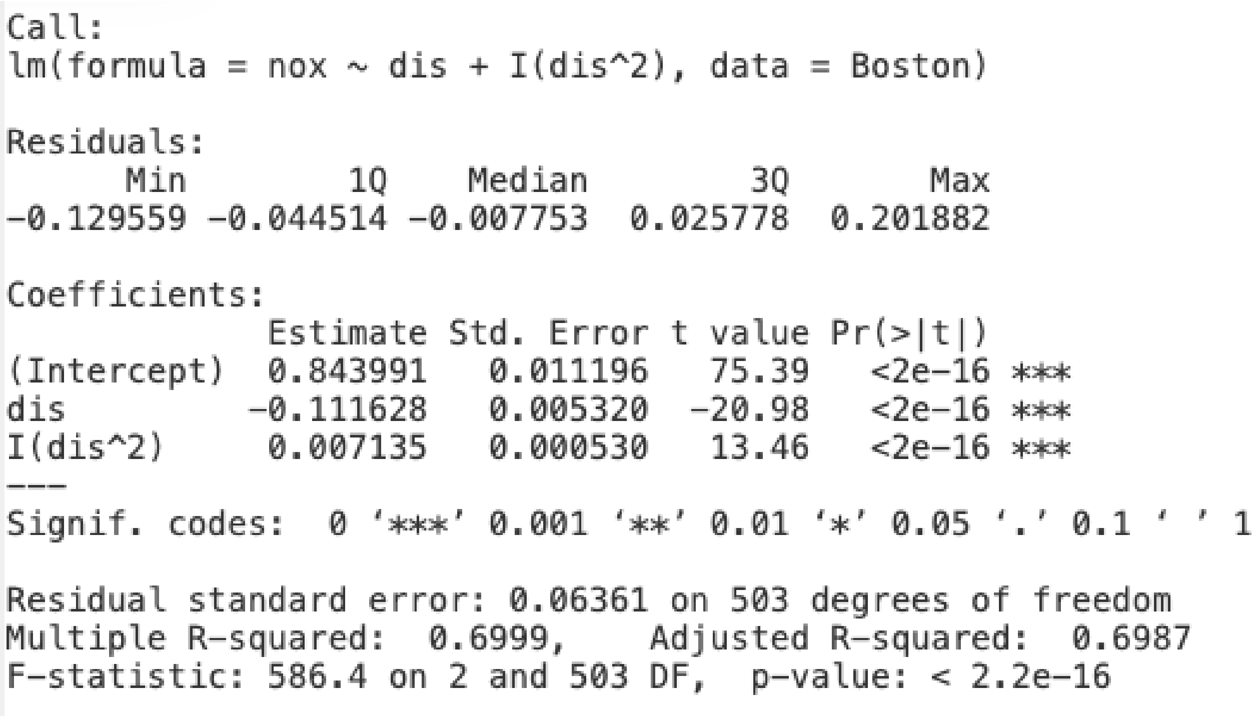


b. Fit a quadratic model.

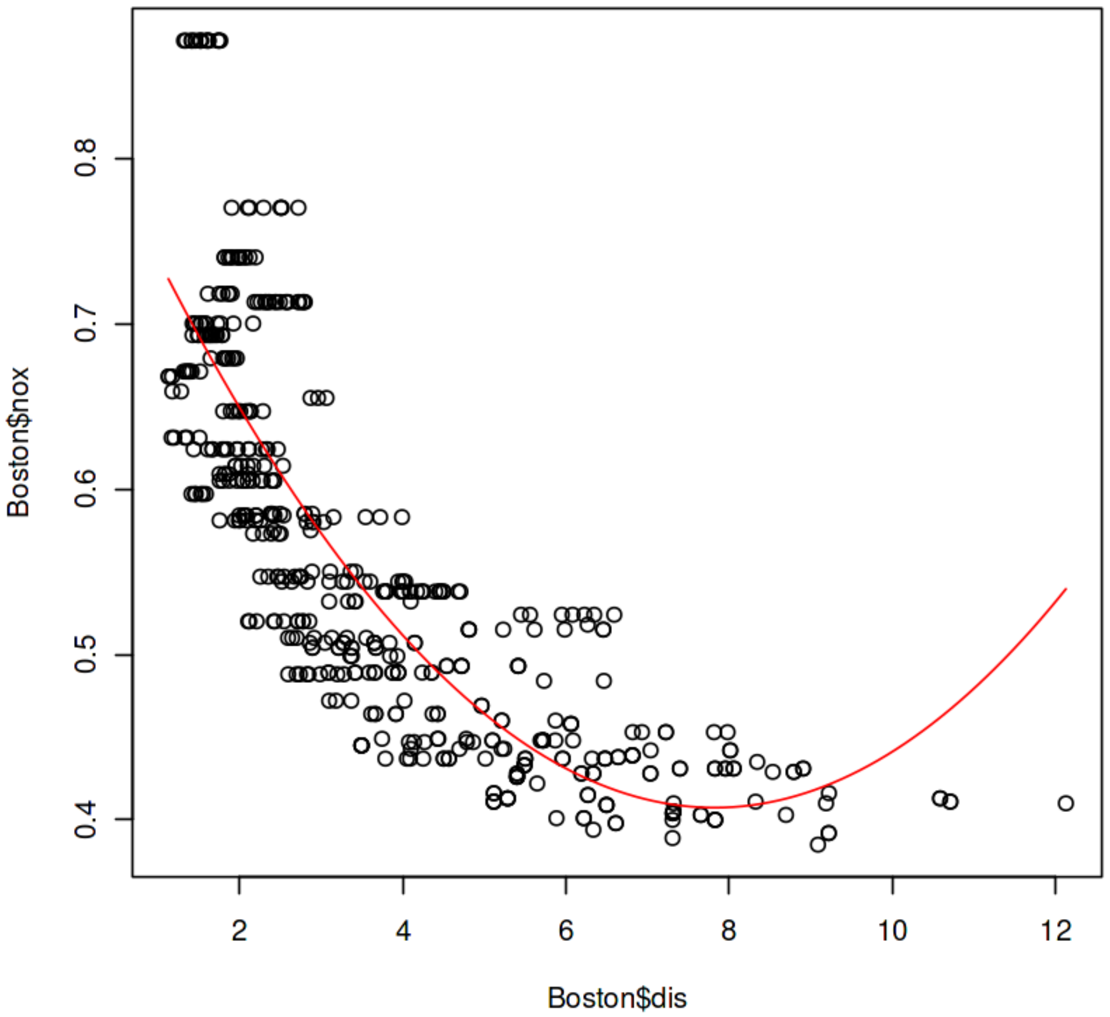
CODE:

```
Boston=read.csv("Boston.csv",header=T,na.strings="?",stringsAsFactors=FALSE)
Model2=lm(nox ~ dis+I(dis^2), data = Boston)
```

```
summary(Model2)
plot(Boston$dis, Boston$nox)
curve(predict(Model2,data.frame(dis=x)),add=TRUE,col="red")
```



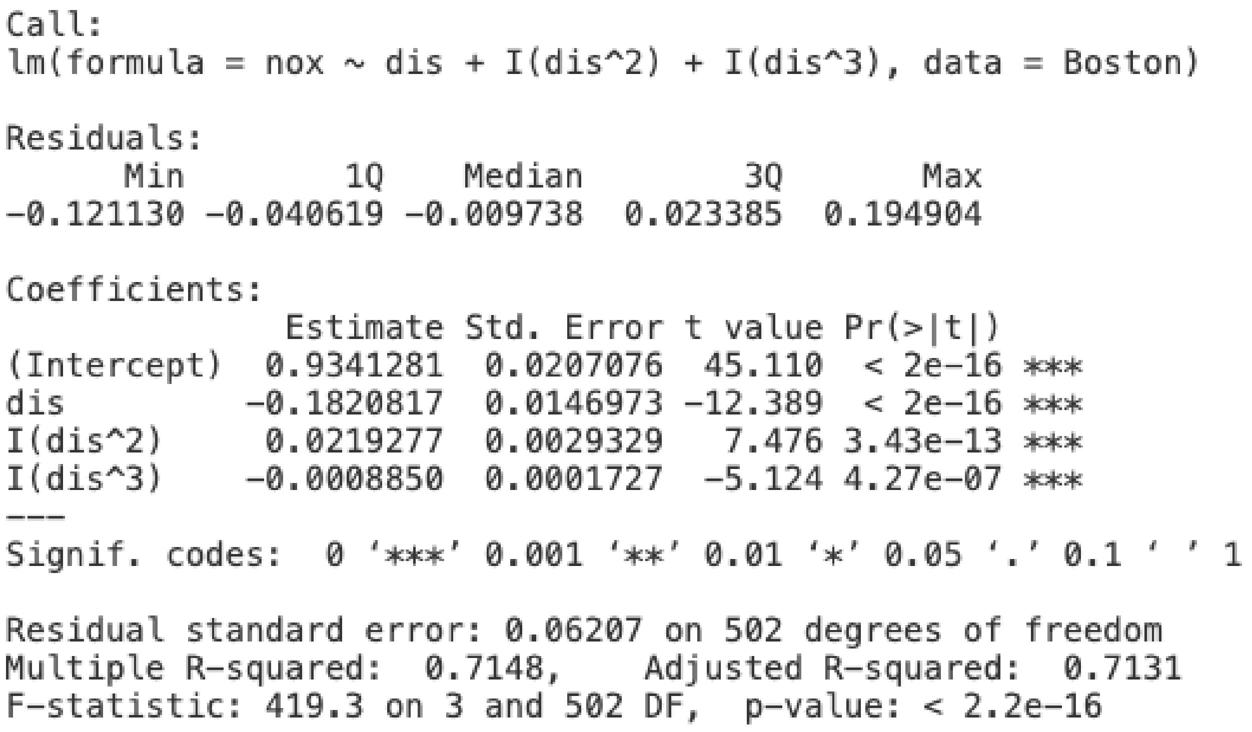
```
Call:
lm(formula = nox ~ dis + I(dis^2), data = Boston)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.129559 -0.044514 -0.007753  0.025778  0.201882 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.843991   0.011196   75.39   <2e-16 ***
dis         -0.111628   0.005320  -20.98   <2e-16 ***
I(dis^2)     0.007135   0.000530   13.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06361 on 503 degrees of freedom
Multiple R-squared:  0.6999,	Adjusted R-squared:  0.6987 
F-statistic: 586.4 on 2 and 503 DF,  p-value: < 2.2e-16
```



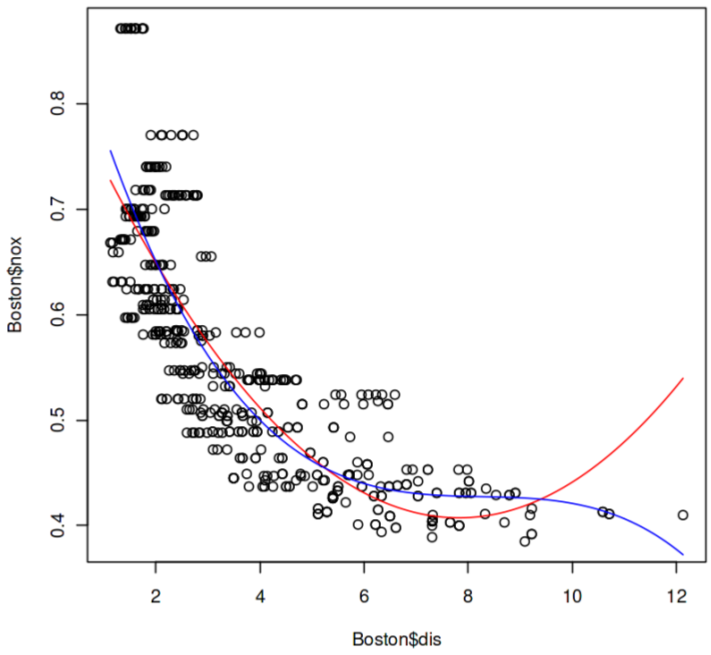
c. Fit a cubic model.

CODE:

```
Boston=read.csv("Boston.csv",header=T,na.strings="?",stringsAsFactors=FALSE)
Model3=lm(nox ~ dis+I(dis^2)+I(dis^3), data = Boston)
summary(Model3)
plot(Boston$dis, Boston$nox)
curve(predict(Model2,data.frame(dis=x)),add=TRUE,col="red")
curve(predict(Model3,data.frame(dis=x)),add=TRUE,col="blue")
```



```
Call:
lm(formula = nox ~ dis + I(dis^2) + I(dis^3), data = Boston)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.121130 -0.040619 -0.009738  0.023385  0.194904 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.9341281  0.0207076  45.110  < 2e-16 ***
dis         -0.1820817  0.0146973 -12.389  < 2e-16 ***
I(dis^2)     0.0219277  0.0029329   7.476 3.43e-13 ***
I(dis^3)    -0.0008850  0.0001727  -5.124 4.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06207 on 502 degrees of freedom
Multiple R-squared:  0.7148,	Adjusted R-squared:  0.7131 
F-statistic: 419.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

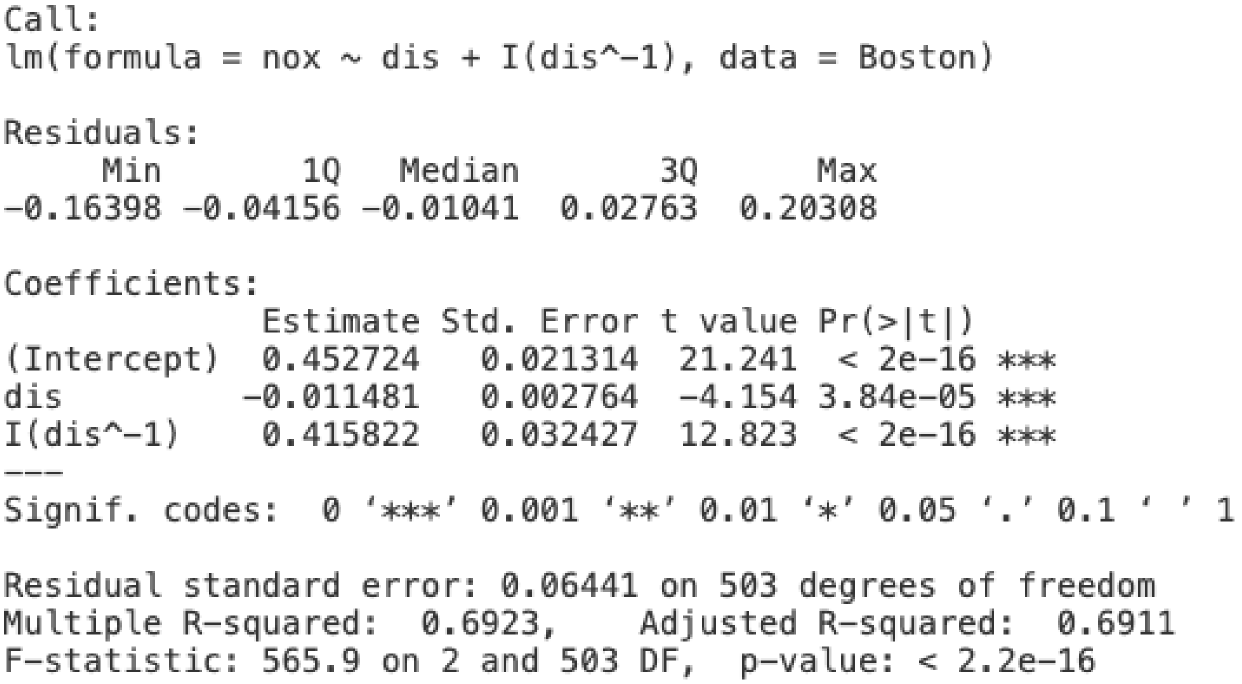


d. Fit a model that uses (dis)-1. Use lm(nox~I(dis^-1)).

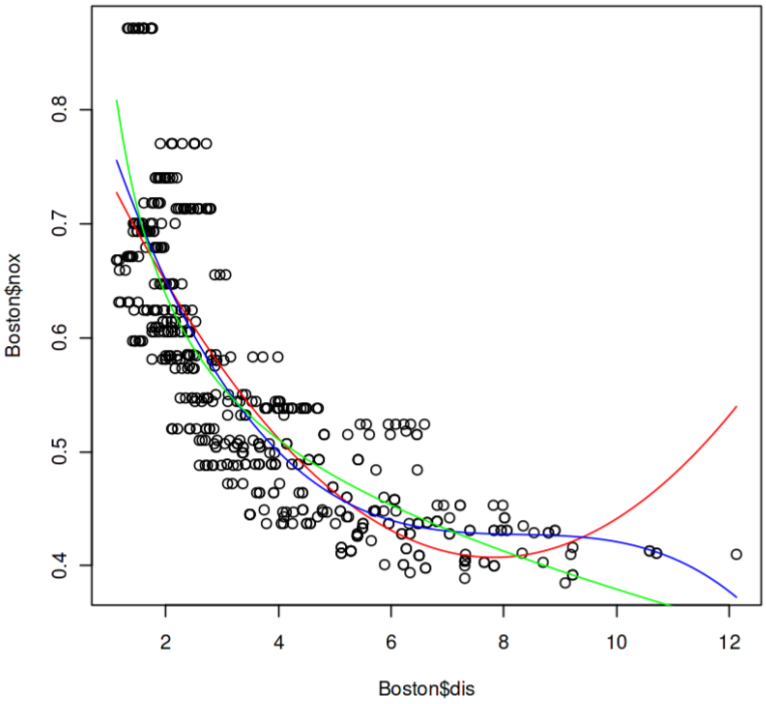
CODE:

Boston=read.csv("Boston.csv",header=T,na.strings="?",stringsAsFactors=FALSE)
Model4=lm(nox ~ dis+I(dis^-1), data = Boston)
summary(Model4)
plot(Boston$dis, Boston$nox)
curve(predict(Model2,data.frame(dis=x)),add=TRUE,col="red")
curve(predict(Model3,data.frame(dis=x)),add=TRUE,col="blue")

curve(predict(Model4,data.frame(dis=x)),add=TRUE,col="green")



```
Call:
lm(formula = nox ~ dis + I(dis^-1), data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16398 -0.04156 -0.01041  0.02763  0.20308 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.452724   0.021314  21.241  < 2e-16 ***
dis         -0.011481   0.002764  -4.154 3.84e-05 ***
I(dis^-1)    0.415822   0.032427  12.823  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06441 on 503 degrees of freedom
Multiple R-squared:  0.6923,	Adjusted R-squared:  0.6911 
F-statistic: 565.9 on 2 and 503 DF,  p-value: < 2.2e-16
```



e. Larry and Mary have a disagreement. Larry likes Model c because it fits the data well and has a small RSE. Mary prefers Model d. Even though Model d has a larger RSE, she claims it is a better fit for the overall situation. Who do you side with? Explain.

Larry asserts that his statement holds merit based on the lower RSE, contending that within the dataset, there are variables exhibiting values below 0.4 as we progress further. He emphasizes the trend illustrated by the curve of model c, supporting his claim. On the other hand, Mary contends that the data on the graph levels off around 0.4, suggesting that the trend depicted by model d is more precise. I align with Larry's viewpoint, given the smaller average error indicated by the RSE value, signifying a more favorable accuracy measure.