Introduction to Statistical Learning Book Exercises

Kevin Sullivan

2025-01-14

Chapter 2: Statistical Learning

Chapter Topics:

1. Prediction
2. Inference
3. Parametric Methods
4. Non-Parametric Methods
5. Trade-Off Between Prediction Accuracy and Model Interpretability
6. Supervised Vs. Unsupervised Learning
7. Assessing Model Accuracy
8. Measuring the Quality of Fit
9. Bias - Variance Trade-Off
10. Classification
11. K-Nearest Neighbors

Practice Question: Load and Perform Exploratory Data Analysis

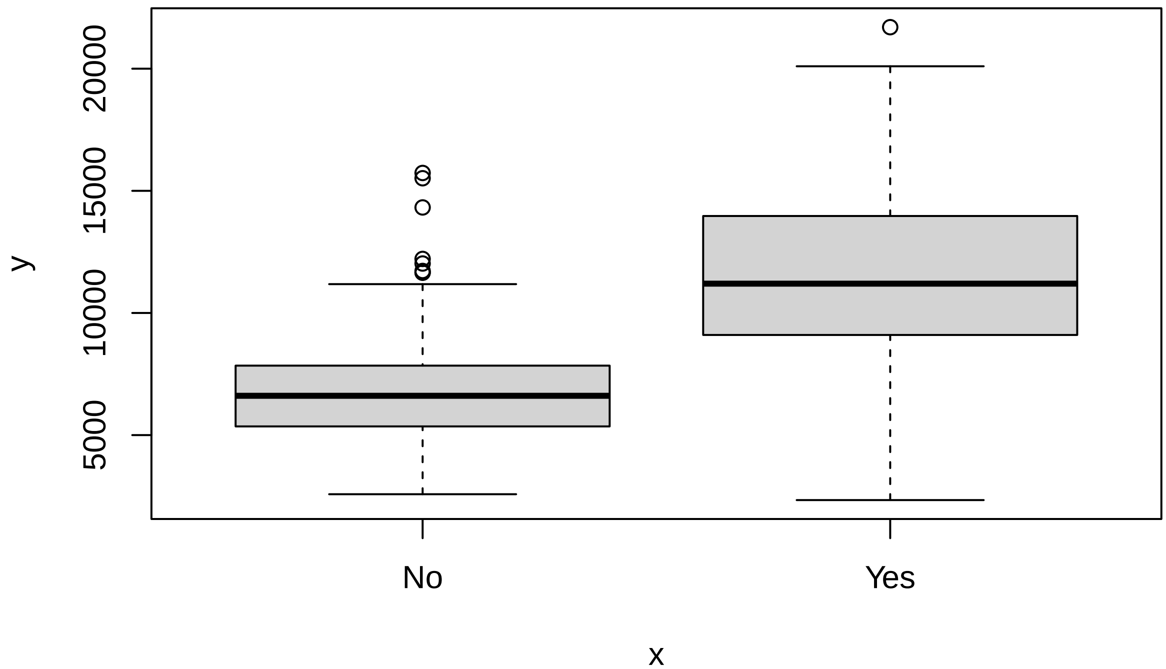
```
# Load Data
chap_two_data = College
```

```
head(chap_two_data)

summary(chap_two_data)
```

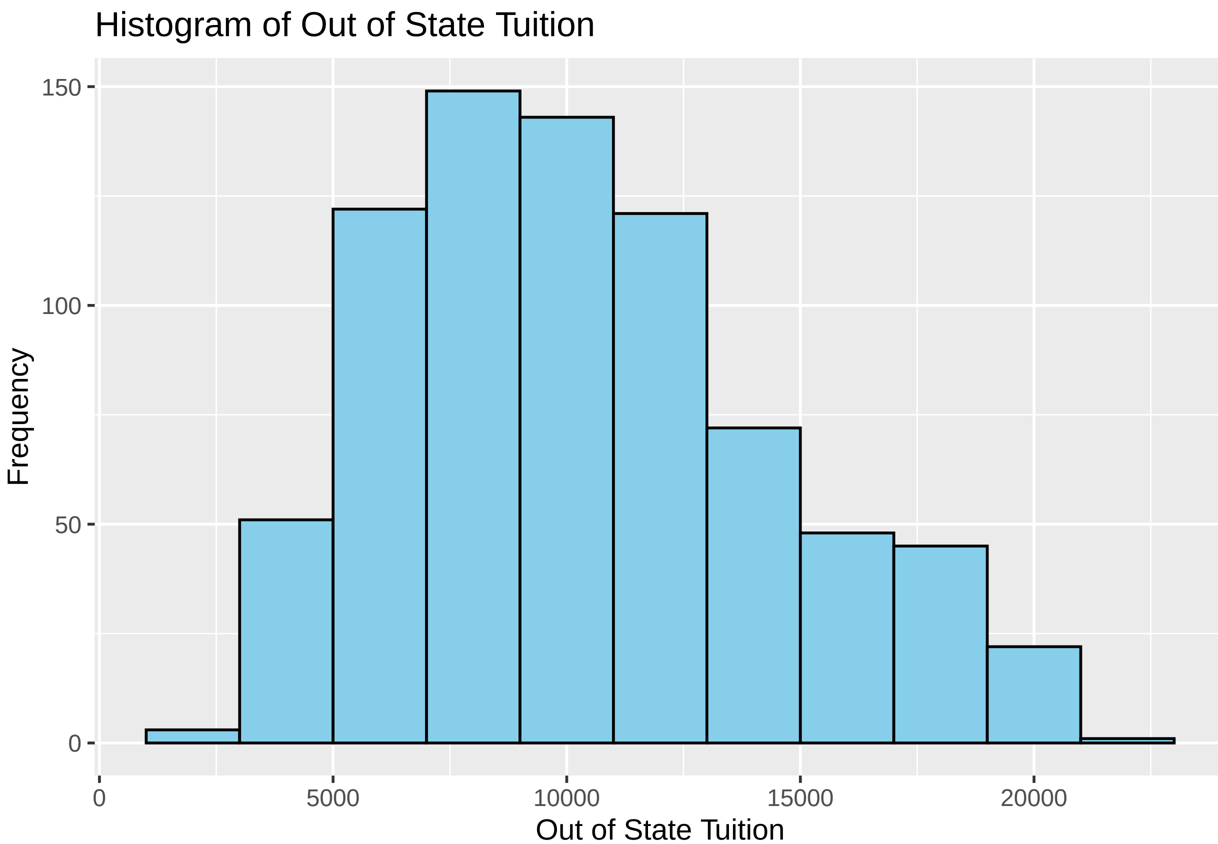
```
# Exploratory Data Analysis

plot(chap_two_data$Private, chap_two_data$Outstate)
```




```r
Elite = rep("No", nrow(chap_two_data))
Elite[chap_two_data$Top10perc > 50] = "Yes"
Elite <- as.factor(Elite)
college = data.frame(chap_two_data, Elite)

college %>%
  ggplot(aes(x = Outstate)) +
  geom_histogram(binwidth = 2000, fill = "skyblue", color = "black") +
  labs(title = "Histogram of Out of State Tuition", x = "Out of State Tuition", y = "Frequency")
```



Chapter 3: Linear Regression

Chapter Topics:

1. Simple Linear Regression
2. Assessing Accuracy of Coefficient Estimates
3. Assessing Accuracy of the Model
4. Multiple Linear Regression
5. Qualitative Predictors
6. Potential Problems
7. Comparison of Linear Regression and KNN

Practice Question

1. This question involves the use of simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the re- sponse?
ii. How strong is the relationship between the predictor and the response?
iii. Is the relationship between the predictor and the response positive or negative?
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

```
model1 = lm(mpg ~ horsepower, data = Auto)

summary(model1)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```r
# MPG associated with a horsepower of 98

mpg_98 = 39.935861 - (0.157845 * 98)

mpg_98
```

```
## [1] 24.46705
```

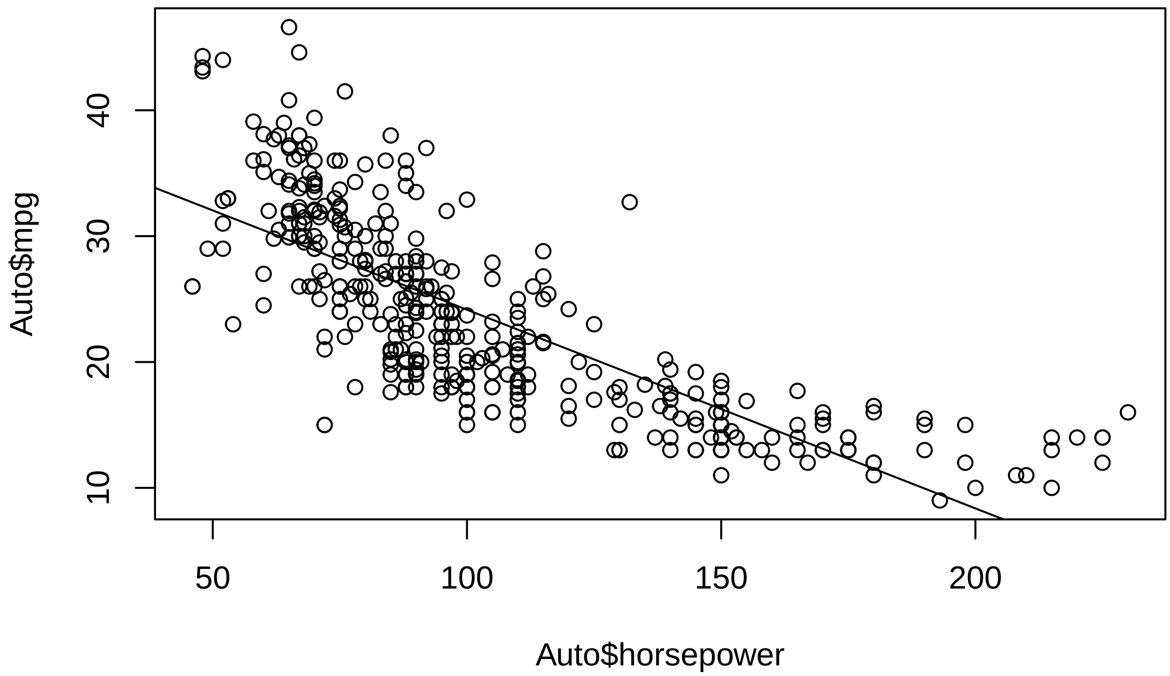
Comment on Model Output:

The horsepower variable has a negative coefficient which indicates that holding all else equal, a one unit increase in horsepower corresponds to a -0.157 unit decrease in mpg. This result is statistically significant at the $p < 0.01$ level. With an R^2 of 0.6049,the model appears to decently capture the variation in the data.

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

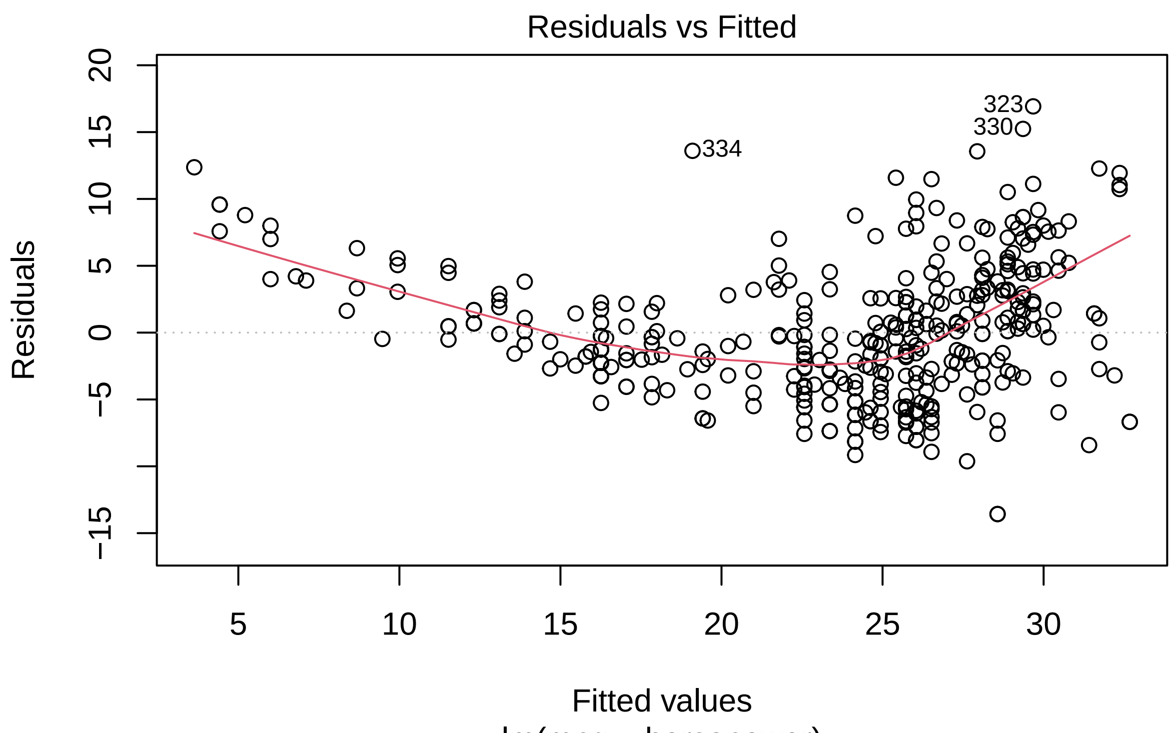
(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

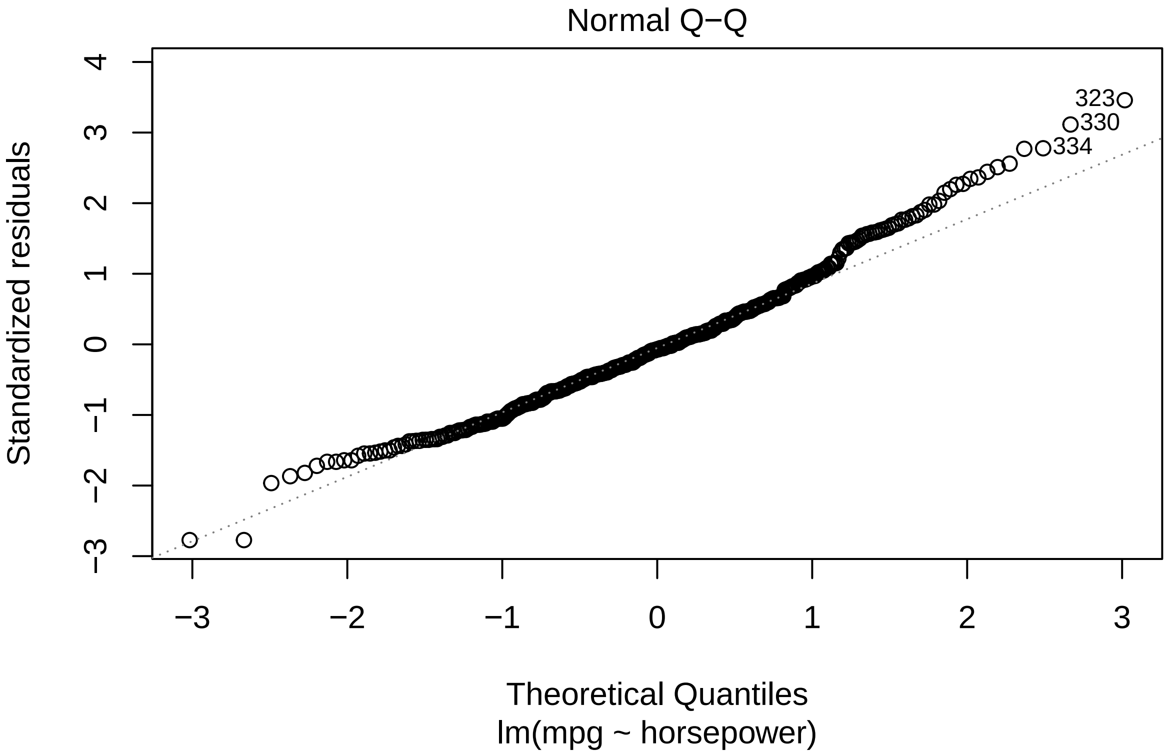
```r
# Plot data with reg line
plot(Auto$horsepower, Auto$mpg)
abline(model1)
```

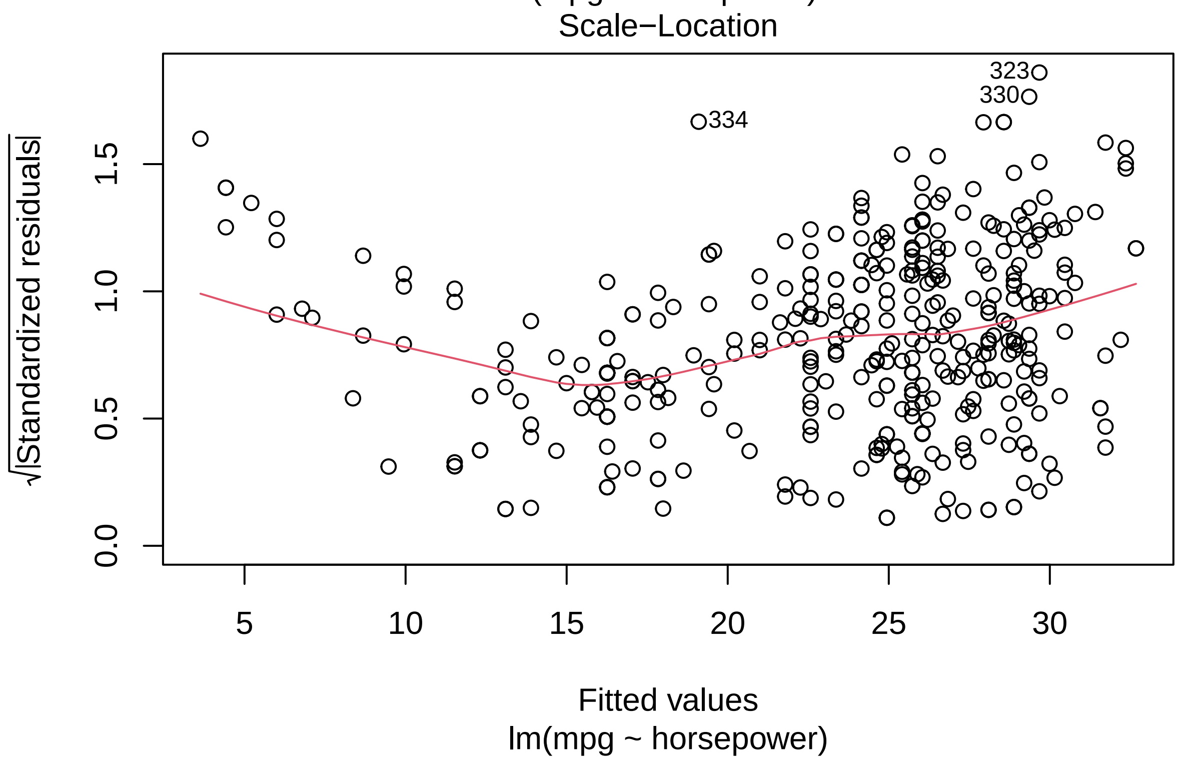


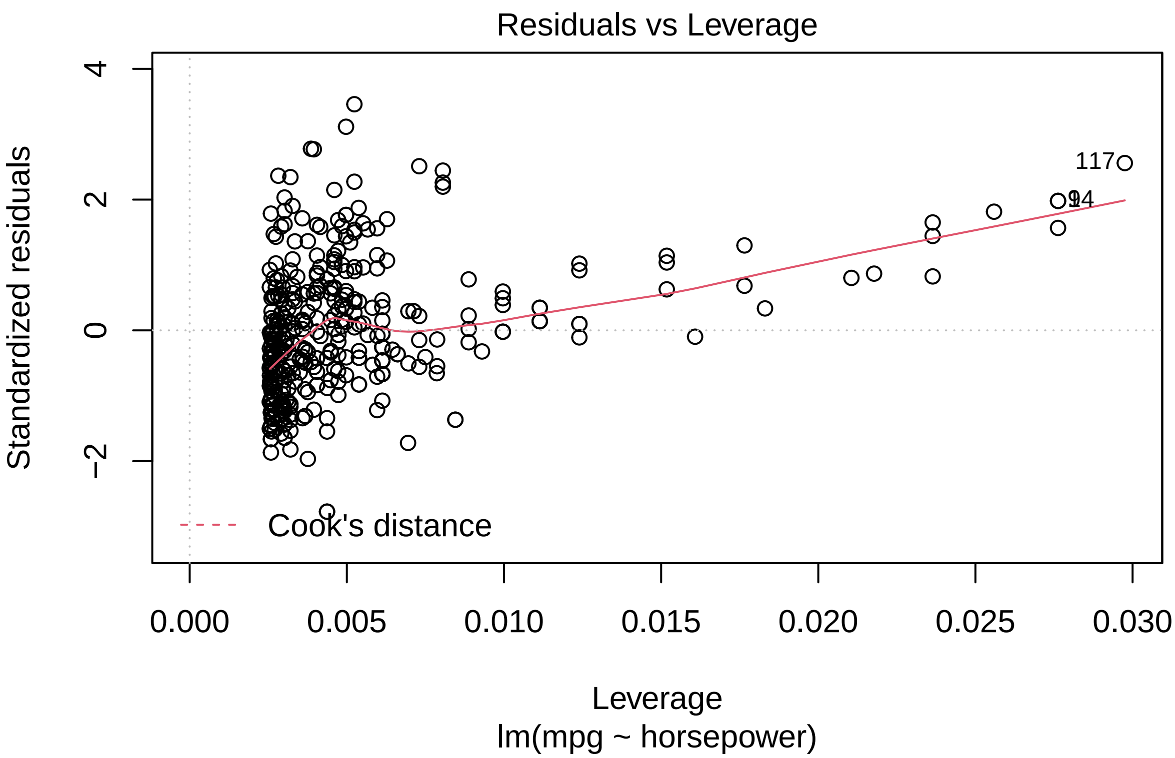

```r
# Diagnostic Plots

plot(model1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ horsepower)
323
330
334


Scale−Location
√|Standardized residuals|
Fitted values
lm(mpg ~ horsepower)
323
330
334



Based on the diagnostic plots it appears that a linear model might not be the best to use on this data. There appears to be a slight curvature to the data along with a fan-shaped pattern to the residuals that indicates that the conditions required for linear modeling may not hold here.

10. This question should be answered using the Carseats data set.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
# Multiple Regression

head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

```
mult_reg_model = lm(Sales ~ Price + Urban + US, data = Carseats)

summary(mult_reg_model)
```

```
##
```

```
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

Price: The negative and statistically significant coefficient indicates that (all else equal) a one unit increase in price corresponds to a -0.05 unit decrease in carseat sales.

UrbanYes: Since Urban is a categorical variable with "no" being the reference category, the result in the model can be interpreted as: Compared to those who do not live in urban areas, those who do live in urban areas buy -0.02 less car seat units. This result is not statistically significant.

USYes: Since US is a categorical variable with "no" being the reference category, the result in the model can be interpreted as: Compared to those who do not live in the US, those who do live in the US buy 1.2 more car seat units. This result is statistically significant.