Dealing with Data II Homework 7 Solutions

Joshua D. Ingram

2022-12-08

Reading - Textbook Chapter 13

a) Section 13.1

b) Section 13.2

Do not pay much attention of the formula, but rather where you can find statistics, such as Rˆ2 or the F statistic, in output.

c) Section 13.3

This covers the overall F-test, but does not talk much about the nested F-test, which we covered in class. However, the section mentions the t-test for testing individual regression coefficients. For testing individual regression coefficients, the t-test and the F-test comparing nested models are the same.

Answer:

d) Section 13.4

We already discussed residual plots when we did simple linear regression at the beginning of the semester, this is just a bit more information.

e) Section 13.5

About categorical predictors and interactions, which we have already covered.

Textbook Exercises - Chapter 13

13.6 - Crime rate and income

Refer to the previous exercise. MINITAB reports the following results for the multiple regression of $y =$ crime rate on $x_1 =$ median income (in thousands of dollars) and $x_2 =$ urbanization.

(This refers to the same dataset that we have analyzed in class. Get it into R and fit and answer all parts of the question using R.)

a) Report the prediction equations relating crime rate to income at

```
crime <- read.csv("/Users/joshuaingram/Main/Projects/masters_coursework/teaching_assistant/dealing_with_
colnames(crime) <- c("county", "crime_rate", "education", "urbanization", "median_income")

fit <- lm(crime_rate ~ median_income + urbanization, data = crime)
summary(fit)
```

```
##
## Call:
## lm(formula = crime_rate ~ median_income + urbanization, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.130 -15.590  -6.484  16.595  48.921
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    39.9723    16.3536   2.444   0.0173 *
## median_income  -0.7906     0.8049  -0.982   0.3297
## urbanization    0.6418     0.1110   5.784 2.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.91 on 64 degrees of freedom
## Multiple R-squared:  0.4669, Adjusted R-squared:  0.4502
## F-statistic: 28.02 on 2 and 64 DF,  p-value: 1.815e-09
```

(i) urbanization levels of 0. Interpret Answer:

$$\hat{y_i} = 39.9723 - 0.7906x_{1,i}$$

Holding urbanization constant at 0, for every thousand dollar increase in median income we expect the crime rate to decrease by 0.79 per thousand people on average.

(ii) urbanization levels of 100. Interpret Answer:

$$\hat{y_i} = 39.9723 - 0.7906x_{1,i} + 64.18$$

Holding urbanization constant at 100, for every thousand dollar increase in median income we expect the crime rate to decrease by 0.79 per thousand people on average. The crime rate for urbanization levels at of 100 is expected to be 64.18 per thousand people greater than urbanization levels of 0.

b)

For the simple regression model relating $y =$ crime rate to $x = income$, MINITAB reports

$$\text{crime } = -11.6 + 2.61 \text{ income}_i$$

Interpret the effect of income, according to the sign of its slope. How does this effect differ from the effect of income in the multiple regression equation?

Answer:

For every one thousand dollar increase in median income, the crime rate increases by 2.61 per thousand people on average. This effect displays a positive relationship between relationship between income and crime rate, whereas the multiple regression shows a negative relationship between crime and median income.

c) Use the estimated slope for income in the simple and multiple regression model to explain the difference in the interpretation of the slope when

(i) ignoring urbanization   Answer:

For every one thousand dollar increase in median income, the crime rate increases by 2.61 per thousand people on average.

(ii) controlling urbanization.   Answer:

Holding urbanization constant, for every thousand dollar increase in median income we expect the crime rate to increase by 0.79 per thousand people on average.

13.11 - Used cars

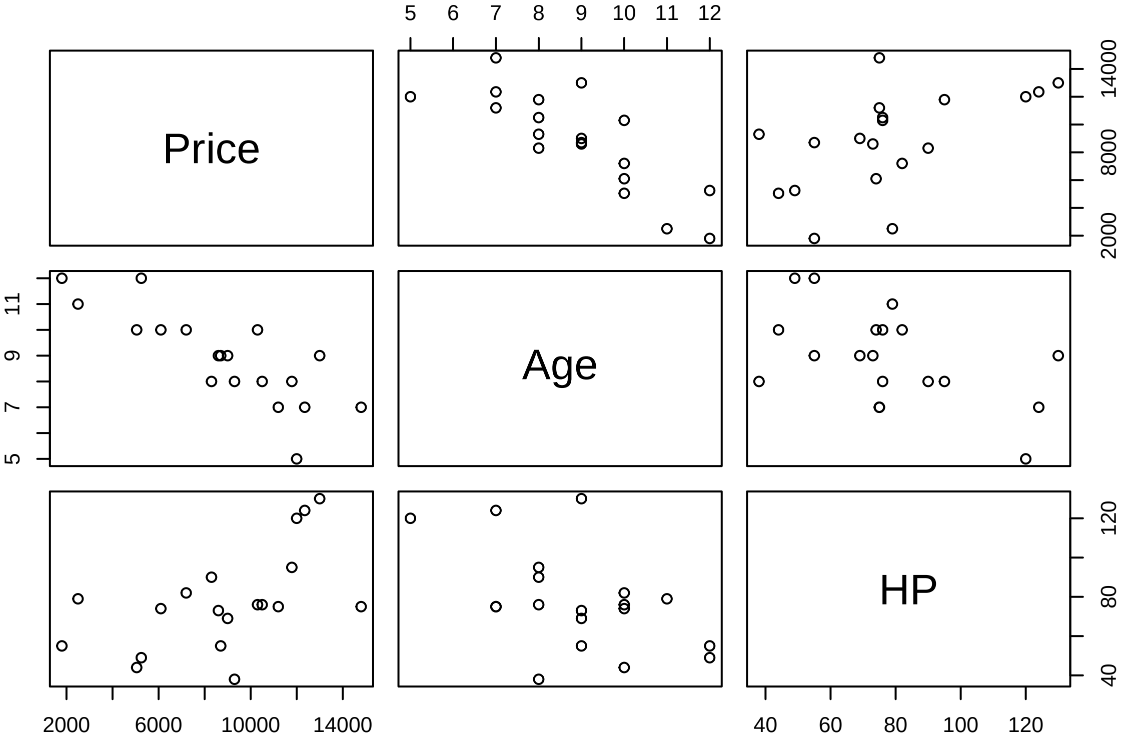
The following data are from a random sample of campus newspaper ads on used cars for sale. Consider the age and horsepower (HP) of a car to predict its selling price.

(Use R where relevant)

a) Construct a scatterplot matrix to investigate the relationship among price, age, and horsepower and interpret.

Answer:

```
cars <- read.csv("/Users/joshuaingram/Main/Projects/masters_coursework/teaching_assistant/dealing_with_
plot(cars[,c(2, 3, 4)])
```



There is a weak to moderate positive relationship between price and HP. There is a moderate negative relationship between price and age. There is a weak negative relationship between age and HP.

b) Find the multiple regression prediction equation for the selling price in terms of age and horsepower of 80 and

```
fit <- lm(Price ~ Age + HP, data = cars)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ Age + HP, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3397.4 -1217.8    43.3   934.1  3379.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19348.71    4053.35   4.774 0.000207 ***
## Age         -1406.30     319.75  -4.398 0.000449 ***
## HP             25.54      22.34   1.143 0.269739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2085 on 16 degrees of freedom
## Multiple R-squared:  0.6869, Adjusted R-squared:  0.6478
## F-statistic: 17.55 on 2 and 16 DF,  p-value: 9.23e-05
```

$$\hat{y_i} = 19,348.71 - 1,406.30 \text{ age}_i + 25.54 \text{ HP}_i$$

(i) is 8 years old, rounded to the nearest hundred? Answer:

$$\hat{y_i} = 19,348.71 - 1,406.30 * 8 + 25.54 * 80 = 10,141.51.$$

(ii) is 10 years old, rounded to the nearest hundred? Answer:

$$\hat{y_i} = 19,348.71 - 1,406.30 * 10 + 25.54 * 80 = 7,328.91.$$

c) Based on this multiple regression, can you predict the price difference between a car with 60 HP and a car with 80 HP without knowing the ages of the two cars? Explain.

Answer:

No, we need to know the age of the cars to predict the price difference. We can do this by holding the age constant, and finding the difference in the effects of prices for a car with 60 HP and a car with 80 HP.

13.41 - U.S. and foreign used cars

Refer to the used car data file from Exercise 13.11. The prediction equation relating y = selling price of used car (in \$) as a function of x_1 = age of car and x_2 = type of car (1 = U.S., 0 = Foreign) is $\hat{y} = 20,493 - 1,185x_1 - 2,379x_2$.

(Use R to verify all outputs. In addition, answer the following question: Fit and plot the model that allows for the effect of age to depend on the type of car. Describe the effect of age on the predicted sales price for cars that are from the US, and compare that to the effect of age when cars are foreign.)

a) Using this equation, find the prediction equation relating selling price and age, separately for U.S. and foreign cars.

Answer:

U.S. Cars:

$$\hat{y}_i = 20,493 - 1,185\text{x}_{1,i} + 2,379$$

Foreign Cars:

$$\hat{y}_i = 20,493 - 1,185\text{x}_{1,i}$$

b) Predict by how much the price changes for a one year increase in the age of the car. Does this apply for both types of cars? Explain.

Answer:

For every one year increase in age of the car, the predicted price decreases by $1,185 on average. This applies to both types of cars.

c) Find the predicted price of a (i) U.S. and (ii) foreign car that is eight years old. Show how the difference between them relates to a parameter estimate for the model.

Answer:

U.S. Cars:

$$\hat{y}_i = 20,493 - 1,185 * 8 + 2,379 = 13,392$$

Foreign Cars:

$$\hat{y}_i = 20,493 - 1,185 * 8 = 1,1013$$

The difference in the prices of the two types of cars is equal to the coefficient for x_2.

Online Lending

Refer to the Online Lending dataset that we briefly analyzed in class. (It is available on Canvas under Files -> Datasets.) Pick a few variables (at least one continuous and one categorical) and fit a linear regression model in R, using the interest rate as the response variable (as we did in class). Write a paragraph about your analysis and what you learned. Include at least one useful plot into your paragraph.

Answer:

```
loans <- read.csv("/Users/joshuaingram/Main/Projects/masters_coursework/teaching_assistant/dealing_with

fit <- lm(Interest.Rate ~ Income + FICO.Score + Home, data = loans)
summary(fit)
```

```
##
## Call:
## lm(formula = Interest.Rate ~ Income + FICO.Score + Home, data = loans)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.4341 -3.5590 -0.2443  2.5076 19.5673
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.248e+01  6.954e+00   8.984 2.25e-16 ***
## Income       1.353e-06  6.507e-06   0.208    0.836
## FICO.Score  -7.080e-02  9.793e-03  -7.230 1.07e-11 ***
## Homeown      1.028e-01  1.175e+00   0.087    0.930
## Homerent     8.040e-01  7.783e-01   1.033    0.303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.05 on 195 degrees of freedom
## Multiple R-squared:  0.2168, Adjusted R-squared:  0.2007
## F-statistic:  13.5 on 4 and 195 DF,  p-value: 9.933e-10
```