Dealing with Data II Homework 3 Solutions

Joshua D. Ingram

2022-10-05

1. Reading

Read Sections 3.3 and 3.4 in the textbook.

2. Textbook Exercises Chapter 3

3.27 - Home selling prices

The House Selling Prices FL data file, preloaded in the *Linear Regression* app and also on the book's website, lists selling prices of homes in Gainesville, Florida, in 2003 and some predictors for the selling price. For the response variable y = selling price in thousands of dollars and the explanatory variable x = size of house in thousands of square feet, $\hat{y} = 9.2 + 77.0x$.

a. How much do you predict a house would sell for if it has

(i) 2,000 square feet Answer:

$$\hat{y} = 9.2 + 77.0(2) = 163.2$$

We predict a 2,000 square foot house in Gainseville, FL to have a selling price of $163,200 on average.

(ii) 3,000 square feet? Answer:

$$\hat{y} = 9.2 + 77.0(3) = 240.2$$

We predict a 3,000 square foot house in Gainseville, FL to have a selling price of $240,200 on average.

b. Using results in part a, explain how to interpret the slope.

Answer:

For every 1,000 square foot increase in the size of the house, we expect the selling price of a house in Gainseville, FL to increase by $77,000 on average.

c. Is the correlation between these variables positive or negative? Why?

Answer:

Positive. As the square footage increases, the price increases by the slope.

d. One home that is 3,000 square feet sold for $300,000. Find the residual, and interpret.

Answer:

$$y_i - \hat{y}_i = 300 - (9.2 + 77.0(3)) = 59.8$$

We predict a 3,000 square foot house in Gainseville, FL to have a selling price of $240,200. For the observed housing price of $300,000, we have a residual of $59,800. We under predicted the housing price by $59,800.

3.34 - Regression between cereal sodium and sugar

The following figure, taken from the *Linear Regression* app, shows the result of a regression analysis of the explanatory variable x = sugar and the response variable y = sodium for the breakfast cereal data set discussed in Chapter 2 (the Cereal data file on the book's website).

a. What criterion is used in finding the line?

Answer:

The sum of squared residuals are minimized.

b. Can you draw a line that will result in a smaller sum of squared residuals?

Answer:

No. Using the ordinary least squares estimator results in the best fit with the smallest sum of squared residuals.

c. Now let's look at a histogram of the residuals. Explain what the two short bars on the far right of the histogram mean in the context of the problem. Which two brands of cereal do they represent? Can you find them on the scatterplot?

Answer:

The two short bars are Raisan Bran and Rice Krispies. They both have over 250 mg of sodium.

3.54 - More firefighters cause worse fires?

Date are available for all fires in Chicago last year on x = number of fire-fighters at the fire and y = cost of damages due to the fire.

a. If the correlation coefficient is positive, does this mean that having more firefighters at a fire causes the damages to be worse? Explain.

Answer:

No. There is likely a confounding variable, and this is only an observational study. An increase in the number of fire fighters could mean that the severity of the fire is worse, meaning there are more damages, and thus more fire fighters are required to be on the scene.

b. Identify a third variable that could be a common cause of x and y. Construct a hypothetical scatterplot, identifying points according to their value on the third variable, to illustrate your argument.

Answer:

A third variable could be the type of fire: Class A ("Ordinary" Fire), Class B (Liquids & Gases), Class C (Electrical Fires), Class D (Metallic Fires), and Class K (Grease Fires or Cooking Fires).

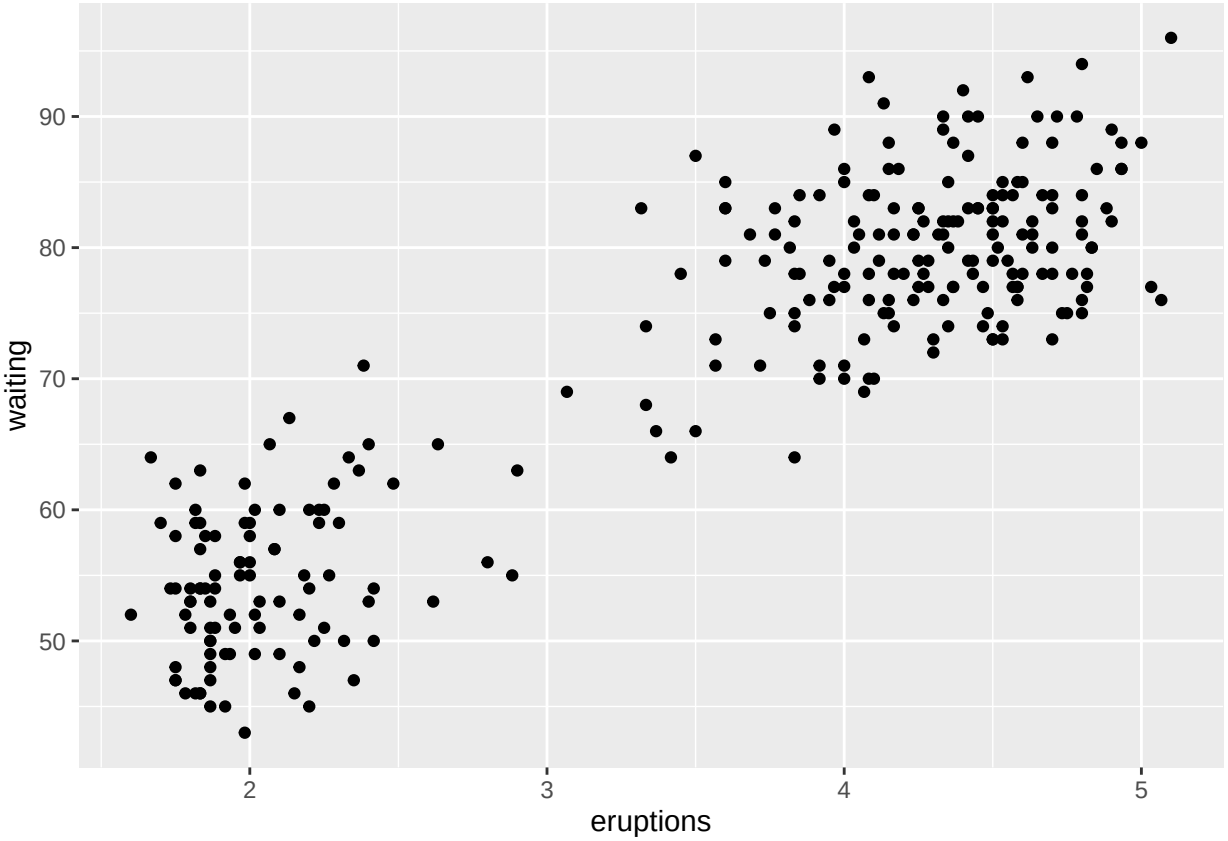
3. Old Faithful

The Old Faithful dataset (available in R by typing faithful in the R prompt, e.g., > faithful) is a dataset about the duration of an eruption of that famous geyser (e.g., 3.6 minutes for the first observation in the dataset) and the waiting time to the next eruption (called "waiting", e.g., 79 minutes for the first observation).

a. Construct an appropriate plot and describe the relationship between the eruption and waiting time.

Answer:

```
ggplot(faithful, aes(x = eruptions, y = waiting)) +
  geom_point()
```



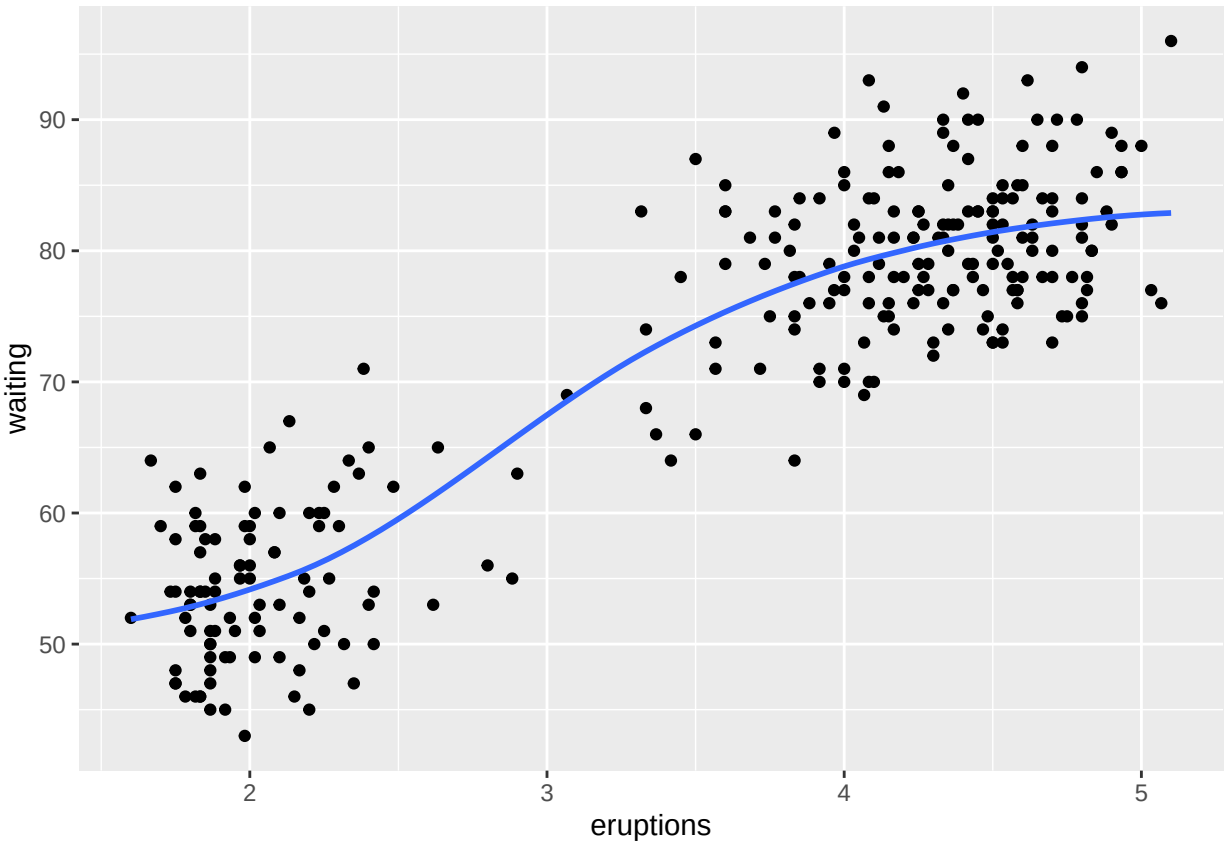
There is a linear relationships between the eruption time and the waiting time between eruptions. There are also two clusters of events, which seem to be long eruptions with long waiting times, and short eruptions with short waiting times. There are fewer observations in the middle values.

b. On your plot, include a smooth trend line to indicate the relationship. Briefly comment whether fitting a linear regression model is appropriate.

Answer:

```
ggplot(faithful, aes(x = eruptions, y = waiting)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



It would be appropriate to fit a linear regression model is appropriate. However, it would be possible to fit two different lines to the groupings of the short eruptions and waiting times, and the long eruptions and waiting times as they would result in two slightly different slopes.

c. Find and display the fitted regression equation (see the Scatterplot handout of how to do this nicely in R Markdown) and explain what each term in the equation stands for, in the context of the problem. i.e., explain what y-hat stands for, state what x stands for, and identify the values for the intercept and slope. Then, interpret the slope in context. Is it meaningful to interpret the intercept?

Answer:

```
fit <- lm(waiting ~ eruptions, data = faithful)
summary(fit)
```

```
##
## Call:
## lm(formula = waiting ~ eruptions, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0796  -4.4831   0.2122   3.9246  15.9719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.4744     1.1549   28.98   <2e-16 ***
## eruptions    10.7296     0.3148   34.09   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.914 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

$$\hat{y}_i = 33.4744 + 10.7296 x_i$$

$\hat{y}_i$ is the predicted waiting time and x_i is the eruption time. 33.4744 (in minutes) is the intercept and 10.7296 (in minutes) is the slope.

For every 1 minute increase in the eruption time, we expect an increase in the waiting time till the next eruption to increase by 10.7296 minutes on average.

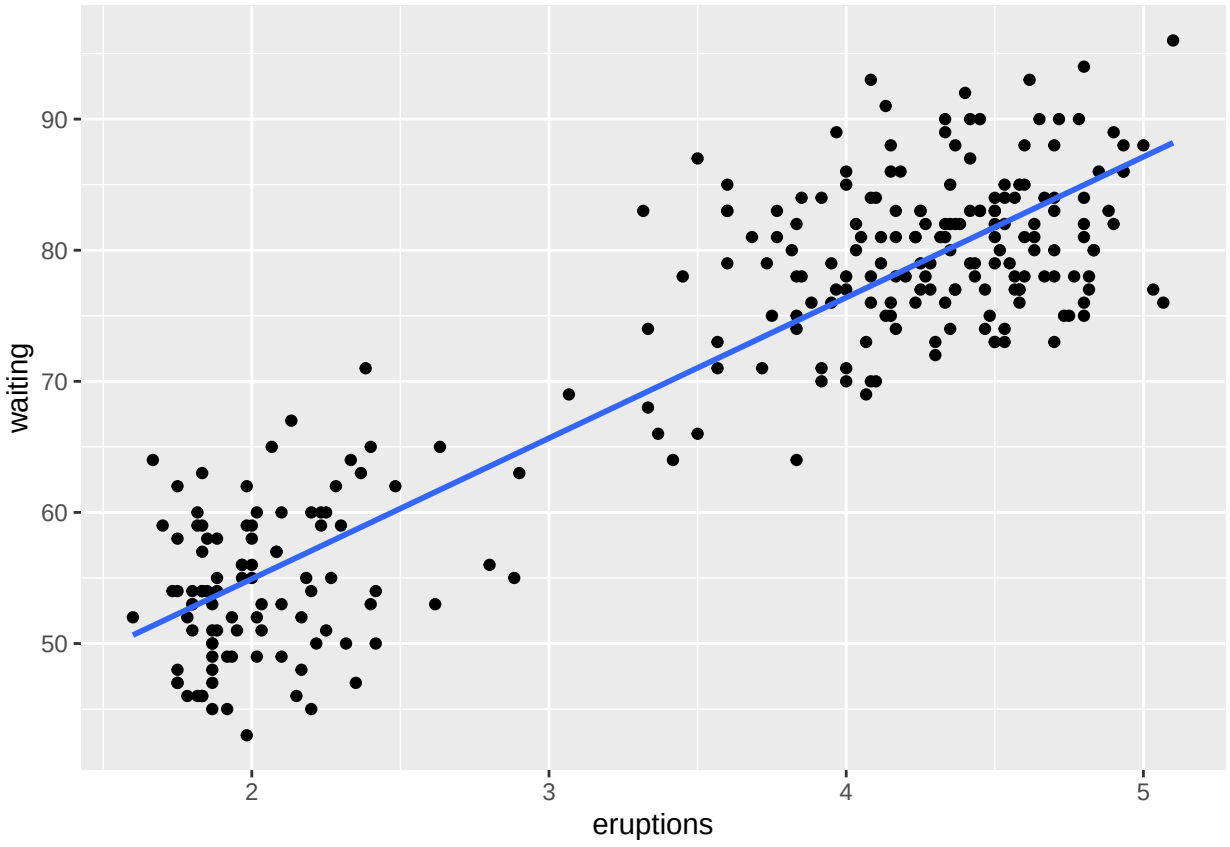
It does not make sense to interpret the intercept, as we cannot have an eruption event occur that lasts 0 seconds.

d. Construct the scatterplot with the fitted regression line superimposed.

Answer:

```
ggplot(faithful, aes(x = eruptions, y = waiting)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



e. Suppose you visit Old Faithful in the winter and just witnessed an eruption of 4 minutes and 15 seconds. Because it's freezing cold (temperatures at Old Faithful in winter can easily drop below 0 degrees F (click here to check out the current temperature or here to visit the live webcam of Old Faithful, which displays prediction of when the next eruption will occur) you want to go inside the visitor center and warm up. Use the prediction equation to predict the waiting time until the next eruption, so you can get out of the visitor center in time to witness the next eruption.

Answer:

```
33.4744 + 10.7296 * 4.25
```

```
## [1] 79.0752
```

We predict a waiting time of 79.0752 minutes until the next eruption, so we should wait nearly 80 minutes in the visitor center.

f. If one eruption lasted 30 seconds longer than another, by how much do you predict the waiting time in between eruptions to change?

Answer:

```
10.7296 * 0.5
```

```
## [1] 5.3648
```

We predict the eruption lasting 30 seconds longer to have a waiting time 5.3648 minutes longer than the shorter eruption.

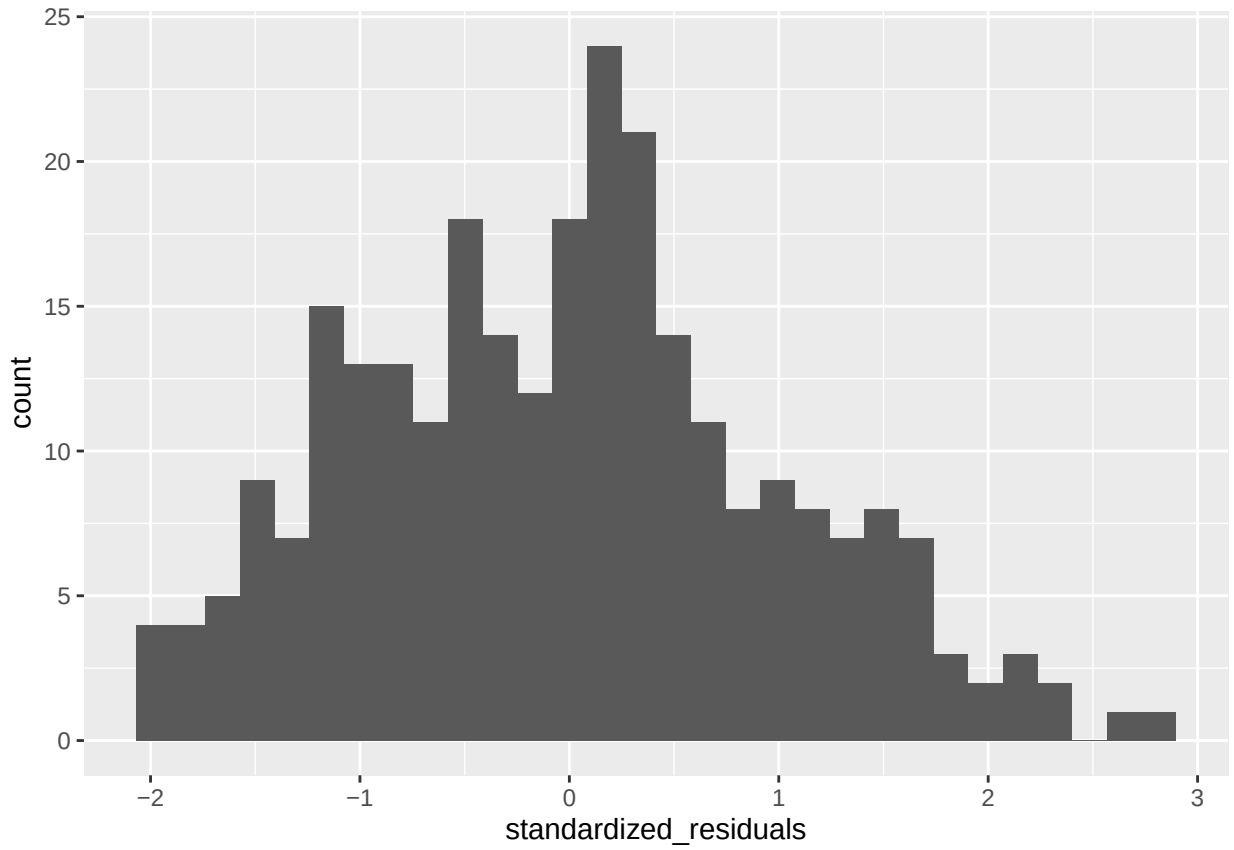
g. Construct a (standardized) residual plot and comment on whether this plot indicates the linear regression is a suitable fit for the dataset.

Answer:

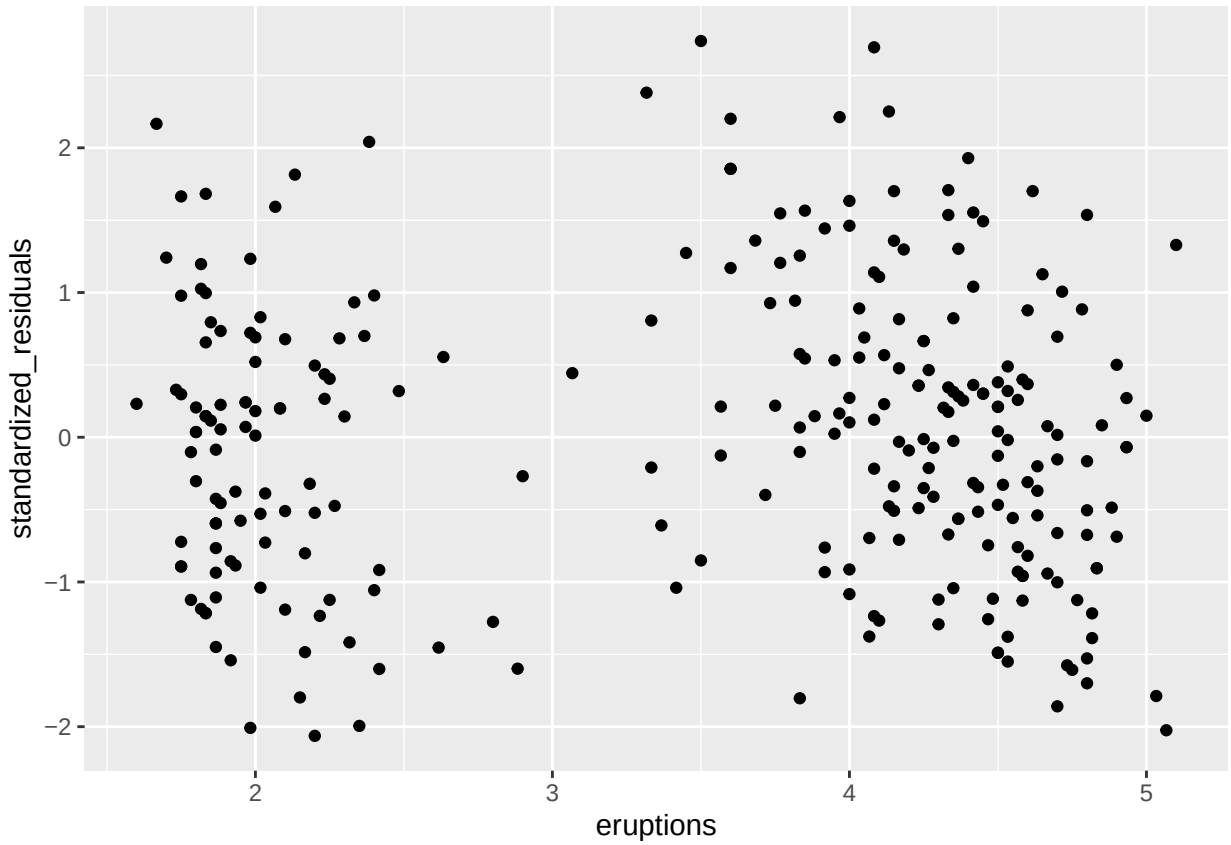
```
faithful$standardized_residuals <- rstudent(fit)

ggplot(data = faithful, aes(x = standardized_residuals)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = faithful, aes(x = eruptions, y = standardized_residuals)) + geom_point()
```



A linear regression model is suitable for the faithful dataset. The majority of the standardized residuals are between -2 and 2, and there are no trends in the residuals vs eruptions plot.