Lab2: Linear Regression

Claire Jellison

9/11/2019

$y^{3.5}$

```
data(quakes)
str(quakes)
```

```
## 'data.frame':    1000 obs. of  5 variables:
##  $ lat     : num  -20.4 -20.6 -26 -18 -20.4 ...
##  $ long    : num  182 181 184 182 182 ...
##  $ depth   : int  562 650 42 626 649 195 82 194 211 622 ...
##  $ mag     : num  4.8 4.2 5.4 4.1 4 4 4.8 4.4 4.7 4.3 ...
##  $ stations: int  41 15 43 19 11 12 43 15 35 19 ...
```

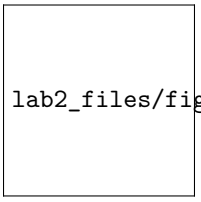
```
?quakes
```

Earthquake detection Included in the data set is a column recording the number of stations that detected each earthquake. This refers to a global network of seismographs and it stands to reason that the larger the quake, the more widely it will be detected.

Exercise 1

Create a plot of the relationship between stations and magnitude. How would you characterize the relationship? (If you see overplotting, you may want to add jitter to your points or make them transparent by playing with the alpha value.)

```
library(ggplot2)
ggplot(quakes, (aes(x = stations, y = mag))) + geom_point(position = "jitter")
```



lab2_files/figure-latex/unnamed-chunk-2-1.pdf

It appears as though there is a fairly linear relationship between the magnitude of the earthquake and the number of stations that it is recorded at, with the greater the magnitude the more stations detecting it.

Exercise 2

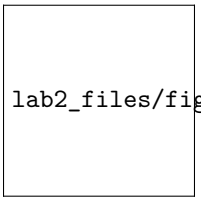
Before you go ahead and fit a linear model to this trend, if in fact there was no relationship between the two, what would you expect the slope to be? What about the intercept?

If there was no relationship between the two variables, we would expect the slope to be zero (since there is no positive or negative correlation between them) and the y intercept should be at the mean magnitude sincethe mean is a pretty optimal prediction given that the other variable doesn't really provide any useful information.

Exercise 3

Ok, now go ahead and fit a linear model called m1 to the trend and add that line to the plot from exercise 1. Interpret your slope and intercept in the context of the problem.

```
m1 <- lm(stations ~ mag, data = quakes)
ggplot(quakes, (aes(x = stations, y = mag))) + geom_point(position = "jitter") +
  geom_smooth(method = "lm")
```




```
summary(m1)
```

```
##
## Call:
## lm(formula = stations ~ mag, data = quakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.871  -7.102  -0.474   6.783  50.244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -180.4243     4.1899  -43.06   <2e-16 ***
## mag           46.2822     0.9034   51.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.5 on 998 degrees of freedom
## Multiple R-squared:  0.7245, Adjusted R-squared:  0.7242
## F-statistic:  2625 on 1 and 998 DF,  p-value: < 2.2e-16
```

Here, the y-intercept is 4.09 indicating that if an earthquake is picked up at no stations the model would predict that it is 4.09. The slope is about 0.016 indicating that for each additinal station that detects a quake the predicted magnitude of the quake increases by 0.016.

Exercise 4

Verify the way that lm() has computed your slope correctly by using R to do the calculation using the equation for the slope based on X and Y.

Exercise 5

Using R, calculate a 95% confidence interval for the slope of the model that you fit in exercise 3. Confirm the calculation using confint().

```
confint(m1, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -188.64628 -172.20238
## mag           44.50944   48.05498
```

Exercise 6

How many stations do you predict would be able to detect an earthquake of magnitude 7.0?

y = 4.0972676 + 0.0156542(stations) stations = (y - 4.0972676)/ (0.0156542)

```
stations = (7.0 - 4.0972676)/ (0.0156542)
stations
```

```
## [1] 185.4283
```

Exercise 7

Questions 1 - 6 in this lab involve elements of data description, inference, and/or prediction. Which was the dominant goal in each question?

Section Three

Exercise 9

Please simulate a data set that has the same number of observations as quakes. To start, generate a vector of x's. You can either generate your own x's or use the exact same x's as the quakes data.

Exercise 10

Next, generate your ()'s (the value of the mean function at the observed x's). Please generate them by writing your own function of the form:

```
f_hat <- function(x) {
  # code goes here
}
```

Exercise 11

Now, generate the y's. Note that you'll need an estimate of (σ^2), $for which you can use (\hat{\sigma}^2 = RSS/n - 2). You can extract the vect$

Exercise 12

Finally, make a plot of your simulated data. How is it similar to the original data? How is it different? How might you change your model to make it more consistent with the data?

Challenge Problem

Use the latitude and longitude data to plot each of these earthquakes in quakes on a map with their magnitude mapped to the size of the plotting character. You may need to add some transparency to prevent overplotting.

One good way to assess whether your fitted model seems appropriate is to simulate data from it and see if it looks like the data that you observed. For the following questions it will be useful to reference the R code provided in the previous two lectures.

The output respose variable would be the average value of a home with the input being a given set of characteristics. So you could give the characteristics of a certain home and the model would predict a value of the home based on the relationships it has seen with previous data.