HW4

Ben Tankus

1/27/2021

ST 517: Data Analytics I Module 4 Homework 1. (8 points) Consider the Sleuth3 dataset case0901, which contains the results of a study of the Meadowfoam plant. Familiarize yourself with the study and load the data.

```
library(Sleuth3)
```

```
## Warning: package 'Sleuth3' was built under R version 4.0.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
head(case0901)
```

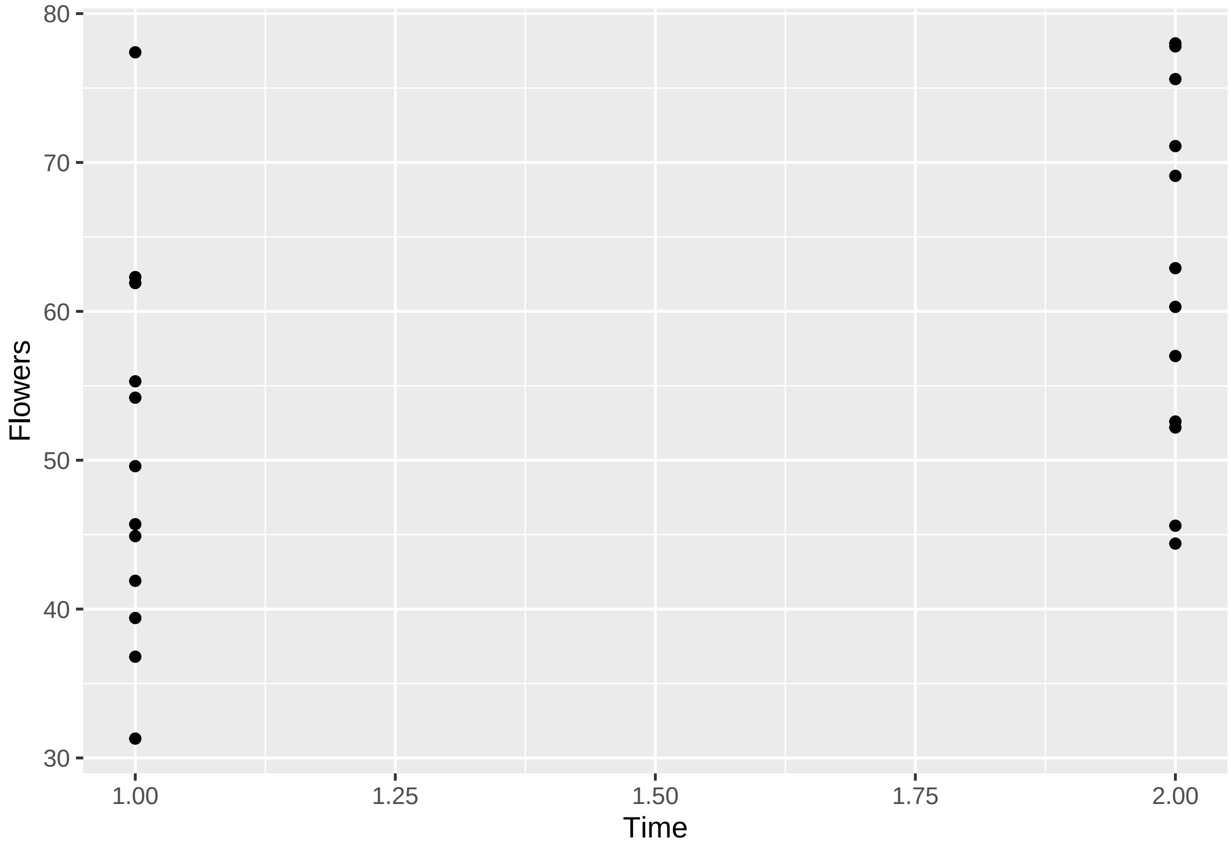
```
##   Flowers Time Intensity
## 1    62.3    1       150
## 2    77.4    1       150
## 3    55.3    1       300
## 4    54.2    1       300
## 5    49.6    1       450
## 6    61.9    1       450
```

```
?case0901
```

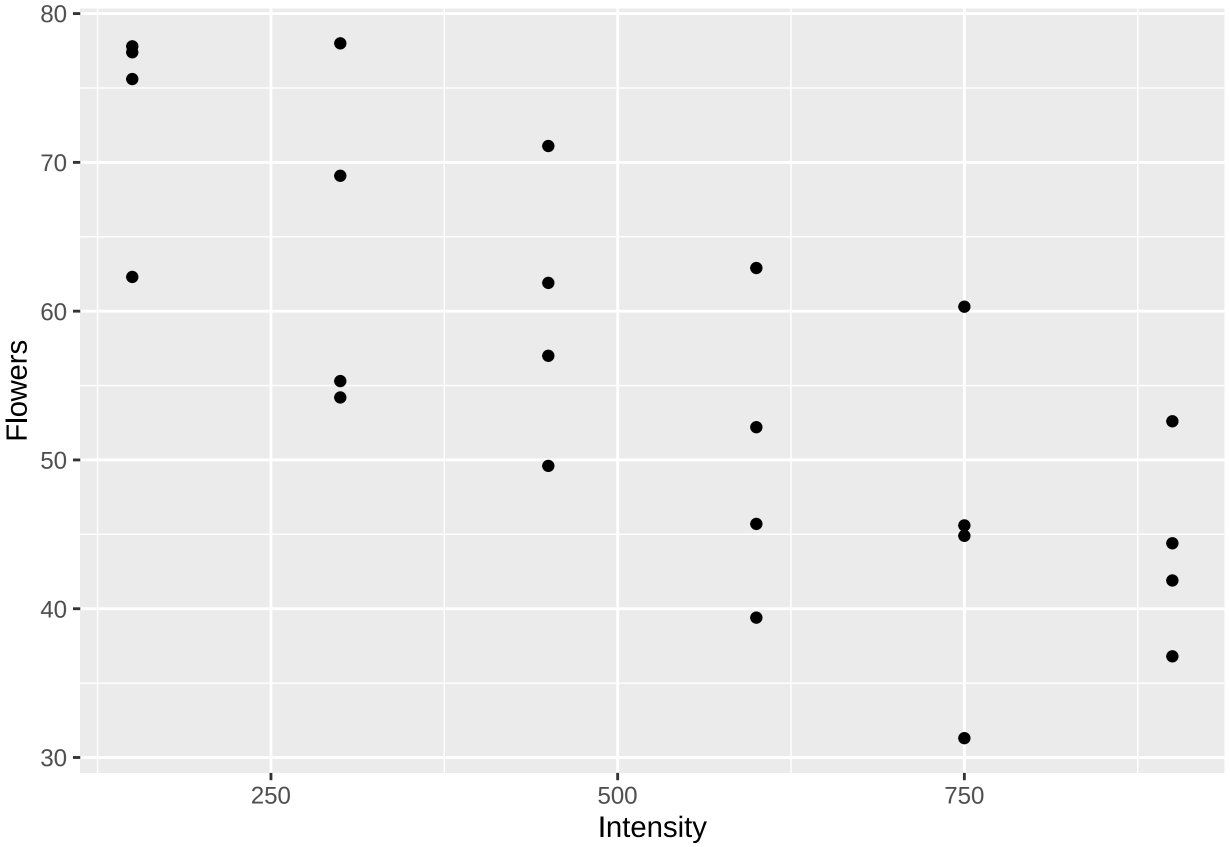
```
## starting httpd help server ... done
```

(a) Create plots of the response variable Flowers against each of the explanatory variables Time and Intensity (provide only your code). Do you think the late (1) or early (2) start time of light intensity regiments led to greater average number of flowers per meadowfoam plant? Do you think flower abundance increased or decreased as the light intensity treatment increased?

```
qplot(Time, Flowers , data = case0901)
```



```
#qplot(Flowers ~ Time, data = case0901, geom = 'boxplot')
qplot(Intensity, Flowers , data = case0901)
```



```
cor(case0901)
```

```
##              Flowers      Time  Intensity
## Flowers    1.0000000 0.4521766 -0.7711649
## Time       0.4521766 1.0000000  0.0000000
## Intensity -0.7711649 0.0000000  1.0000000
```

Looking at the plots, late has *slighly* higher average number of flowers, and as intensity increases, flower count decreases. The spread of flower responses to the time input is large, therefore it will be difficult to confirm statistical significance.

(b) Write out the the multiple linear regression model in statistical notation, where Flowers is the response, and Time and Intensity are the explanatory variables. Give the assumed distribution of the is. Then fit the model with lm(), and give sigma. Note: the Time term should be an indicator variable; you accomplish this in R by making it a factor.

- Y = Flowers
- X_1 = Time
- X_2 = Intensity

$$\mu(Y|X_1, X_2) = \beta_1 X_1 + \beta_2 X_2$$

### NOTE: There is no interaction term as the covariance of time to intensity is 0.

(c) Suppose we fit a model with an interaction term. In non-technical terms what will the interaction coefficient tell us? Suppose we fit separate models for Time = 1 and Time = 2 observations; what would we see if the interaction from the full model (which includes Time) is statistically significant?

(d) Give the model that includes an interaction term, and then fit it. Give the p-value from the test of this term's significance. What do you conclude?

2. (2 points) Now suppose a graduate student in your department tells you he has three observations (table below) he forgot to include in the original dataset.

|Obs.#| Flowers| Time| Intensity | |25 | 80 | 2| 150 | |26 | 80 | 2 | 900 | |27 | 40 | 2 | 1200 |

You can add these observations to your data with:

```r
case0901_updated <- rbind(case0901,
data.frame(Flowers = c(80, 80, 40), Time = c(2, 2, 2), Intensity = c(150, 900, 1200)))
```

(a) Create a Flowers vs. Intensity plot of the new 27 observation dataset.

(b) Of these three new observations, which has the greatest leverage? Explain why in non-technical terms, referencing the Flowers vs. Intensity plot.

(c) Of these three new observations, which has the greatest Cook's D statistic? Explain why in nontechnical terms, referencing the Flowers vs. Intensity plot.

(d) Of these three new observations, which has neither the greatest leverage nor the greatest Cook's D? Explain why in non-technical terms, referencing the Flowers vs. Intensity plot.