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
```

```
library(broom)
```

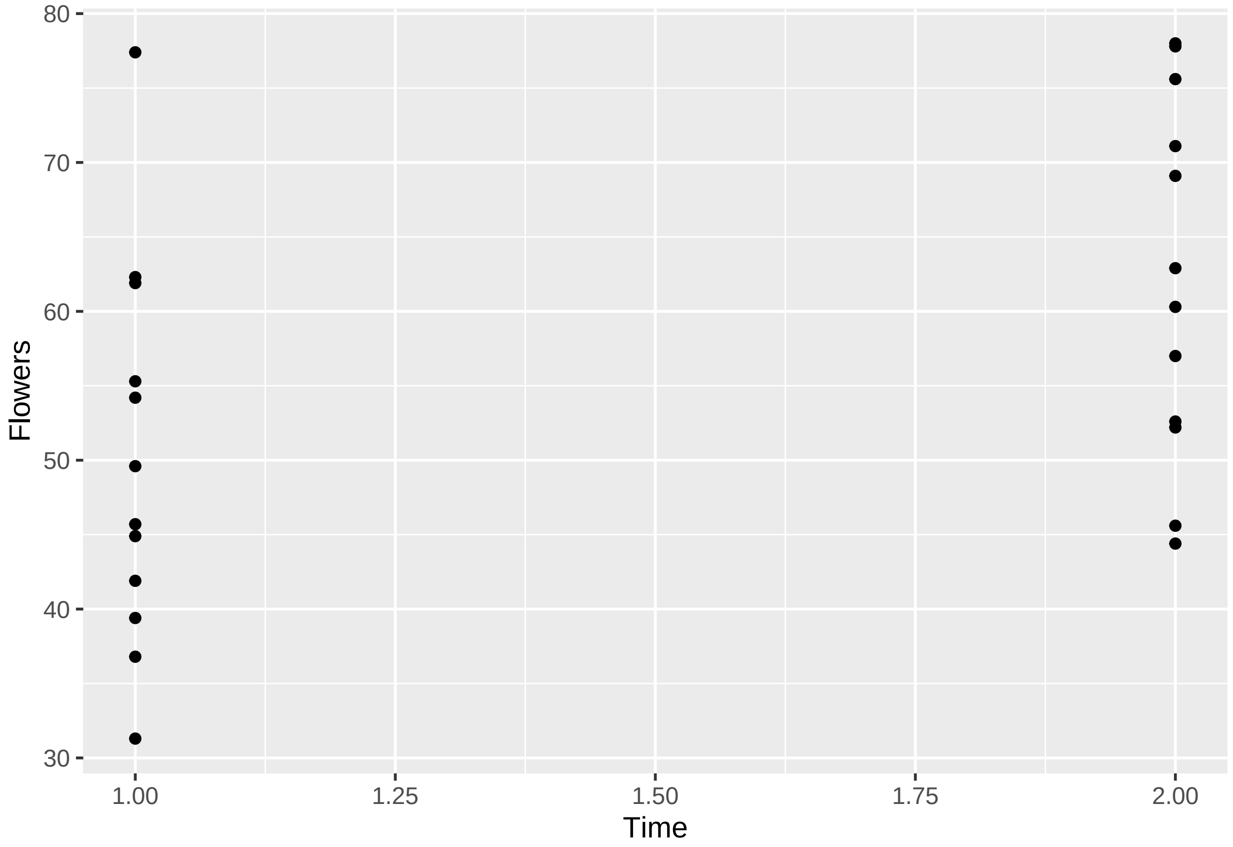
```
## Warning: package 'broom' was built under R version 4.0.3
```

```
flowerData <- case0901
head(flowerData)
```

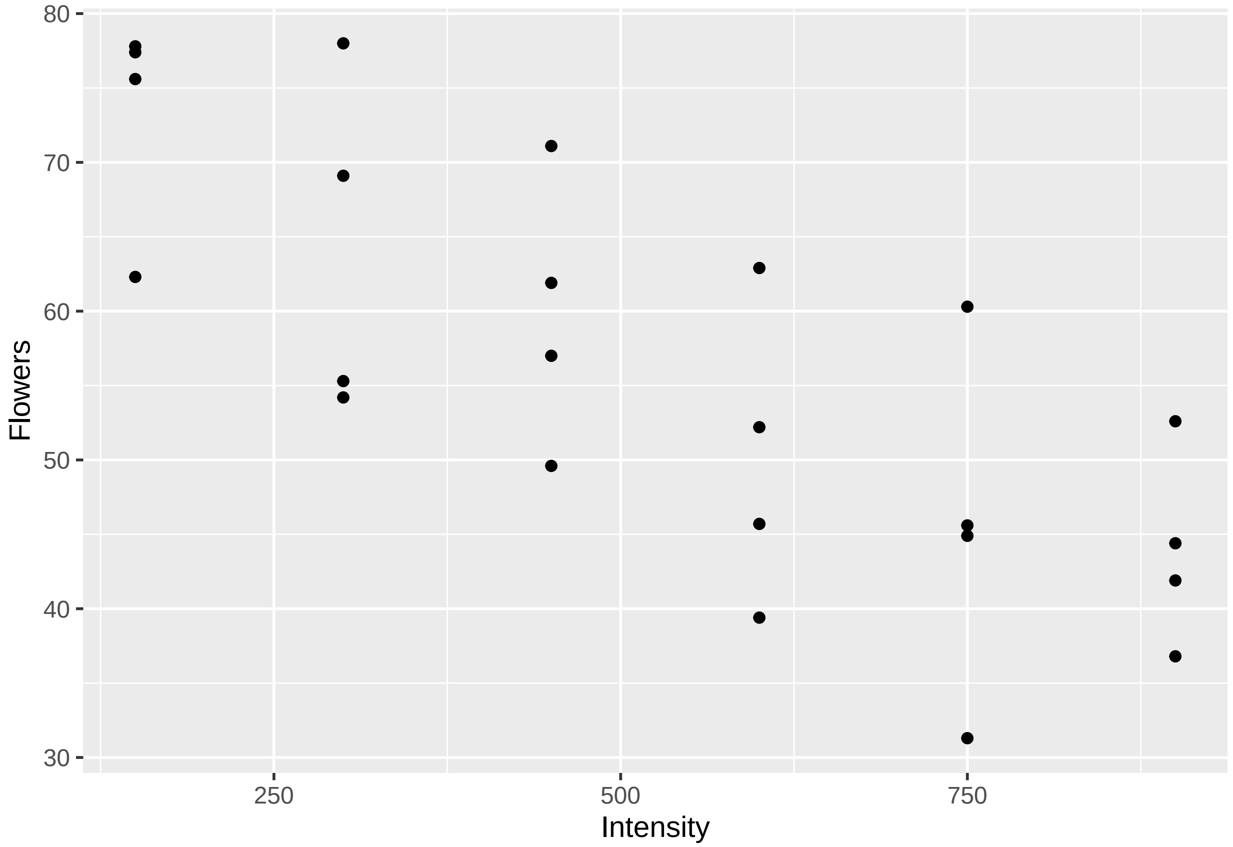
```
##   Flowers Time Intensity
## 1    62.3    1       150
## 2    77.4    1       150
## 3    55.3    1       300
## 4    54.2    1       300
## 5    49.6    1       450
## 6    61.9    1       450
```

(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 = flowerData)
```



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



```
cor(flowerData)
```

```
##              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 $\epsilon_i s$. 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

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

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

The assumed distribution of $\epsilon_i s$ is normal around zero.

$\hat{\sigma}$ = 6.441

```r
flowerData[’Timefactor’] <- factor(flowerData$Time)

fit <- lm(Flowers ~ Timefactor + Intensity, data = flowerData)

summary(fit)
```

```
##
## Call:
## lm(formula = Flowers ~ Timefactor + Intensity, data = flowerData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.652 -4.139 -1.558  5.632 12.165
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.305833   3.273772  21.781 6.77e-16 ***
## Timefactor2 12.158333   2.629557   4.624 0.000146 ***
## Intensity   -0.040471   0.005132  -7.886 1.04e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.441 on 21 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:   0.78
## F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

(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?

If the interaction is statistically significant, the interaction term will tell us the additional affect of intensity when plants recieve the late treatment compared to the early treatment.

If the interaction is not statistically significant the interaction term will equal zero.

If we fit two seperate models, one for Time = 1 and another for Time = 2, the model will have two coefficients (slope, and intercept), along with only one explanitory variable (Intensity). This differs from the original model with three coefficients and two explanitory variables (time and intensity). The estimates of the intensity coefficient will be identical between the original and seperate models, whereas the coefficients for each time parameter will obviously change.

As expected, the time = 2 model has a larger intercept value of 83.14 **(compared to** 71.62 **time = 1). This is a reaction to the the larger mean flower response of the time = 2 coefficient.**

```
flowerSplit <- split(flowerData, flowerData['Time'])
df1 <- data.frame(flowerSplit[1])
df2 <- data.frame(flowerSplit[2])

fitTime1 <- lm(X1.Flowers ~ X1.Intensity, data = df1)
summary(fitTime1)
```

```
##
## Call:
## lm(formula = X1.Flowers ~ X1.Intensity, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.516 -4.276 -2.220  4.874 11.938
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.623333   4.688128  15.278 2.93e-08 ***
## X1.Intensity -0.041076   0.008025  -5.118 0.000452 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.122 on 10 degrees of freedom
## Multiple R-squared:  0.7237, Adjusted R-squared:  0.6961
## F-statistic:  26.2 on 1 and 10 DF,  p-value: 0.0004518
```

```
fitTime2 <- lm(X2.Flowers ~ X2.Intensity, data = df2)
summary(fitTime2)
```

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

```
## lm(formula = X2.Flowers ~ X2.Intensity, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2067 -3.9067 -0.4667  5.4733  7.0533
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  83.146667   3.968633  20.951 1.36e-09 ***
## X2.Intensity -0.039867   0.006794  -5.868 0.000158 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.029 on 10 degrees of freedom
## Multiple R-squared:  0.775,  Adjusted R-squared:  0.7525
## F-statistic: 34.44 on 1 and 10 DF,  p-value: 0.0001577
```

(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?

```
fitInteraction <- lm(Flowers ~ Timefactor + Intensity + Timefactor*Intensity, data = flowerData)
summary(fitInteraction)
```

```
##
## Call:
## lm(formula = Flowers ~ Timefactor + Intensity + Timefactor *
##     Intensity, data = flowerData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.516 -4.276 -1.422  5.473 11.938
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           71.623333   4.343305  16.491 4.14e-13 ***
## Timefactor2           11.523333   6.142360   1.876   0.0753 .
## Intensity             -0.041076   0.007435  -5.525 2.08e-05 ***
## Timefactor2:Intensity  0.001210   0.010515   0.115   0.9096
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.598 on 20 degrees of freedom
## Multiple R-squared:  0.7993, Adjusted R-squared:  0.7692
## F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07
```

Pvalue of time and intensity is 0.91 suggesting we accept the null hypothesis that there is no interaction.

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.

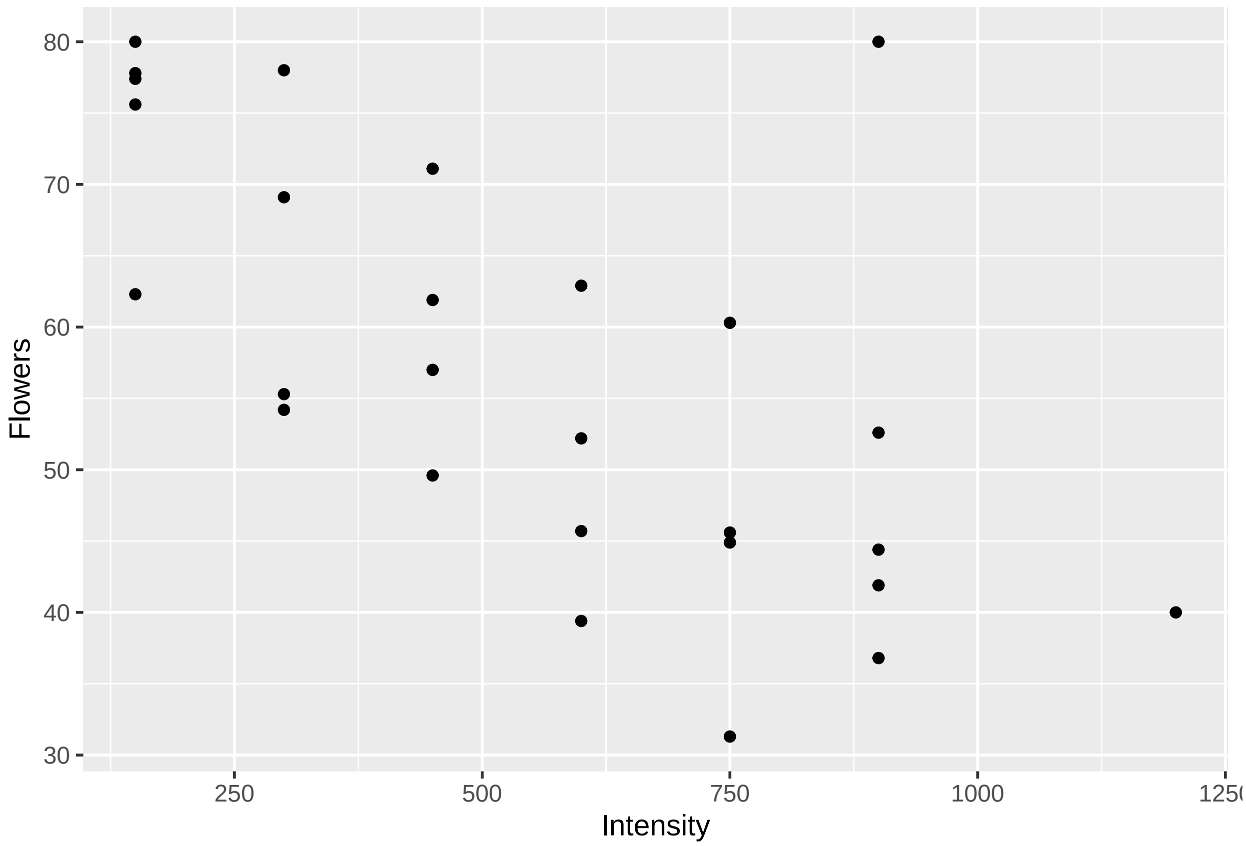
|Obs.#| Flowers| Time| Intensity | |25 | 80 | 2| 150 | |26 | 80 | 2 | 900 | |27 | 40 | 2 | 1200 |

You can add these observations to your data with:

```
flowerData_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.

```
qplot(Intensity, Flowers, data = flowerData_updated)
```



```
augFit <- lm(Flowers ~ factor(Time) + Intensity, data = flowerData_updated)

#Calc leverage
flowerData_updated_aug <- augment(augFit)
flowerData_updated_aug[flowerData_updated_aug$.hat == max(flowerData_updated_aug$.hat),] # max leverage
```

```
## # A tibble: 1 x 8
##   Flowers ‘factor(Time)‘ Intensity .fitted .std.resid  .hat .sigma .cooksd
##     <dbl> <fct>              <dbl>   <dbl>      <dbl> <dbl>  <dbl>   <dbl>
## 1      40 2                   1200    41.2     -0.159 0.241   8.86 0.00265
```

```
flowerData_updated_aug[flowerData_updated_aug$.cooksd == max(flowerData_updated_aug$.cooksd),] # max le
```

```
## # A tibble: 1 x 8
##   Flowers ‘factor(Time)‘ Intensity .fitted .std.resid  .hat .sigma .cooksd
##     <dbl> <fct>              <dbl>   <dbl>      <dbl> <dbl>  <dbl>   <dbl>
## 1      80 2                    900    51.6       3.48 0.114   6.25   0.520
```

```
flowerData_updated_aug
```

```
## # A tibble: 27 x 8
##    Flowers ‘factor(Time)‘ Intensity .fitted .std.resid   .hat .sigma  .cooksd
##      <dbl> <fct>              <dbl>   <dbl>      <dbl>  <dbl>  <dbl>    <dbl>
##  1    62.3 1                    150    63.1    -0.0996 0.145    8.86 0.000561
##  2    77.4 1                    150    63.1     1.78   0.145    8.25 0.180
##  3    55.3 1                    300    57.9    -0.315  0.106    8.84 0.00390
##  4    54.2 1                    300    57.9    -0.449  0.106    8.82 0.00792
##  5    49.6 1                    450    52.7    -0.370  0.0858   8.84 0.00428
##  6    61.9 1                    450    52.7     1.11   0.0858   8.63 0.0388
##  7    39.4 1                    600    47.5    -0.971  0.0858   8.69 0.0295
##  8    45.7 1                    600    47.5    -0.211  0.0858   8.85 0.00139
##  9    31.3 1                    750    42.2    -1.33   0.106    8.53 0.0698
## 10    44.9 1                    750    42.2     0.325  0.106    8.84 0.00415
## # ... with 17 more rows
```

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

The observation of (Flowers: 40, Time: 2, Intensity: 1200) has the greatest leverage. There are a few points that are further from the data set (Ex. 80, 2, 900) but the (40,2,1200) has the most *leverage* because it has the most unique combination of values. 40 is a very low value for flowers, and 1200 is a very high value for intensity. The point with the largest cook's distance (80,2,900) has a unique flowers value, but a relatively common Intensity value.

(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.

The observation of (Flowers: 80, Time: 2, Intensity: 900) has the greatest cook's distance. This point will have the greatest influence on the regression model because it is far away from the estimate line, *and* it is far away from the other points. If this point is removed there will be a large impact on the regression model.

(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.

The remaining point is (Flowers: 80, Time: 2, Intensity: 150). This point has low cook's D because if it were removed, the model would only minimally change. It has low leverage because it does not have a unique combination of values and is surrounded by 3 other points.