ST517-HW4

Nora Quick

1. Consider the `Sleuth3` dataset `case0901`, which contains the results of a study of the Meadow-foam plant. Familiarize yourself with the study and load the data.

```
# ?case0901
# case0901
# head(case0901)
```

   (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(Intensity, Flowers, data = case0901)
```

Based on the graph comparing Flowers and Time it appears that the early (=2) start time of light intensity regiments lead to a grater acerage number of flowers.

Based on the graph comparing Flowers to Intensity it appears that the flower abundance decreased ad the light intensity treatment increased.

   (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 $\hat{}\sigma$. Note: the Time term should be an indicator variable; you accomplish this in R by making it a factor.

```
flower_fit1 <- lm(Flowers ~ Intensity + factor(Time), data = case0901)
summary(flower_fit1)
```

```
##
## Call:
## lm(formula = Flowers ~ Intensity + factor(Time), data = case0901)
##
## 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 ***
## Intensity     -0.040471   0.005132  -7.886 1.04e-07 ***
## factor(Time)2 12.158333   2.629557   4.624 0.000146 ***
## ---
```

```
## 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
```

The statistical notation of this is Flowers = $\beta 0 + \beta 1$Time + $\beta 2$Intensiry. The assumed distribution is a Normal Distribution. The sigma of the fit is about 6.44.

(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 we fit the model with an interaction term we will be indicating that one of the explanatory relies on the other. In the case of this we would indicate that Time is an indication of how much Intensity would affect the Flower. In other words the intensity doesn't matter until we say how much time that intensity was applied.

If we seperated the models into Time = 1 and Time = 2 we would probably see that intensity doesn't affect the flowers as much if they are all started later. But with high intensity early on the intensity likely affects the flowers. I believe that separating them out would move the model closer to a Normalized Distribution.

(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?

```
flower_fit2 <- lm(Flowers ~ Intensity * factor(Time), data = case0901)
summary(flower_fit2)
```

```
##
## Call:
## lm(formula = Flowers ~ Intensity * factor(Time), data = case0901)
##
## 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 ***
## Intensity               -0.041076   0.007435  -5.525 2.08e-05 ***
## factor(Time)2           11.523333   6.142360   1.876   0.0753 .
## Intensity:factor(Time)2  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
```

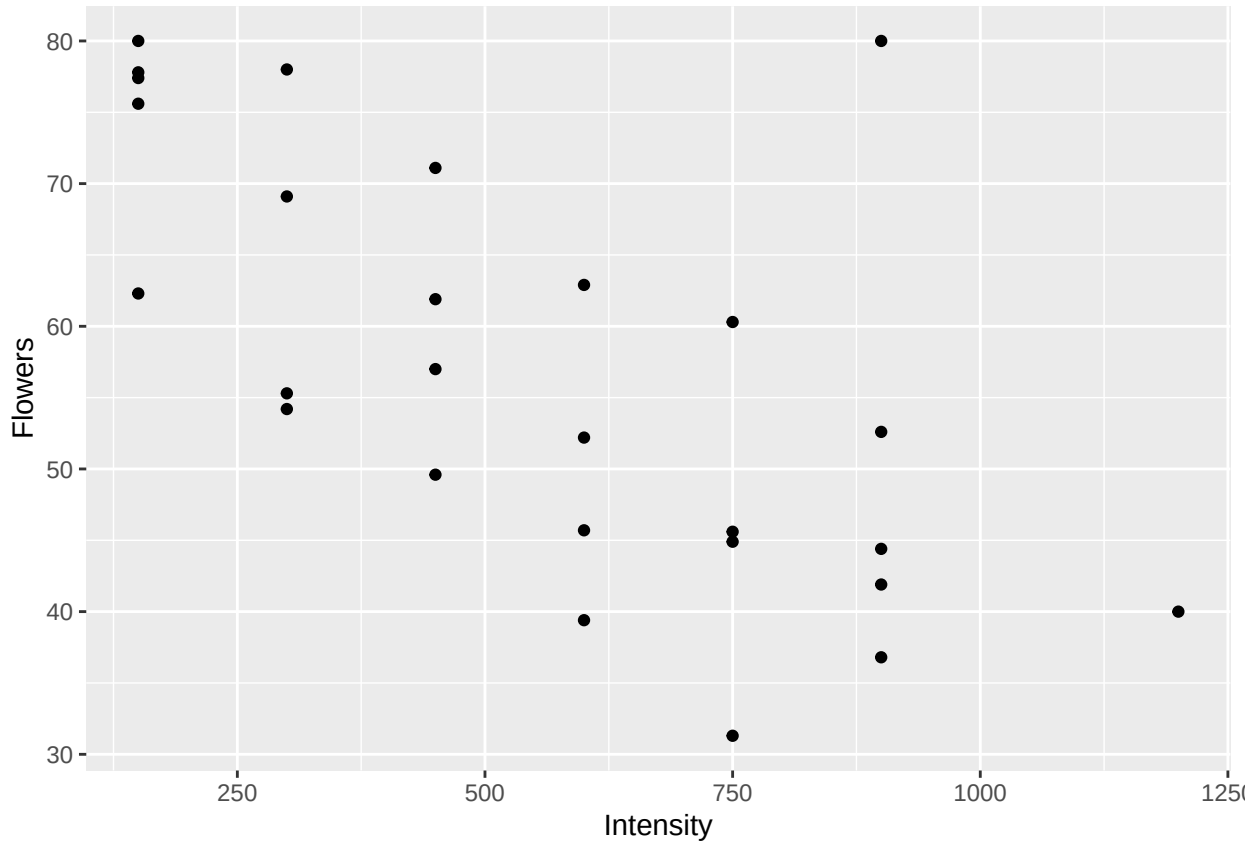
The p-value of the variable is about 0.91. I conclude that there's weak evidance that the two are independent. In other words Time and Intensity do not intaract with Flowers independently.

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

```
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.

```
qplot(Intensity, Flowers, data = case0901_updated)
```



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

As we can see from the Flowers vs. Intensity plot the point that has the greatest leverage is the point at flower = 40, intensity = 1200. We can also find this by fitting the updated data, augmenting it, and checking .hat for >2.

This has the largest leverage because it is farther away and solo from any other intensity. We can also see from the code provided that it is simply the largest Intensity added to the new data with a low Flower value. High Intensity and and lower Flowers provides the greatest leverage.

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

As we can see from the Flowers vs. Intensity plot the point that has the greatest Cook's D statistic is flower = 80, intensity = 900. We can also find this by fitting the updated data, augmentint it, and checking .cooksd for >2.

This is the largest Cook's D because it is the furthest away from it's row in intensity 900. We can also see from the code provided that it is the middle term added to Intensity but it is one of the two higher Flowers values. So semi-high Intensity and high Flowers provides greatest Cook's D.

(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.

As we can see from the Flowers vs. Intensity plot the point that has neither the greatest leverage nor the greatest Cook's D is the point at flower = 80, intensity = 150.

This is neither the greatest leverage nor the greatest Cook's D because it is the lighest Flower at the lowest Intensity. We can also see this in the updated code. Because it is lowest Intensity and highest Flower it doesn't fall under leverage or Cook's D.