Business Analytics - Assignment 1 **Nicolas Tachet - nt2479**
Prof. Jacob Leshno
Due Date: 09/22/17

As suggested on the assignement, I submitted two files (my R code and this pdf with my answers). Don't hesitate to refer to my R code for more clarity.

Linear Regression

1)Load the data from Crops.csv into R. You can use setwd() to set the current working directory. Print a summary of the variables.

We load the data using the read.csv method and then display the summary:

```
     Yield              Water          Herbicide        Fertilizer
appropriateFertilizer
 Min.   :-0.1437   Min.   : 5.0    Min.   : 1.0    Min.   : 1.00
Min.   :0.0000
 1st Qu.:12.1552   1st Qu.:15.0    1st Qu.: 3.0    1st Qu.: 3.75
1st Qu.:0.0000
 Median :16.2610   Median :27.5    Median : 5.5    Median : 6.50
Median :0.0000
 Mean   :17.2652   Mean   :27.5    Mean   : 6.5    Mean   : 6.50
Mean   :0.3333
 3rd Qu.:20.8657   3rd Qu.:40.0    3rd Qu.: 8.0    3rd Qu.: 9.25
3rd Qu.:1.0000
 Max.   :41.0415   Max.   :50.0    Max.   :20.0    Max.   :12.00
Max.   :1.0000
```

Thanks to this brief summary we see there is a huge difference between the smallest crops and the longest one. Even if these values might be outliers, we can definitely see there might exist a good "recipe" to harvest big crops. Let's try to find it.

2) Regress the yield on the amount of water used. Explain and interpret the results.

Here we want to regress the yield on the amount of water used. Let's plot this regression and the yield against the amount of water to see how it fits.

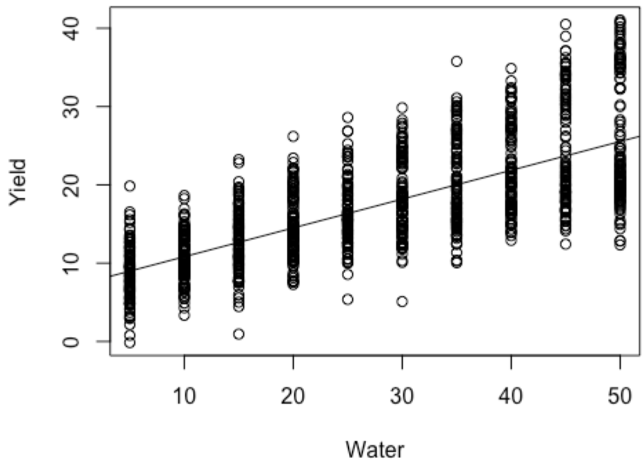

Figure 1: Plot Yield against Water/Regression

It seems that it follows the trend in a good way. The summary of the regression suggests the same thing:

```
Call:
lm(formula = Yield ~ Water)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2156  -3.8619  -0.8595   3.4678  16.8022

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.14521    0.32533   21.96   <2e-16 ***
Water        0.36800    0.01049   35.09   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 5.217 on 1198 degrees of freedom
Multiple R-squared:  0.5069,    Adjusted R-squared:  0.5065
F-statistic:  1232 on 1 and 1198 DF,  p-value: < 2.2e-16
```

Indeed, both t-values are rather important which means that p-values are really low. Thus the coefficents are significant. Our F-Statistic score is rather important which suggests the same thing.
If we look at the R-squared, we see that this is not really close to 1. However, we saw during a previous class that you could make interesting predictions even with a smal R-squared. Thus, it is a good start.

3) Regress the yield on the amount of fertilizer used. Explain and interpret the results.

Let's do the same with the amount of fertlizer.

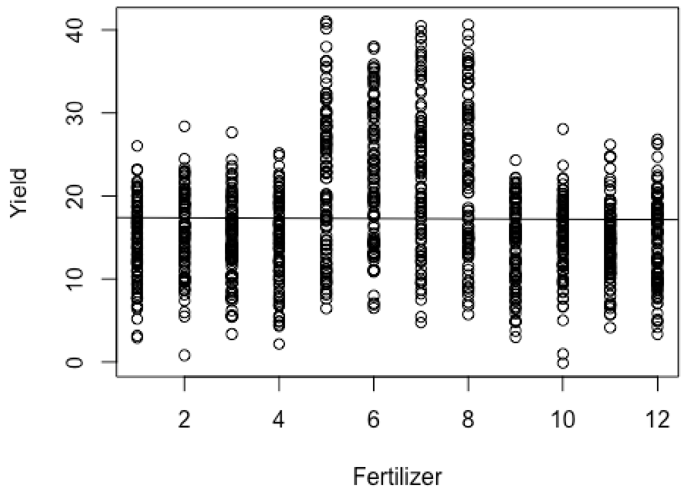

Figure 2: Plot Yield against Fertilizer/Regression

Obviously, a linear regression does not explain the model and it is going to be more difficult to make this variable relevant. The summary of the regression suggests the same thing:

```
Call:
lm(formula = Yield ~ Fertilizer)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3315  -5.1610  -0.9798   3.5892  23.7430

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.40913    0.45721  38.077   <2e-16 ***
Fertilizer  -0.02214    0.06212  -0.356    0.722
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 7.429 on 1198 degrees of freedom
Multiple R-squared:  0.000106,  Adjusted R-squared:  -0.0007286
F-statistic: 0.127 on 1 and 1198 DF,  p-value: 0.7216
```

Here, the F-statistic is really low. A linear regression on this variable is not relevant. However, it seems that for certain values of fertilizer, the yield is more important. This suggests that with a little piece of work we could make this variable more interesting.

4) Regress the yield on the amount of herbicide used. Explain and inter-

pret the results.

Let's do the same with the amount of Herbicide.

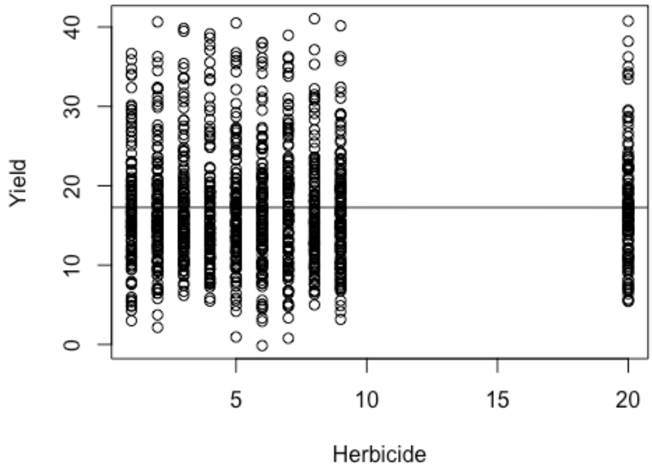

Figure 3: Plot Yield against Herbicide/Regression

Obviously, a linear regression does not explain the model. The summary of the regression suggests the same thing:

```
Call:
lm(formula = Yield ~ Herbicide)

Residuals:
     Min       1Q   Median       3Q      Max
-17.410   -5.111   -1.009    3.598   23.778

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.273662   0.346445  49.860   <2e-16 ***
Herbicide   -0.001298   0.041859  -0.031    0.975
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 7.429 on 1198 degrees of freedom
Multiple R-squared:  8.022e-07, Adjusted R-squared:  -0.0008339
F-statistic: 0.000961 on 1 and 1198 DF,  p-value: 0.9753
```

Here, the F-statistic is really close to 0. A linear regression on this variable is not relevant. Moreover, it seems impossible to find any pattern with this variable that could help us.

5)Regress the yield on all the variables. Explain and interpret the results.

If we regress the yield on all the variables, we have the following result:

```
Call:
lm(formula = Yield ~ Herbicide + Fertilizer + Water)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2898  -3.9300  -0.8841   3.4496  16.8113

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.297540   0.472309  15.451   <2e-16 ***
Herbicide    -0.001298   0.029415  -0.044    0.965
Fertilizer   -0.022138   0.043657  -0.507    0.612
Water         0.368001   0.010494  35.068   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 5.221 on 1196 degrees of freedom
Multiple R-squared:  0.507,     Adjusted R-squared:  0.5058
F-statistic:   410 on 3 and 1196 DF,  p-value: < 2.2e-16
```

Only two variables are significant here, the same variables than the ones from our first model (the Water model). Fertilizer and Herbicides are not relevant here (same result as in questions 3 and 4). The F-Statistic is smaller than in the first model which means that adding non relevant values weakened the model instead of improving it.

6) The farmer suspects that high levels of fertilizer may not be effective. To check this conjecture, plot the yield against the amount of fertilizer used. Explain why the plot is consistent with the regression results.

Again let's plot the yield against the amount of fertilizer.

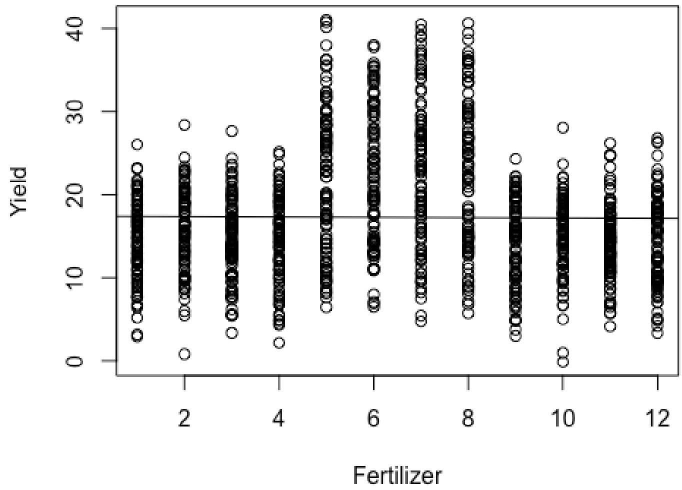

Figure 4: Plot Yield against Fertilizer/Regression

We clearly see that for a range of Fertilizer from 5 to 8, the yield is greater. The farmer is right: high or low levels of fertilizers are ineffective. Actually, this explains why the regression was not relevant. Indeed, the Fertilizer variable is more a binary input: either it is effective or it is not. A regression is useless in this case. To this extent, the plot is consistent with the regression.

7) Based on the plot, create an indicator appropriateFertilizer whose value is 1 when the amount of fertilizer is appropriate, and 0 when the amount of fertilizer is too high or too low. Regress Yield on the indicator you created and interpret the results.

We are going to select only the crops with the right fertilizer values by creating a new dummy variable called appropriateFertilizer. Let's regress the yield on the indicator we created.

```
Call:
lm(formula = Yield ~ crops$appropriateFertilizer)

Residuals:
     Min       1Q   Median       3Q      Max
-17.8090  -4.4348   0.3434   4.3392  18.5079

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  14.6311     0.2272   64.40   <2e-16 ***
crops$appropriateFertilizer   7.9025     0.3935   20.08   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 6.426 on 1198 degrees of freedom
```

```
Multiple R-squared:  0.2519,    Adjusted R-squared:  0.2512
F-statistic: 403.3 on 1 and 1198 DF,  p-value: < 2.2e-16
```

It's getting better. We eventually found a usefulness for the Fertilizer variable. Indeed, the p-value is really low which means the variable we created is significant as shown by this plot.

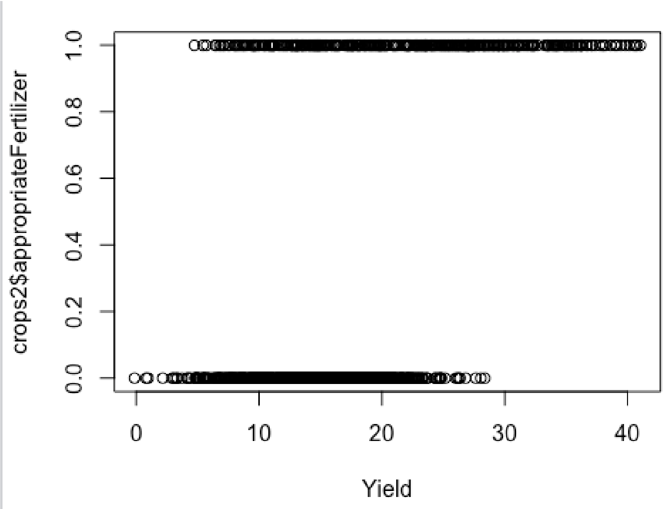

Figure 5: Plot appFertilizer against Yield

8) The farmer suggests that an appropriate amount of fertilizer should raise the effectivness of watering the crops. Run a regression with an interaction between water and appropriateFertilizer to check this. Interpret the results.

The farmer suggests that an appropriate amount of fertilizer should raise the effectivness of watering the crops. Thus, let's see if it exists a correlation between appropriateFertilizer and Water.

```
Call:
lm(formula = crops$Yield ~ crops$Water * crops$appropriateFertilizer,
    data = crops2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3417  -1.9775  -0.1309   1.9996  11.2450

Coefficients:
                                      Estimate Std. Error t value Pr(>|t
(Intercept)                           7.270364   0.231286  31.435
<2e-16 ***
crops$Water                           0.267662   0.007455  35.904
<2e-16 ***
crops$appropriateFertilizer          -0.375471   0.400599  -0.937
0.349
```

```
crops$Water:crops$appropriateFertilizer   0.301016    0.012912   23.312
<2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3.028 on 1196 degrees of freedom
Multiple R-squared:  0.8341,    Adjusted R-squared:  0.8337
F-statistic:  2005 on 3 and 1196 DF,  p-value: < 2.2e-16
```

As expected, there is a correlation between our variables (the t-value is equal to 23.312 which is a rather high value). However, the variable appropriateFertilizer is no longer relevant. We should remove it from our model.

9) Select a collection of variables and interaction terms to use as predictors for the yield. Run the regression, and interpret the results. Explain why you chose this regression model.

Thanks to question 9, we know now that we should use the same variables but we should remove the appropriateFertilizer (just let the correlated variable). As we know that Herbicide is not relevant for the model, let's just forget it. For our new regresion, we have the following summary:

```
Call:
lm(formula = Yield ~ Water + Water:appropriateFertilizer, data = crops2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2702  -1.9420  -0.1097   1.9924  11.3523

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.145207   0.188834   37.84   <2e-16 ***
Water                       0.271238   0.006404   42.35   <2e-16 ***
Water:appropriateFertilizer 0.290288   0.005977   48.57   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3.028 on 1197 degrees of freedom
Multiple R-squared:  0.834,     Adjusted R-squared:  0.8337
F-statistic:  3007 on 2 and 1197 DF,  p-value: < 2.2e-16
```

We see that the F-Statistic is 3007 which is far more greater than for the previous regression. We improved the model by removing the appropriateFertilizer variable.

10) For this model, what is a 99% confidence interval for the regression coefficients. Interpret the results.

For this model, here is the 99% confidence interval:

```
> confint(fitting ,level=0.99)
                                 0.5 %     99.5 %
(Intercept)                  6.6580255  7.6323893
Water                        0.2547149  0.2877612
Water:appropriateFertilizer  0.2748678  0.3057085
```

Thus, there is 99% probability that the calculated confidence interval from some future experiment encompasses the true value of the parameters we are trying to evaluate.

11) For this model, what is a 90% prediction for the yeild of a single sample that gets water=30,fertilizer =5 and herbicide =5? Interpret the results.

Here, we test our model with a random set of inputs (we call it "testdata"). This function provides us the yield of the crops that we might expect from this set of inputs according to our model.

```
> predict(fitting, testdata, interval="predict", level=0.9)
       fit       lwr      upr
1 23.99099  19.00036  28.98163
```

According to this result, there is 90% chance that the yield of the future crops belongs to the interval $[19.000; 28.98]$