AIR Data Analysis Q1

Leonid Rempel

4/9/2020

Introduction

Hello AIR! Here I will be answering Quesion 1 with R, and I will use R markdown to make it look nice while showing you what I did. You may also run the R document attatched.

Lets first pull up the Iris dataset and summarize what we have. Before I do this, I will also load ggplot2 with some thematic options, in addition to stargazer, to make everyting look as beautiful as possible.

```
### Checking for and loading our packages
packages <- c("ggplot2", # For graphs
              "ggthemes", # For thematic purposes
              "stargazer", # For charts/regression output
              "datasets")

for (i in 1:length(packages)) {
    if (!packages[i] %in% rownames(installed.packages())) {
      install.packages(packages[i]
                       , repos = "http://cran.rstudio.com/"
                       , dependencies = TRUE
                       )
    }
    library(packages[i], character.only=TRUE)
  }
```

```
### loading iris
data("iris")
stargazer(iris, align = TRUE, header = FALSE, title = "Summary")
```

Table 1: Summary

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Sepal.Length	150	5.843	0.828	4.300	5.100	6.400	7.900
Sepal.Width	150	3.057	0.436	2.000	2.800	3.300	4.400
Petal.Length	150	3.758	1.765	1.000	1.600	5.100	6.900
Petal.Width	150	1.199	0.762	0.100	0.300	1.800	2.500

```
ftable(iris$Species)
```

setosa versicolor virginica

```
50         50         50
```

a) From the summary we see that there are 50 irises per specie.

Part b) Scatterplot

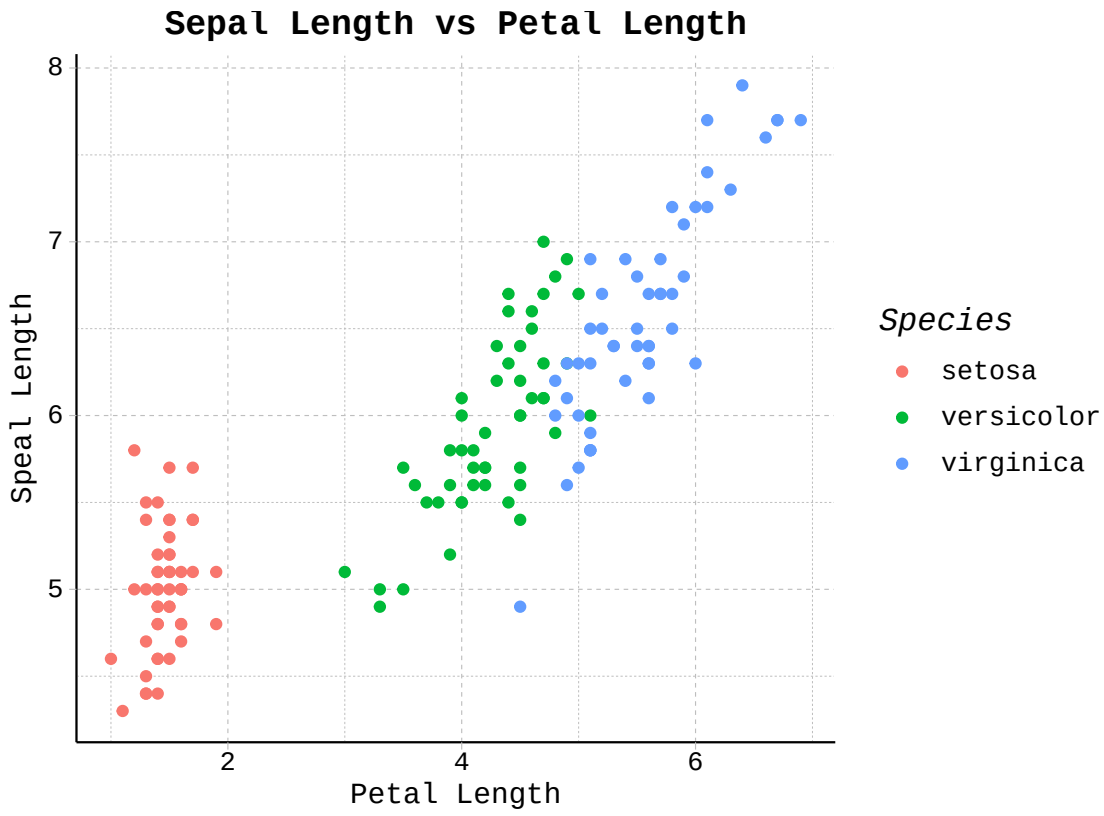
```
ggplot(iris, aes(x = Petal.Length, y = Sepal.Length)) +

  #coloring according to species
  geom_point(aes(color = factor(Species))) +

  #my preferred theme
  theme_pander(base_family = 'mono') +

  #labeling the graph
  labs(x = 'Petal Length', y = 'Speal Length',
           title = 'Sepal Length vs Petal Length',
           color = 'Species') +

  #this is just formatting the graph a little more
  theme(axis.title.x = element_text(size = 11, vjust=0, hjust = 0.5, family = 'mono'),
        axis.title.y = element_text(size = 11, vjust=2, hjust = 0.5, family = 'mono'),
        plot.title = element_text(size = 13, hjust = 0.5, vjust = 1),
        plot.margin = unit(c(0.5, 1, 0.5, 1), "cm"),
        axis.line = element_line(linetype = 'solid', size = 0.4),
        plot.caption = element_text(size = 10, hjust = 0),
        axis.text.y = element_text(hjust = 0.5))
```



Lets make our observations now! As we can see, there seems to be an overall positive relationship between petal length and sepal length. Furthermore, the strength of this relationship varies by specie. As for the sepal and petal lengths, the virginicia has the largest measurements followed by the versicolor and the setosa.

Part c) Multiple Regression

Our next task is a multiple regression of petal length, petal width and sepal width on sepal length. Lets see if taking account of these extra variables changes the relationship exhibited above in the scatterplot.

```
options(scipen = 3) # 3 decimal level percision
# our model
reg <- lm(Sepal.Length ~ Petal.Length + Petal.Width + Sepal.Width, iris)

# this is the regression output
stargazer(reg, title="Multiple Regression Results", align = TRUE, header = FALSE)
```

Table 2: Multiple Regression Results

	Dependent variable:
	Sepal.Length
Petal.Length	0.709^{***}
	(0.057)
Petal.Width	-0.556^{***}
	(0.128)
Sepal.Width	0.651^{***}
	(0.067)
Constant	1.856^{***}
	(0.251)
Observations	150
R^2	0.859
Adjusted R^2	0.856
Residual Std. Error	0.315 (df = 146)
F Statistic	295.539^{***} (df = 3; 146)
Note:	*p<0.1; **p<0.05; ***p<0.01

Part d) Interpretation of Multiple Regression

Looking at the above table, we see that, holding petal width and sepal width constant, a 1 cm increase in petal length is associated with, on average, a 0.709 cm increase in sepal length. This effect is statistically significant beyond the 1% level.

Part e) EXTRA CREDIT

I would like to see the effect by species. As we saw in the scatter plot, the relationship is different across species, so lets look at how different it is.

Table 3: Multiple Regression Results with Species

	Dependent variable:
	Sepal.Length
Petal.Length	0.829^{***}
	(0.069)
Petal.Width	-0.315^{**}
	(0.151)
Sepal.Width	0.496^{***}
	(0.086)
Speciesversicolor	-0.724^{***}
	(0.240)
Speciesvirginica	-1.023^{***}
	(0.334)
Constant	2.171^{***}
	(0.280)
Observations	150
R^2	0.867
Adjusted R^2	0.863
Residual Std. Error	0.307 (df = 144)
F Statistic	188.251^{***} (df = 5; 144)
Note:	*p<0.1; **p<0.05; ***p<0.01

R chose the setosa as our base categorical variable. Thus, all else equal, we see that Versicolors have, on average, a Sepal length that is 0.724 cm smaller than the Setosa and Virginicas have a sepal length that is, on average, 1.023 cm smaller than the Setoa. The effect of petal length on predicting speal length is still significant and now predicts a 0.829 cm increase in sepal length for each 1 cm increase in petal length, all else equal.

That concludes Q1. The other two questions are done in Python.