An Exploration into Three Closely Related Varieties of Iris

Jordan Schupbach

March 16, 2017

1 Introduction

This famous iris data set, collected by Anderson (Anderson 1936) and used by Fisher (Fisher 1936), gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. They are extremely similar species of iris, but do have some blank differences. In Plot 1 we give images of the three iris species.

2 Description of Data

This dataset has 1 categorical variable and 4 quantitative variables. Whether a variable is the explanatory variable or the dependent variable may depend on the model being considered motivated by some question of interest. Before looking at any models, we want to get familiar with the data. In the very least, we will want to summarize the data either with plots or with summary statistics. The first few lines of the dataset are given below:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.10	3.50	1.40	0.20	setosa
4.90	3.00	1.40	0.20	setosa
4.70	3.20	1.30	0.20	setosa
4.60	3.10	1.50	0.20	setosa
5.00	3.60	1.40	0.20	setosa
5.40	3.90	1.70	0.40	setosa
4.60	3.40	1.40	0.30	setosa

Table 1: First 7 rows of the dataset

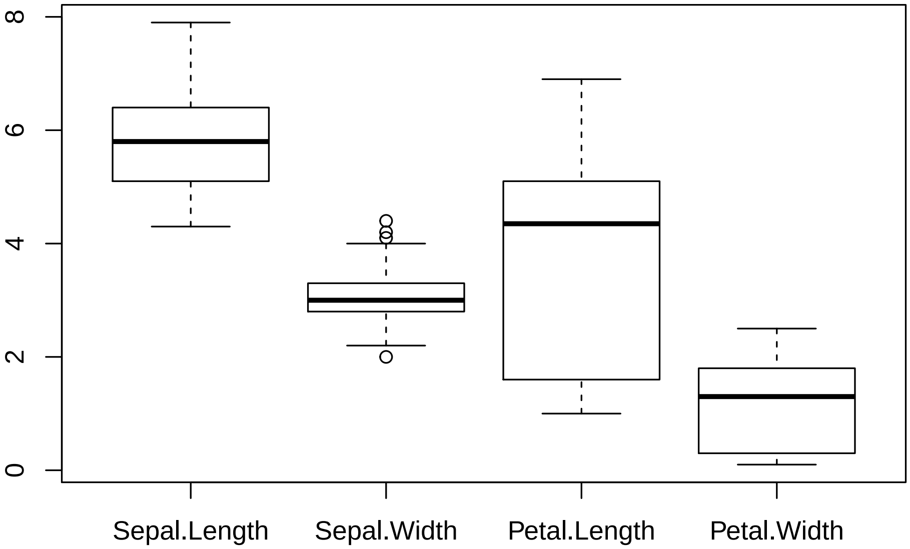
First, let's explore the whole dataset. We can print summary tables by each variable using the stargazer function out of a package by the same name. This is the output:

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

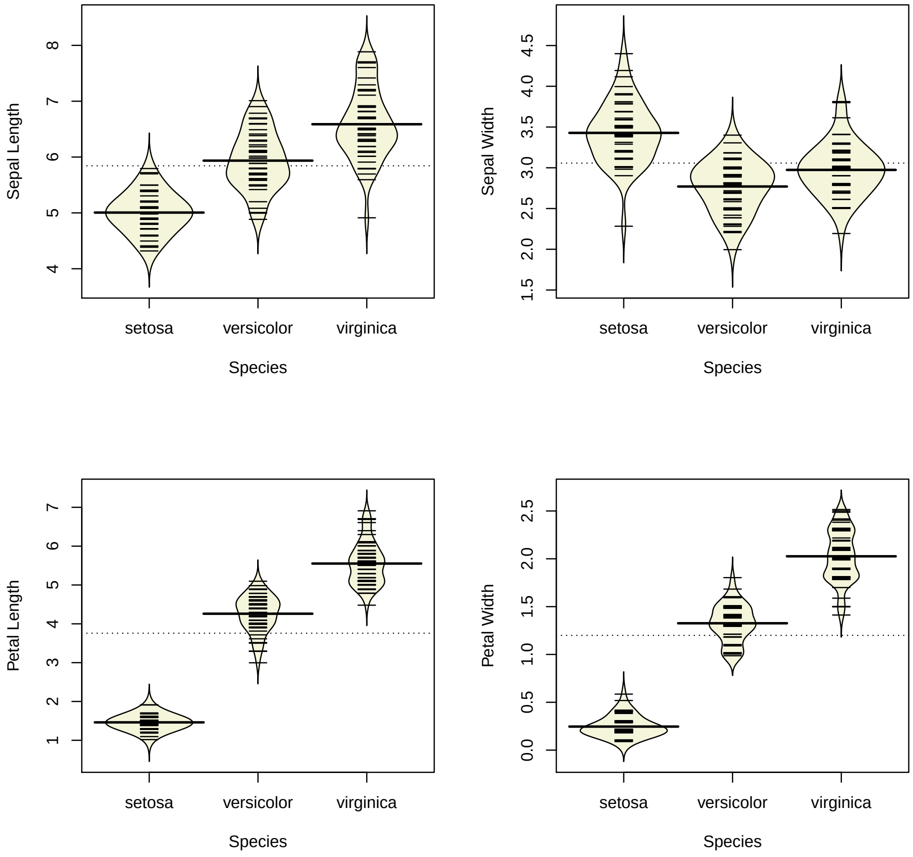
Table 2: Caption here

In Table 2, we see the summary output of each variable. We also may want to subset our data by species and look at summary statistics. Consider the beanplot's below.

We can also visualize these summary statistics with a box-plot. The box-plot given in Figure blank



In Table 2 we see the summary output of each variable. We also may want to subset our data by species and look at summary statistics. Consider the beanplot's below.



Though this marginal information may be useful in an analysis, when exploring multivariate data such as this we may want to consider some pair-wise relationships

We can see in Figure blank that a clear relationship between sepal width and sepal length. We also see that this relationship may differ by species of iris.

3 A Simple Regression

To explore the relationship we observed in figure blank, we will build a regression model. We can write out the relationship as:

$$SepalLength = \beta_0 + \beta_1 SepalWidth + \beta_2 SepalWidth \times$$

We get the following model summary output.

In Table blank, we see that

We can also conduct an Analysis of Variance (ANOVA) to test

We might wonder if we need to model the interaction term

We can see from the ANOVA output that there isn't much evidence that the slope of the relationship between Sepal Width and Sepal Length differs by species. We might then test the additive effects. We give type II sums of squares in the additive model to get two tests we are interested in.

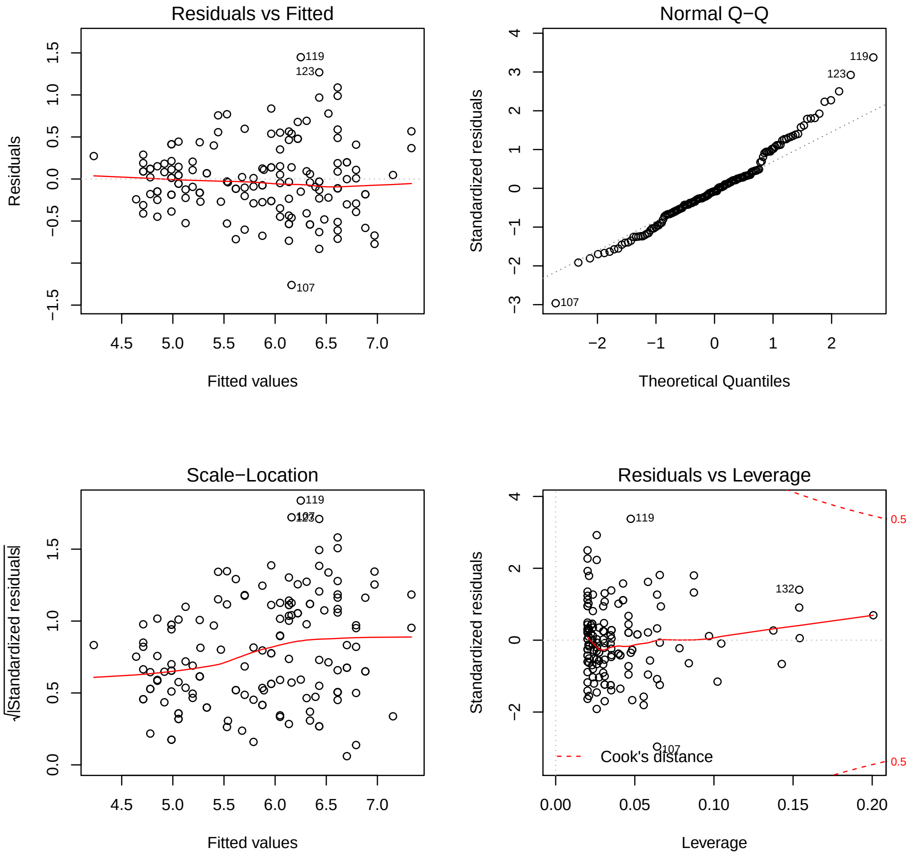
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.6390	0.5715	4.62	0.0000
Sepal.Width	0.6905	0.1657	4.17	0.0001
Speciesversicolor	0.9007	0.7988	1.13	0.2613
Speciesvirginica	1.2678	0.8162	1.55	0.1225
Sepal.Width:Speciesversicolor	0.1746	0.2599	0.67	0.5028
Sepal.Width:Speciesvirginica	0.2110	0.2558	0.83	0.4106

Table 3: A descriptive caption

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sepal.Width	1	1.41	1.41	7.30	0.0077
Species	2	72.75	36.38	188.11	0.0000
Sepal.Width:Species	2	0.16	0.08	0.41	0.6668
Residuals	144	27.85	0.19		

Table 4: A descriptive caption

We see that the additive model seems appropriate. The diagnostic plots for this model are given in table



	Sum Sq	Df	F value	Pr(>F)
Sepal.Width	10.95	1	57.10	0.0000
Species	72.75	2	189.65	0.0000
Residuals	28.00	146		

Table 5: A descriptive caption

4 Appendix

4.1 R-code

4.2 Document Creation Info

- **platform**:

  booktabs

  - **version**: R version 3.3.2 (2016-10-31)
  - **system**: x86_64, linux-gnu
  - **ui**: X11
  - **language**: en_US
  - **collate**: en_US.UTF-8
  - **tz**: Navajo
  - **date**: 2017-03-16

- **packages**:

package	*	version	date	source
assertthat		0.1	2013-12-06	CRAN (R 3.3.2)
backports		1.0.5	2017-01-18	CRAN (R 3.3.2)
base64enc		0.1-3	2015-07-28	CRAN (R 3.3.2)
beanplot	*	1.2	2014-09-19	CRAN (R 3.3.2)
car	*	2.1-4	2016-12-02	CRAN (R 3.3.2)
codetools		0.2-15	2016-10-05	CRAN (R 3.3.1)
colorspace		1.3-2	2016-12-14	CRAN (R 3.3.2)
DBI		0.5-1	2016-09-10	CRAN (R 3.3.2)
devtools	*	1.12.0	2016-12-05	CRAN (R 3.3.2)
digest		0.6.12	2017-01-27	CRAN (R 3.3.2)
dplyr		0.5.0	2016-06-24	CRAN (R 3.3.2)
evaluate		0.10	2016-10-11	CRAN (R 3.3.2)
foreign		0.8-67	2016-09-13	CRAN (R 3.3.1)
ggplot2	*	2.2.1	2016-12-30	CRAN (R 3.3.2)

package	*	version	date	source
gtable		0.2.0	2016-02-26	CRAN (R 3.3.2)
htmltools		0.3.5	2016-03-21	CRAN (R 3.3.2)
htmlwidgets		0.8	2016-11-09	CRAN (R 3.3.2)
httr		1.2.1	2016-07-03	CRAN (R 3.3.2)
jsonlite		1.3	2017-02-28	cran ((**???**))
knitr	*	1.15.1	2016-11-22	CRAN (R 3.3.2)
lattice		0.20-34	2016-09-06	CRAN (R 3.3.1)
lazyeval		0.2.0	2016-06-12	CRAN (R 3.3.2)
lme4		1.1-12	2016-04-16	CRAN (R 3.3.2)
magrittr		1.5	2014-11-22	CRAN (R 3.3.2)
MASS	*	7.3-45	2015-11-10	CRAN (R 3.2.5)
Matrix		1.2-7.1	2016-09-01	CRAN (R 3.3.1)
MatrixModels		0.4-1	2015-08-22	CRAN (R 3.3.2)
memoise		1.0.0	2016-01-29	CRAN (R 3.3.2)
mgcv		1.8-16	2016-11-07	CRAN (R 3.3.2)
minqa		1.2.4	2014-10-09	CRAN (R 3.3.2)
mnormt		1.5-5	2016-10-15	CRAN (R 3.3.2)
munsell		0.4.3	2016-02-13	CRAN (R 3.3.2)
nlme		3.1-129	2017-01-19	CRAN (R 3.3.2)
nloptr		1.0.4	2014-08-04	CRAN (R 3.3.2)
nnet		7.3-12	2016-02-02	CRAN (R 3.2.5)
pander	*	0.6.0	2015-11-23	CRAN (R 3.3.2)
pbkrtest		0.4-6	2016-01-27	CRAN (R 3.3.2)
plotly	*	4.5.6	2016-11-12	CRAN (R 3.3.2)
plyr		1.8.4	2016-06-08	CRAN (R 3.3.2)
printr	*	0.0.6	2017-01-30	Github (yihui/printr@42f100e)
psych	*	1.6.12	2017-01-08	CRAN (R 3.3.2)
purrr		0.2.2	2016-06-18	CRAN (R 3.3.2)
quantreg		5.29	2016-09-04	CRAN (R 3.3.2)
R6		2.2.0	2016-10-05	CRAN (R 3.3.2)
Rcpp		0.12.9	2017-01-14	CRAN (R 3.3.2)
rmarkdown		1.3	2016-12-21	CRAN (R 3.3.2)
rprojroot		1.2	2017-01-16	CRAN (R 3.3.2)
scales		0.4.1	2016-11-09	CRAN (R 3.3.2)
SparseM		1.74	2016-11-10	CRAN (R 3.3.2)
stargazer	*	5.2	2015-07-14	CRAN (R 3.3.2)
stringi		1.1.2	2016-10-01	CRAN (R 3.3.2)
stringr		1.1.0	2016-08-19	CRAN (R 3.3.2)
tibble		1.2	2016-08-26	CRAN (R 3.3.2)

package	*	version	date	source
tidyr		0.6.1	2017-01-10	CRAN (R 3.3.2)
viridisLite		0.1.3	2016-03-12	CRAN (R 3.3.2)
withr		1.0.2	2016-06-20	CRAN (R 3.3.2)
xtable	*	1.8-2	2016-02-05	CRAN (R 3.3.2)
yaml		2.1.14	2016-11-12	CRAN (R 3.3.2)

References

Anderson, Edgar. 1936. "The Species Problem in Iris." *Annals of the Missouri Botanical Garden* 23 (3). Missouri Botanical Garden Press: 457–509. http://www.jstor.org/stable/2394164.

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2). Blackwell Publishing Ltd: 179–88. doi:10.1111/j.1469-1809.1936.tb02137.x.