

# Predicting Iris Species

Jordan Schupbach

September 7, 2017

# 1 Introduction

The famous (Fisher 1936) (or (Anderson 1936)) iris data set gives the measurements (in centimeters) of the sepal length and width, and petal length and width for 50 flowers, each from one of three species of iris. The species are *I. setosa*, *I. versicolor*, and *I. virginica*. We provide images for each of the three species in Figure 1. Researchers are interested in whether they can identify the species from the set of the four measurements taken on the plants. For this study, the focus is only on predicting *I. Setosa* versus *I. Versicolor*.

## 1.1 Description of Data

This dataset has 1 categorical variable (species) and 4 quantitative variables (sepal length and width, and petal length and width; all measured in centimeters). Again, the researchers are interested in predicting species from the four measurements. We provide the first six lines of the dataset in Table 1 and a five-number summary of the data in Table 2. We also may want to subset our data by species and look at summary statistics. Means for the two species of interest are given in Table 3.

These data were gathered by Dr. Edgar Anderson. The *Iris setosa* and *Iris versicolor* were found growing together in the same colony. The third species, *Iris virginica*, was from a different colony, as it would be a “circumstance which might considerably disturb both the mean values and their variabilities.”

## 2 Statistical Methods Used

To classify the two species of interest, *I. setosa* and *I. versicolor*, we use the method of linear discriminant analysis following (Fisher 1936). One assumption of the method is that the covariates be independently normally distributed. We check the assumption of normality by plotting histograms of four variables grouped by species type in Figure 1. Further, we provide beanplots from the `beanplot` R package (Kampstra 2008), which assume a normal kernel density. The two figures would suggest the assumption of normality is reasonably met.

Linear discriminant analysis allows us to write an objective function of the form

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4 \tag{1}$$

for which we classify an observation as either *I. Setosa* or *I. Versicolor* depending on its value.

### 3 Summary of Statistical Findings

We provide means for the two groups in Table 3. The LDA was conducted in R (R Core Team 2017) using the `lda()` function from the `MASS` package (Venables and Ripley 2002). Coefficients of discrimination are given in Table 4. We see that we can write equation 1 as

$$X = -1.77x_1 - 0.3x_2 + 3.04x_3 + 2.14x_4 \quad (2)$$

Following (Fisher 1936), we can take the coefficient for sepal width to be unity (see Table 5) and rewrite equation 2 as

$$X = x_1 + 0.17x_2 - 1.71x_3 - 1.21x_4 \quad (3)$$

These equations can be used to score each observation on this new scale, and classify the observation as either *I. setosa* or *I. virginica*. Figure 5 shows the values of the scores for the two species according to equation 2. We can see from the plot that the scores do an impressive job in separating *I. setosa* and *I. virginica*. A confusion matrix is given in Table 6. We see that the equation can perfectly separate *I. setosa* plants from *I. virginica*.

### 4 Scope of Inference

Since we did not randomly assign any treatment to treatment groups, no causal inference can be inferred in this study. Further, since observations did not come from a random sample, we cannot infer to a larger population than that of the sample.

## 5 Appendix

### 5.1 Tables

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
4	4.60	3.10	1.50	0.20	setosa
5	5.00	3.60	1.40	0.20	setosa
6	5.40	3.90	1.70	0.40	setosa

Table 1: First 6 lines of the dataset

Table 2: Summary table of the four quantitative variables

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

	Sepal.Width	Sepal.Length	Petal.Width	Petal.Length
setosa	3.43	5.01	0.25	1.46
versicolor	2.77	5.94	1.33	4.26

Table 3: Group means

	LD1
Sepal.Width	-1.77
Sepal.Length	-0.30
Petal.Width	3.04
Petal.Length	2.14

Table 4: Coefficients of linear discriminants

	LD1
Sepal.Width	1.00
Sepal.Length	0.17
Petal.Width	-1.71
Petal.Length	-1.21

Table 5: Coefficients of linear discriminants with first scaled to unity

	setosa	versicolor
setosa	50	0
versicolor	0	50

Table 6: Confusion matrix

## 5.2 Figures



Figure 1: Three species of iris: I. Setosa, I. Virginica, I. Versicolor

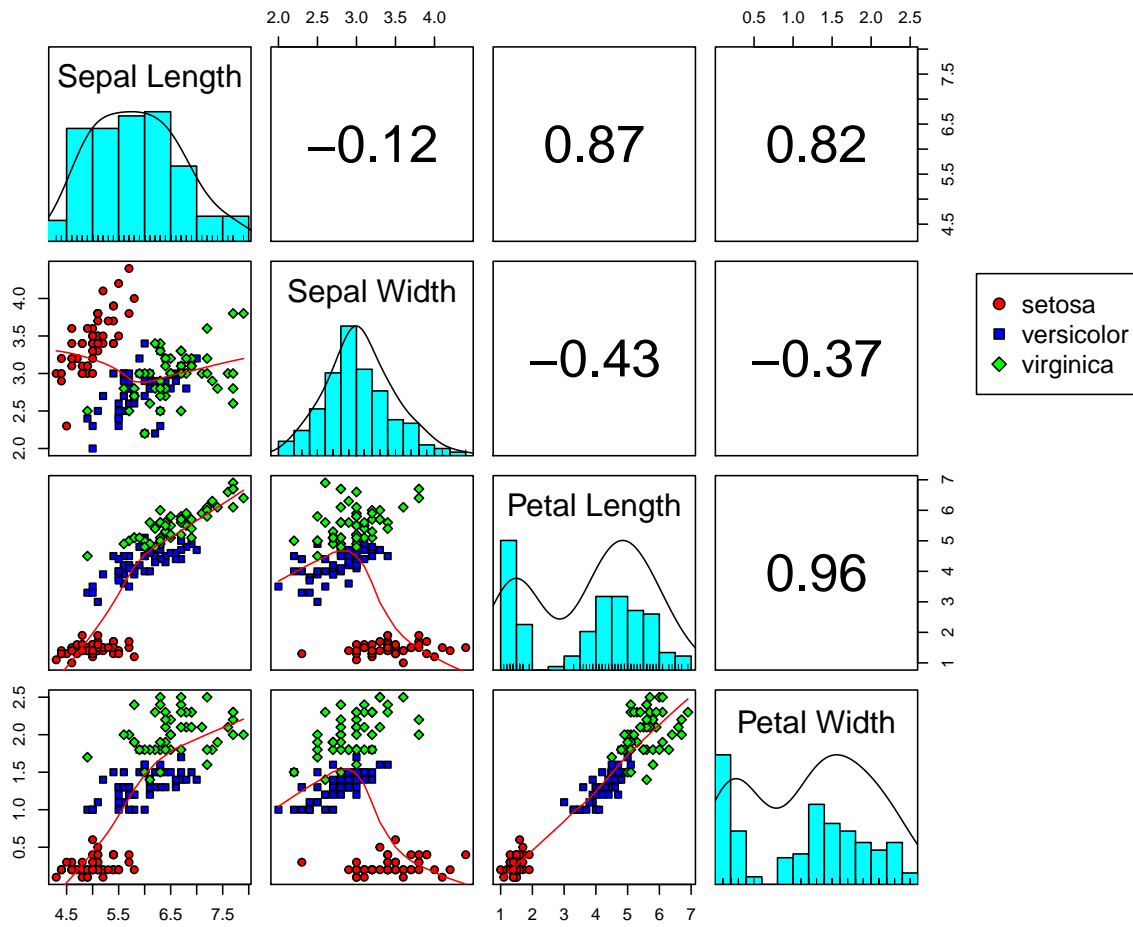


Figure 2: A matrix of scatterplots of the quantitative variables

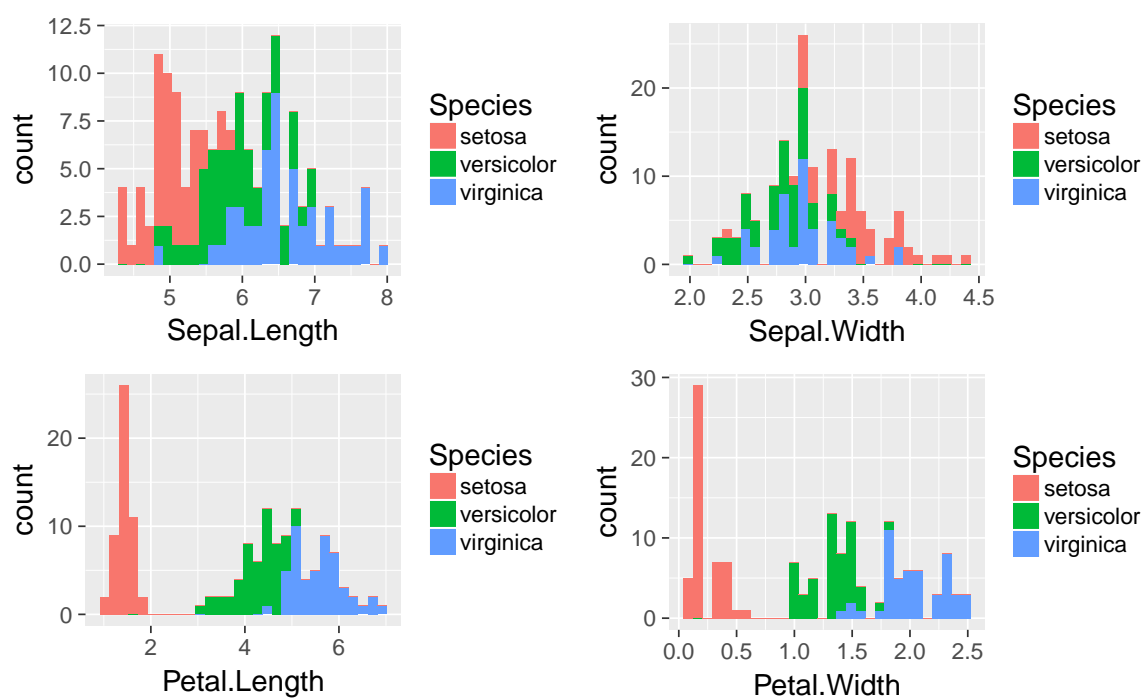


Figure 3: Histograms of covariates separated by species

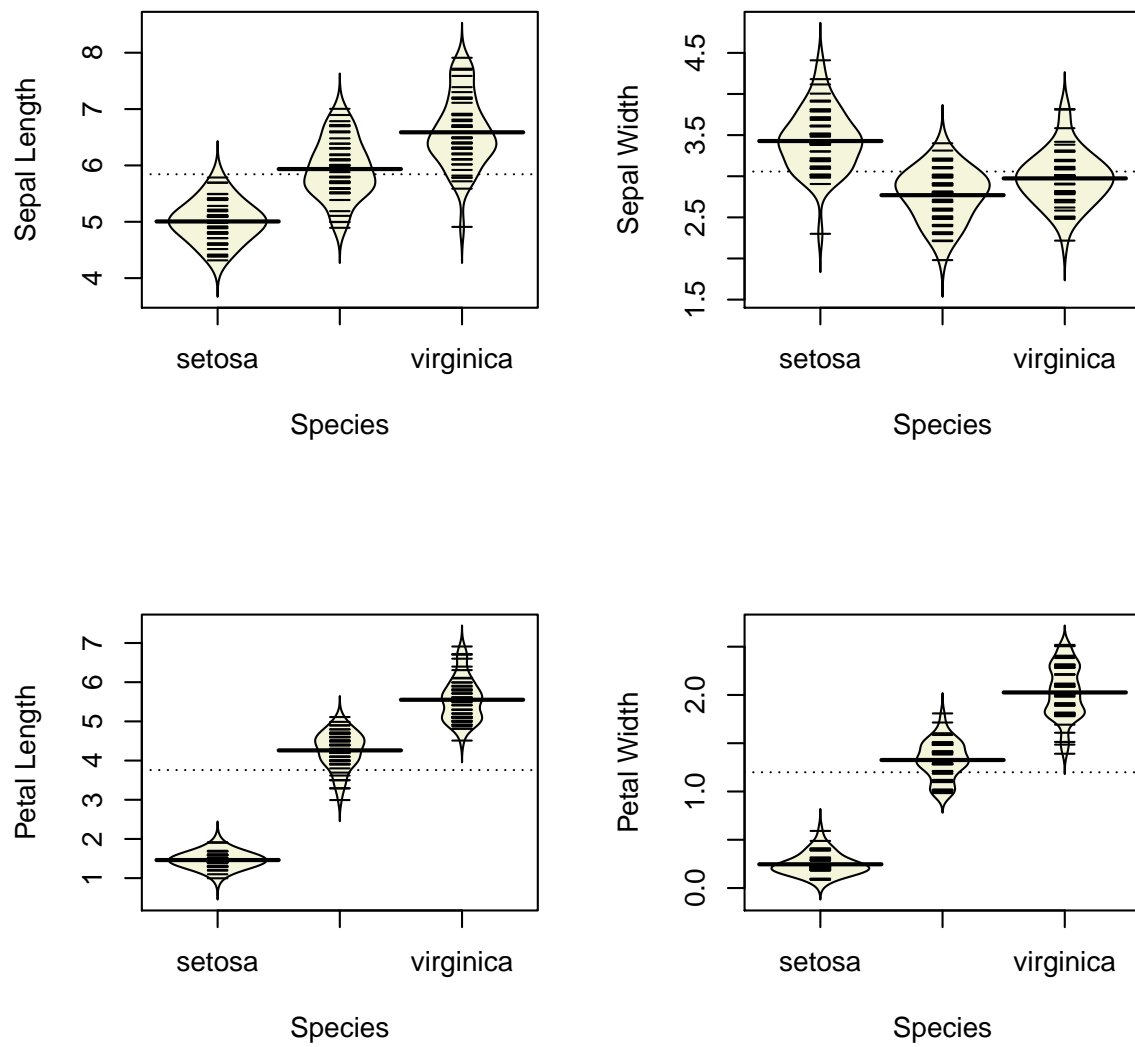


Figure 4: Beanplots of the four quantitative variables by species



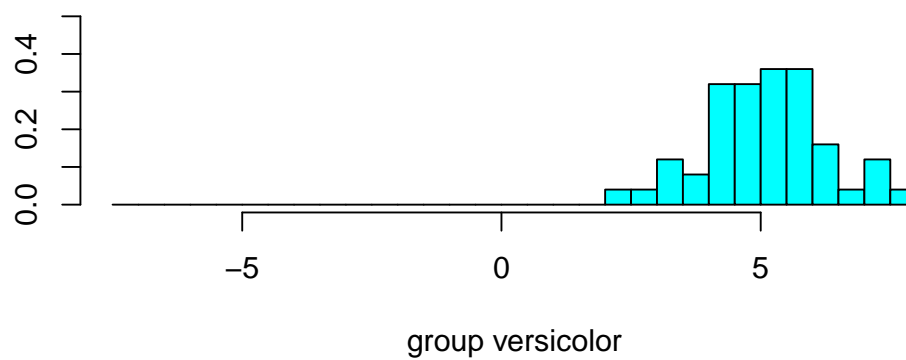
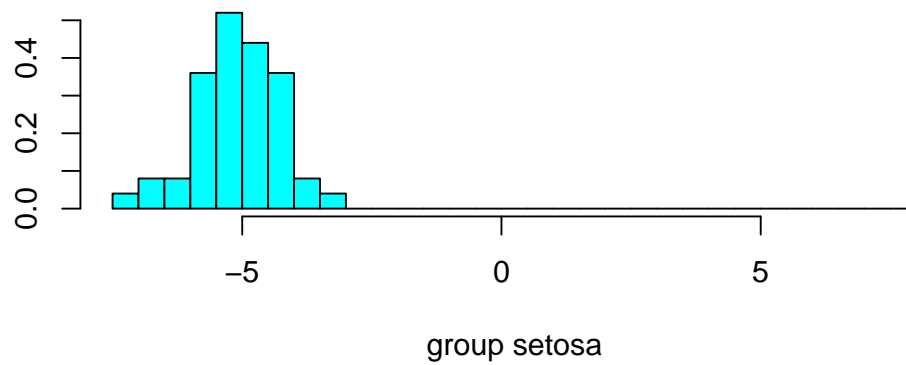


Figure 5: Linear discriminant scores for setosa and versicolor

## 5.3 R-code

```
# Subset the data
iris_sub <- subset(iris, iris$Species == "setosa" |
                  iris$Species == "versicolor")
iris_sub$Species <- factor(iris_sub$Species)
head(iris_sub)

# LDA
z <- lda(Species ~ Sepal.Width + Sepal.Length + Petal.Width +
        Petal.Length, iris_sub, prior = c(1,1)/2)
preds <- predict(z)$class

## Tables:
# Table 1
head(iris)
print(xtable(head(iris), caption = "First 6 lines of the dataset",
              label = "data-head"),
      comment = FALSE)

# Table 2
summary(iris)
stargazer(iris, iqr = TRUE,
          type = "latex", header = FALSE, label = "summary",
          title = "Summary table of the four quantitative variables")

# Table 3
z$mean
print(xtable(z$mean, label = "means-tbl", caption = "Group means"),
      comment = FALSE)

# Table 4
z$scaling
print(xtable(z$scaling, label = "scaling-tbl",
            caption = "Coefficients of linear discriminants"),
      comment = FALSE)

# Table 5
(1/z$scaling[1]) * z$scaling
print(xtable((1/z$scaling[1]) * z$scaling, label = "scaling-unity-tbl",
            caption = "Coefficients of linear discriminants with first scaled to unity"),
```

```

    comment = FALSE)

# Table 6: Confusion Matrix
table(preds, iris_sub$Species)
print(xtable(table(preds, iris_sub$Species), label = "confusion",
    caption = "Confusion matrix"),
    comment = FALSE)

## Figures
# Figure 1:
fig1 <- rasterGrob(readPNG("images/iris_setosa.png"), interpolate=TRUE)
fig2 <- rasterGrob(readPNG("images/iris_virginica.png"), interpolate=TRUE)
fig3 <- rasterGrob(readPNG("images/iris_versicolor.png"), interpolate=TRUE)
grid.arrange(fig1, fig2, fig3, ncol=3)

# Figure 2:
par(mar = c(5,4,4,3))
pairs.panels(iris[,1:4],
    labels = c("Sepal Length",
               "Sepal Width",
               "Petal Length",
               "Petal Width"),
    bg=c("red",
         "blue",
         "green")[iris$Species],
    pch=20+as.numeric(iris$Species),
    ellipses = FALSE,
    oma=c(4,4,6,12))
par(xpd=TRUE)
legend(.9*par("usr")[2], ## 1.05 times x1 limit
    .7*par("usr")[4], ## .8 times y2 limit
    levels(iris$Species),
    pch = c(21,22,23),
    pt.bg = c("red", "blue", "green"))

# Figure 3:
plot1 <- ggplot(iris, aes(x=Sepal.Length))
plot1<- plot1 + geom_histogram(aes(fill = Species))
plot2 <- ggplot(iris, aes(x=Sepal.Width))
plot2<- plot2 + geom_histogram(aes(fill = Species))
plot3 <- ggplot(iris, aes(x=Petal.Length))

```

```

plot3<- plot3 + geom_histogram(aes(fill = Species))
plot4 <- ggplot(iris, aes(x=Petal.Width))
plot4<- plot4 + geom_histogram(aes(fill = Species))
theme_set(theme_gray(base_size = 18))
m<-grid.arrange(plot1,plot2, plot3,plot4, ncol=2)
print(m)

# Figure 4:
par(mfrow = c(2,2))
beanplot(Sepal.Length ~ Species, data = iris, ylab = "Sepal Length",
         xlab = "Species", col = "beige", method = "jitter")
beanplot(Sepal.Width ~ Species, data = iris, ylab = "Sepal Width",
         xlab = "Species", col = "beige", method = "jitter")
beanplot(Petal.Length ~ Species, data = iris, ylab = "Petal Length",
         xlab = "Species", col = "beige", method = "jitter")
beanplot(Petal.Width ~ Species, data = iris, ylab = "Petal Width",
         xlab = "Species", col = "beige", method = "jitter")

# Figure 5:
plot(z)

```

## 5.4 Document Creation Info

- platform:
  - booktabs
  - version: R version 3.4.1 (2017-06-30)
  - system: x86\_64, linux-gnu
  - ui: X11
  - language: en\_US
  - collate: en\_US.UTF-8
  - tz: America/Denver
  - date: 2017-09-07

- packages:

package	*	version	date	source
backports		1.0.5	2017-01-18	CRAN (R 3.4.0)

package	*	version	date	source
base	*	3.4.1	2017-07-08	local
beanplot	*	1.2	2014-09-19	CRAN (R 3.4.0)
car	*	2.1-4	2016-12-02	CRAN (R 3.4.0)
colorspace		1.3-2	2016-12-14	CRAN (R 3.4.0)
compiler		3.4.1	2017-07-08	local
datasets	*	3.4.1	2017-07-08	local
devtools	*	1.13.1	2017-05-13	CRAN (R 3.4.0)
digest		0.6.12	2017-01-27	CRAN (R 3.4.0)
evaluate		0.10	2016-10-11	CRAN (R 3.4.0)
foreign		0.8-69	2017-06-21	CRAN (R 3.4.0)
ggplot2	*	2.2.1	2016-12-30	CRAN (R 3.4.0)
graphics	*	3.4.1	2017-07-08	local
grDevices	*	3.4.1	2017-07-08	local
grid	*	3.4.1	2017-07-08	local
gridExtra	*	2.2.1	2016-02-29	CRAN (R 3.4.0)
gtable		0.2.0	2016-02-26	CRAN (R 3.4.0)
htmltools		0.3.6	2017-04-28	CRAN (R 3.4.0)
knitr	*	1.16	2017-05-18	CRAN (R 3.4.0)
labeling		0.3	2014-08-23	CRAN (R 3.4.0)
lattice		0.20-35	2017-03-25	CRAN (R 3.3.3)

package	*	version	date	source
lazyeval		0.2.0	2016-06-12	CRAN (R 3.4.0)
lme4		1.1-13	2017-04-19	CRAN (R 3.4.0)
magrittr		1.5	2014-11-22	CRAN (R 3.4.0)
MASS	*	7.3-47	2017-04-21	CRAN (R 3.4.1)
Matrix		1.2-10	2017-04-28	CRAN (R 3.4.0)
MatrixModels		0.4-1	2015-08-22	CRAN (R 3.4.0)
memoise		1.1.0	2017-04-21	CRAN (R 3.4.0)
methods	*	3.4.1	2017-07-08	local
mgcv		1.8-18	2017-07-28	CRAN (R 3.4.1)
minqa		1.2.4	2014-10-09	CRAN (R 3.4.0)
mnormt		1.5-5	2016-10-15	CRAN (R 3.4.0)
munsell		0.4.3	2016-02-13	CRAN (R 3.4.0)
nlme		3.1-131	2017-02-06	CRAN (R 3.4.0)
nloptr		1.0.4	2014-08-04	CRAN (R 3.4.0)
nnet		7.3-12	2016-02-02	CRAN (R 3.4.0)
nvimcom	*	0.9-34	2017-08-28	local
pander	*	0.6.0	2015-11-23	CRAN (R 3.4.0)
parallel		3.4.1	2017-07-08	local
pbkrtest		0.4-7	2017-03-15	CRAN (R 3.4.0)
plyr		1.8.4	2016-06-08	CRAN (R 3.4.0)

package	*	version	date	source
png	*	0.1-7	2013-12-03	CRAN (R 3.4.0)
printr	*	0.1	2017-05-19	Github (yihui/printr@7bc168b) CRAN (R 3.4.0)
psych	*	1.7.5	2017-05-03	
quantreg		5.33	2017-04-18	CRAN (R 3.4.0)
Rcpp		0.12.10	2017-03-19	CRAN (R 3.4.0)
rlang		0.1.1	2017-05-18	CRAN (R 3.4.0)
rmarkdown	*	1.5	2017-04-26	CRAN (R 3.4.0)
rprojroot		1.2	2017-01-16	CRAN (R 3.4.0)
scales		0.4.1	2016-11-09	CRAN (R 3.4.0)
SparseM		1.77	2017-04-23	CRAN (R 3.4.0)
splines		3.4.1	2017-07-08	local
stargazer	*	5.2	2015-07-14	CRAN (R 3.4.0)
stats	*	3.4.1	2017-07-08	local
stringi		1.1.5	2017-04-07	CRAN (R 3.4.0)
stringr		1.2.0	2017-02-18	CRAN (R 3.4.0)
tibble		1.3.1	2017-05-17	CRAN (R 3.4.0)
tools		3.4.1	2017-07-08	local
utils	*	3.4.1	2017-07-08	local
withr		1.0.2	2016-06-20	CRAN (R 3.4.0)
xtable	*	1.8-2	2016-02-05	CRAN (R 3.4.0)

package	*	version	date	source
yaml		2.1.14	2016-11-12	CRAN (R 3.4.0)

## 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.

Kampstra, Peter. 2008. “Beanplot: A Boxplot Alternative for Visual Comparison of Distributions.” *Journal of Statistical Software, Code Snippets* 28 (1): 1–9. <http://www.jstatsoft.org/v28/c01/>.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. <http://www.stats.ox.ac.uk/pub/MASS4>.