Iris Data LDA

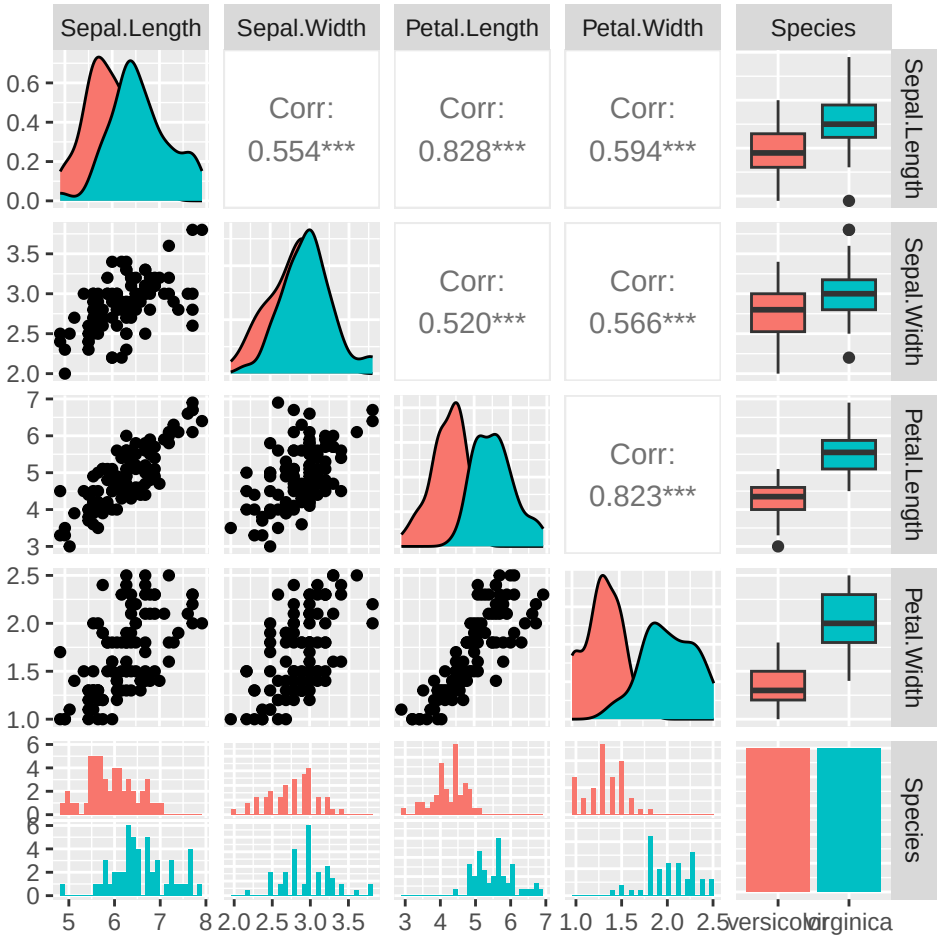
Carson Slater

2024-07-12

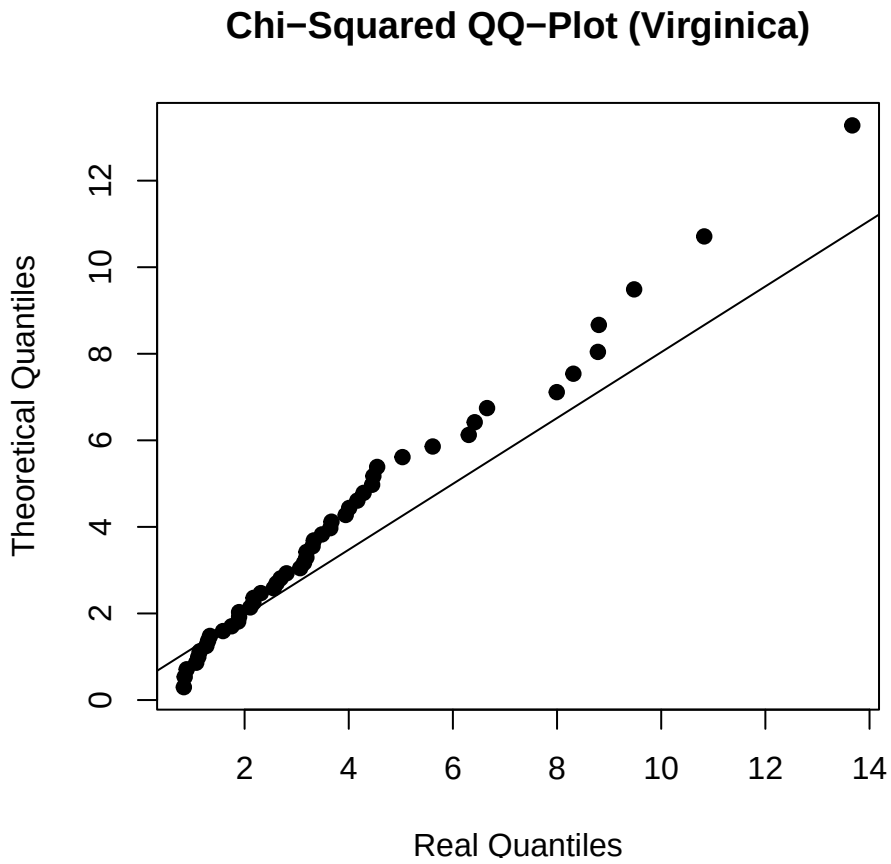
Colophon

Loading the Data, Checking Assumptions

```
data("iris")
iris <- iris |>
  dplyr::filter(Species != c("setosa")) |>
  mutate(Species = droplevels(Species))

GGally::ggpairs(iris, aes(fill = Species))
```

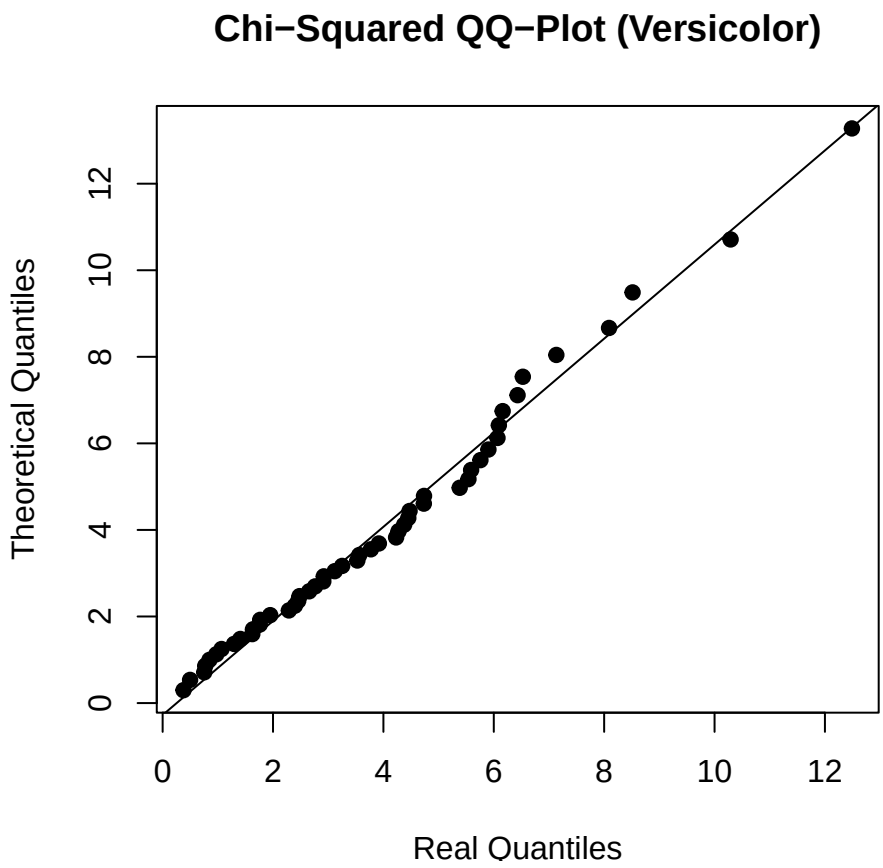


```r
iris_virginica <- iris |> dplyr::filter(Species == "virginica")
iris_versicolor <- iris |> dplyr::filter(Species == "versicolor")

xbar1 <- colMeans(iris_virginica[,-5])
sigma1 <- cov(iris_virginica[,-5])

xbar2 <- colMeans(iris_versicolor[,-5])
sigma2 <- cov(iris_versicolor[,-5])

dist1 <- mahalanobis(as.matrix(iris_virginica[,-5]), center = xbar1, cov = sigma1)


qqplot(dist1, qchisq(ppoints(nrow(iris_virginica)), ncol(iris_virginica[,-5])),
       main = "Chi-Squared QQ-Plot (Virginica)",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
       pch = 19)
qqline(dist1, distribution = \(prob) qchisq(prob, df = ncol(iris_virginica[,-5])))
```




```r
dist2 <- mahalanobis(as.matrix(iris_versicolor[,-5]), center = xbar2, cov = sigma2)

qqplot(dist2, qchisq(ppoints(nrow(iris_versicolor)), ncol(iris_versicolor[,-5])),
       main = "Chi-Squared QQ-Plot (Versicolor)",
       ylab = "Theoretical Quantiles",
       xlab = "Real Quantiles",
       pch = 19)
```

```r
qqline(dist2, distribution = \(prob) qchisq(prob, df = ncol(iris_versicolor[,-5])))
```



We would have reason to believe the `iris` for `versicolor` data has a multivariate normal distribution, as the QQ-plot of the mahalanobis distances resembles a χ^2_4 distribution. We would want to suspect that the `virginica` observations do not come from a multivariate normal distribution.

Creating Cross Validation Folds

```r
set.seed(613)

n_folds <- 10

# create 10 folds
folds <- createFolds(iris$Species, k = n_folds, list = TRUE, returnTrain = FALSE)
```

Performing the Linear Discriminant Analysis

```r
apparent_error_rates <- numeric(n_folds)
preds <- vector("list", length = n_folds)
posteriors <- vector("list", length = n_folds)

for(i in 1:10) {

  test_indices <- folds[[i]]
```

```r
  train_indices <- setdiff(1:nrow(iris), test_indices)

  train_data <- iris[train_indices, ]
  test_data <- iris[test_indices, ]

  trained_lda <- MASS::lda(Species ~ ., train_data)
  predictions <- predict(trained_lda, test_data)

  preds[[i]] <- predictions$class
  posteriors[[i]] <- predictions$posterior

  apparent_error_rate <- mean(preds[[i]] != test_data$Species)
  apparent_error_rates[i] <- apparent_error_rate

}
# apparent error rate
mean(apparent_error_rates)
```

```
## [1] 0.03
```

The apparent error rate is 0.03.