# Cushing's Syndrome: LDA, QDA, and Logistic Regression

Melissa Swann

March 2, 2020

## Background

The data set Cushings (in library MASS) has data on three types of the disease Cushing's syndrome (http://endocrine.niddk.nih.gov/pubs/cushings/cushings.htm), and the levels of two steroid metabolites used to diagnose the disease.

If left untreated, the pituitary keeps producing massive amounts of cortisol (a natural steroid) and you get symptoms of obesity, high blood pressure, acne and the various side effects of steroids. In the past, patients usually died of a heart attack. For centuries, even if diagnosed, it was impossible to treat. Now using micro surgery, there is a procedure to remove the small tumor without even any scarring (though the nose and sinuses and into the pituitary!).

In the library *nnet* there is a function multinom that does logistic regression on more than 2 response categories. The syntax is similar to glm. After loading library(nnet), look at help(multinom).

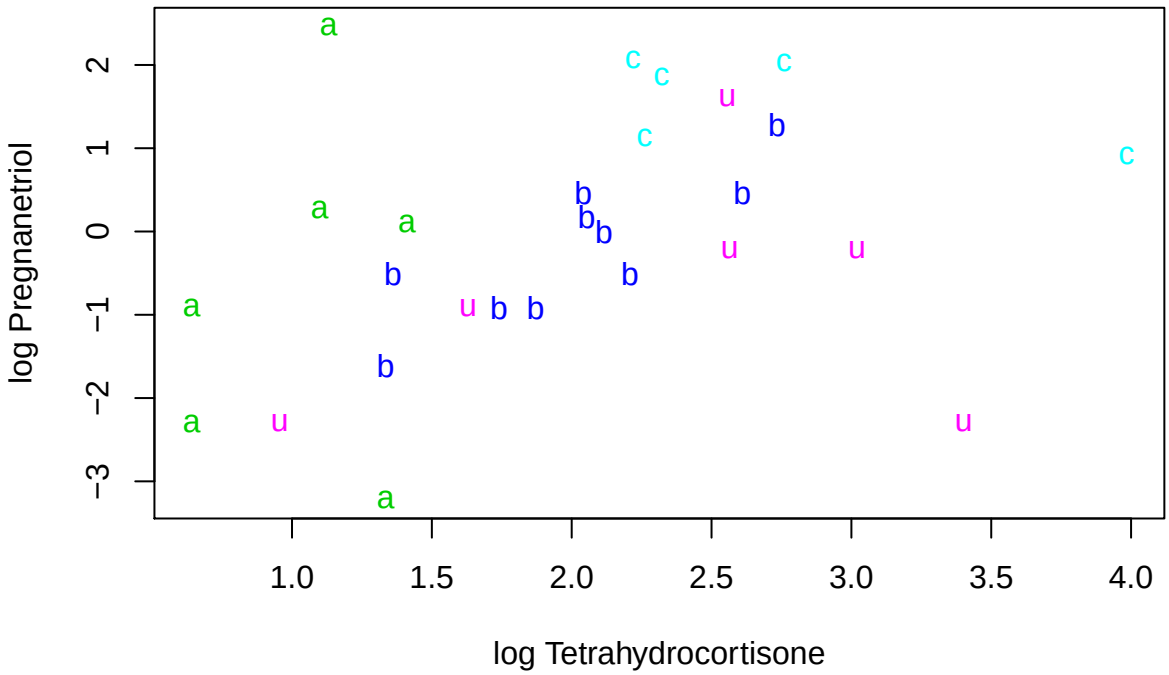
I want to compare LDA, QDA (letting the assumption about equal variances go) and multinomial logistic to the prediction of Cushings. The two predictor variables are highly right skewed, so I'll take the logs of the steroid levels.

```
Cf <- Cushings # dataframe with the logs of the steroid levels
names(Cf) <- c("logTetrahydrocortisone", "logPregnanetriol", "Type")
Cf$logTetrahydrocortisone <- log(Cushings$Tetrahydrocortisone)
Cf$logPregnanetriol <- log(Cushings$Pregnanetriol)
```

## Data Exploration

Let's take a look at the placement of the four groups ("u" is unknown):

```
plot(Cf[,1], Cf[,2], type="n",xlab = "log Tetrahydrocortisone", ylab ="log Pregnanetriol")
text(Cf[,1], Cf[,2],labels=Cf$Type,col=as.numeric(Cf$Type)+2)
```




These data do not appear to be naturally separated into clear, disparate groups in terms of steroid metabolite levels. We so see that Type A patients seem distinct from Type C patients. Type A patients tend to have low log levels of Tetrahydrocortisone (below 1.5) and variable log levels of Pregnanetriol. In contrast Type C patients appear to have moderately high levels of log Tetrahydrocortisone (typically between 2.0 and 3.0 with one observation at approximately 4.0) and relatively high (positive) log levels of Pregnanetriol. Type B patients look to have variable levels of both, though Type B patients seem unlikely to have extreme levels of either steroid metabolite. Type B does appear to overlap with both Type A and Type C. The unknown (Type U) patients do not exhibit any clear trend for either metabolite, and most of them could be reasonably classified as at least two of the other types. There is one patient of unknown type with a high log Tetrahydrocortisone level of approximately 3.4 and a log Pregnanetriol level of approximately -2.2, which is not consistent with any other type.

Next, I will drop the "u"s from the data set to look just at the patients with known types of Cushing's syndrome.

```
Cf_u_included <- Cf
Cf <- subset(Cf, Type != "u")
Cf$Type=factor(Cf$Type)
```

## Fitting Models

I want to fit and compare three models: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and Multinomial on this new data set Cf. I'll call these models Z.l, Z.q, and Z.m, respectively.

```
# lda
Z.l <- lda(Type ~ logTetrahydrocortisone + logPregnanetriol, data = Cf)
Z.l
```

```
## Call:
## lda(Type ~ logTetrahydrocortisone + logPregnanetriol, data = Cf)
##
## Prior probabilities of groups:
##         a         b         c
```

```
## 0.2857143 0.4761905 0.2380952
##
## Group means:
##   logTetrahydrocortisone logPregnanetriol
## a               1.043285       -0.6034147
## b               2.007256       -0.2060408
## c               2.709728        1.5998004
##
## Coefficients of linear discriminants:
##                              LD1        LD2
## logTetrahydrocortisone 1.7511836 -0.9907670
## logPregnanetriol       0.2341177  0.7874075
##
## Proportion of trace:
##    LD1    LD2
## 0.9299 0.0701
```

```
# qda
Z.q <- qda(Type ~ logTetrahydrocortisone + logPregnanetriol, data = Cf)
Z.q
```

```
## Call:
## qda(Type ~ logTetrahydrocortisone + logPregnanetriol, data = Cf)
##
## Prior probabilities of groups:
##         a         b         c
## 0.2857143 0.4761905 0.2380952
##
## Group means:
##   logTetrahydrocortisone logPregnanetriol
## a               1.043285       -0.6034147
## b               2.007256       -0.2060408
## c               2.709728        1.5998004
```

```
# multinomial
Z.m <- multinom(Type ~ logTetrahydrocortisone + logPregnanetriol, data = Cf)
```

```
## # weights:  12 (6 variable)
## initial  value 23.070858
## iter  10 value 6.623970
## iter  20 value 6.214841
## iter  30 value 6.182968
## iter  40 value 6.172650
## iter  50 value 6.167699
## iter  60 value 6.162723
## iter  70 value 6.156685
## iter  80 value 6.155298
## iter  90 value 6.153807
## iter 100 value 6.152597
## final  value 6.152597
## stopped after 100 iterations
```

```
Z.m
```

```
## Call:
## multinom(formula = Type ~ logTetrahydrocortisone + logPregnanetriol,
##     data = Cf)
```

```
##
## Coefficients:
##   (Intercept) logTetrahydrocortisone logPregnanetriol
## b   -19.09536               13.70353       -0.2491835
## c   -27.95158               15.58629        3.3210951
##
## Residual Deviance: 12.30519
## AIC: 24.30519
```

## Visualization Prep

In order to plot the posterior probabilities, I'll create a grid 100 x 100 on the two X variables:

```
xp <- seq(0.6, 4.0, length = 100)
yp <- seq(-3.25, 2.45, length = 100)
cushT <- expand.grid(logTetrahydrocortisone = xp, logPregnanetriol = yp)
```

Now I can predict each of the models on the grid. I'll call the results Z.lpred Z.qpred and Z.mpred.

For each row, and each model, I will now have three probabilities – for types "a", "b" and "c". The lines separating the predictions can be found by finding where, for example, the probability for the third class minus the max of the other two is positive.

```
# LDA predictions
lda.predict <- predict(Z.l, newdata=cushT)
Z.lpred <- lda.predict$posterior
# the probabilities are found in the "posterior" component

Z.p1l = Z.lpred[,1]-pmax(Z.lpred [,2],Z.lpred [,3])
Z.p1l = matrix(Z.p1l, nrow=100, ncol=100) # Make the vector a 100x100 matrix

Z.p3l = Z.lpred[,3]-pmax(Z.lpred [,2],Z.lpred [,1])
Z.p3l = matrix(Z.p3l, nrow=100, ncol=100) # Make the vector a 100x100 matrix
```

```
# QDA predictions
qda.predict <- predict(Z.q, newdata=cushT)
Z.qpred <- qda.predict$posterior
# the probabilities are found in the "posterior" component

Z.p1q = Z.qpred[,1]-pmax(Z.qpred [,2],Z.qpred [,3])
Z.p1q = matrix(Z.p1q, nrow=100, ncol=100) # Make the vector a 100x100 matrix

Z.p3q = Z.qpred[,3]-pmax(Z.qpred [,2],Z.qpred [,1])
Z.p3q = matrix(Z.p3q, nrow=100, ncol=100) # Make the vector a 100x100 matrix
```

```
# Multinomial predictions
Z.mpred <- predict(Z.m, type="probs", newdata=cushT)

Z.p1m = Z.mpred[,1]-pmax(Z.mpred [,2],Z.mpred [,3])
Z.p1m = matrix(Z.p1m, nrow=100, ncol=100) # Make the vector a 100x100 matrix

Z.p3m = Z.mpred[,3]-pmax(Z.mpred [,2],Z.mpred [,1])
Z.p3m = matrix(Z.p3m, nrow=100, ncol=100) # Make the vector a 100x100 matrix
```

## Visualization

For each model, I can plot these two lines by using the contour function. To do that, I'll need to turn each of these 6 vectors into 100 by 100 matrices. I'm only looking to see where each function switches from negative to positive (where it equals 0).

But first, I'll calculate the mean steroid levels for all three types, so I can include those points in the plots.

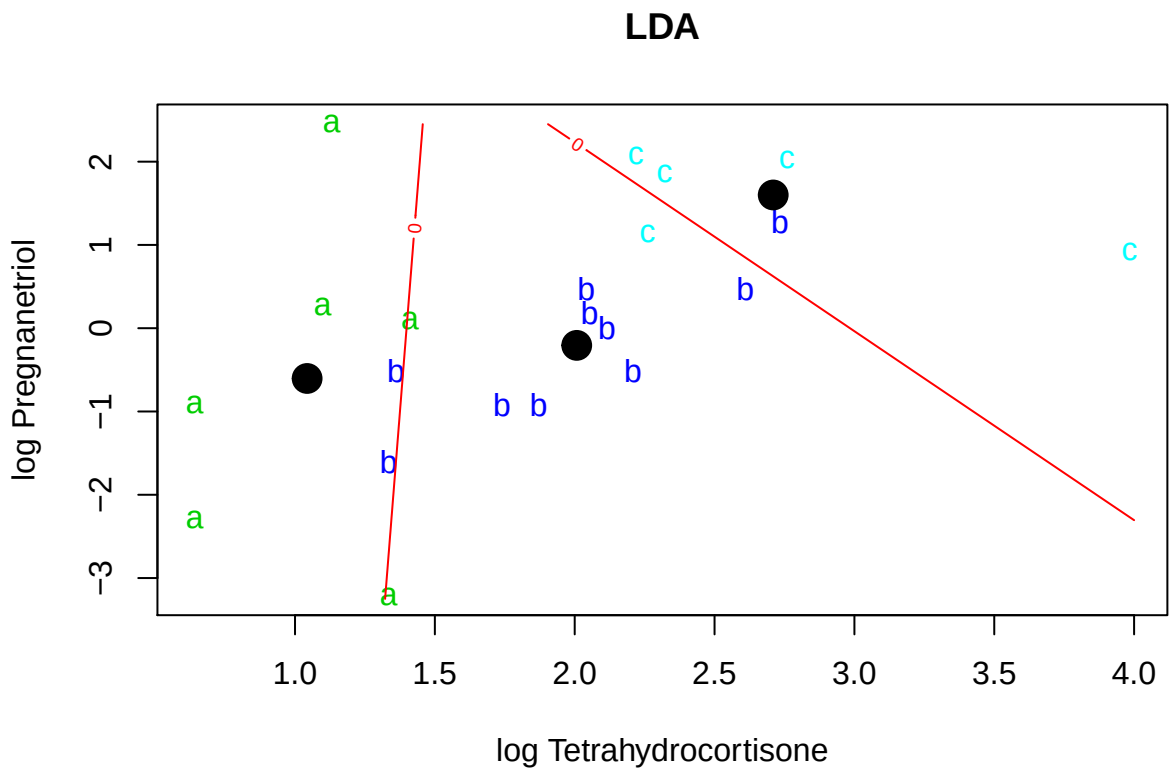
```
# Mean steroid levels for Type A, Type B, and Type C.
TypeA <- subset(Cf, Type=="a")
meanA <- c(mean(TypeA$logTetrahydrocortisone), mean(TypeA$logPregnanetriol))

TypeB <- subset(Cf, Type=="b")
meanB <- c(mean(TypeB$logTetrahydrocortisone), mean(TypeB$logPregnanetriol))

TypeC <- subset(Cf, Type=="c")
meanC <- c(mean(TypeC$logTetrahydrocortisone), mean(TypeC$logPregnanetriol))
```

Now, here are the plots of the boundaries based on our LDA, QDA, and multinomial logistic models.

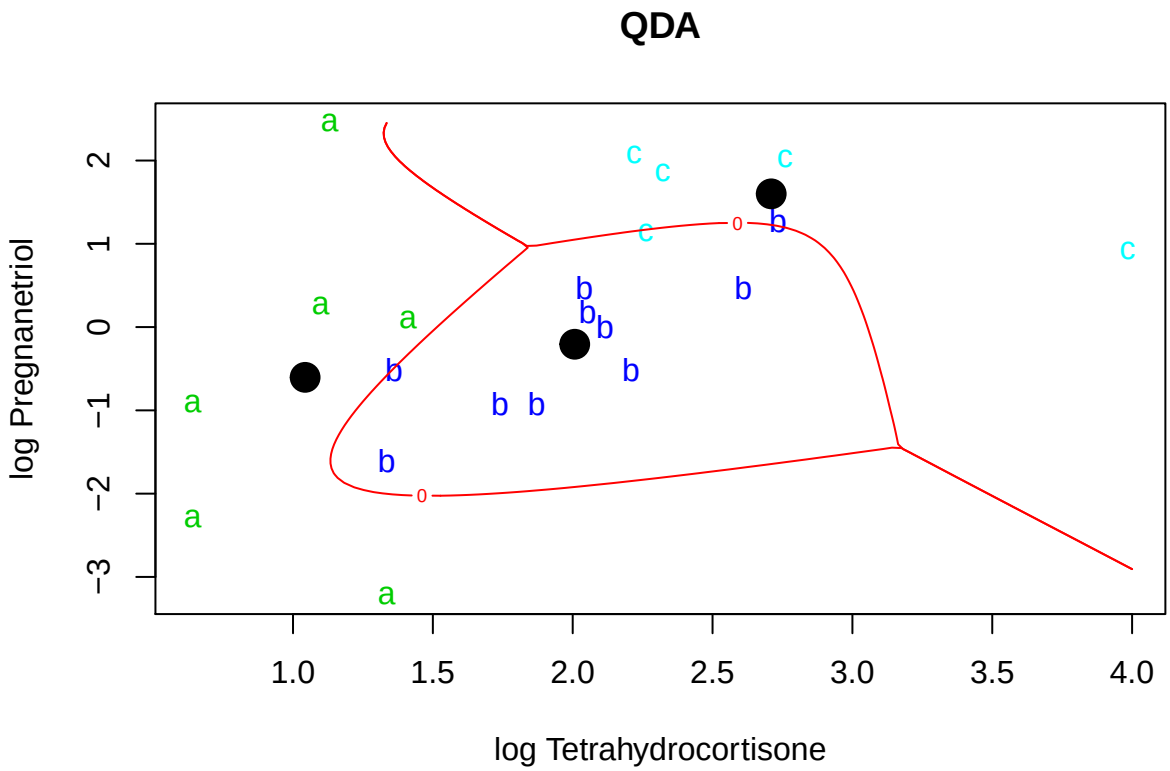
```
# LDA
plot(Cf[,1], Cf[,2], type="n",xlab = "log Tetrahydrocortisone", ylab ="log Pregnanetriol",
     main = "LDA")
text(Cf[,1], Cf[,2],labels=Cf$Type,col=as.numeric(Cf$Type)+2)
contour(xp,yp,Z.p1l,levels=0,add=T,col="red")
contour(xp,yp,Z.p3l,levels=0,add=T,col="red")
points(meanA[1], meanA[2], pch=19, cex=2)
points(meanB[1], meanB[2], pch=19, cex=2)
points(meanC[1], meanC[2], pch=19, cex=2)
```




The LDA model mischaracterizes six points out of the 21 points in the dataset.

Now to QDA:

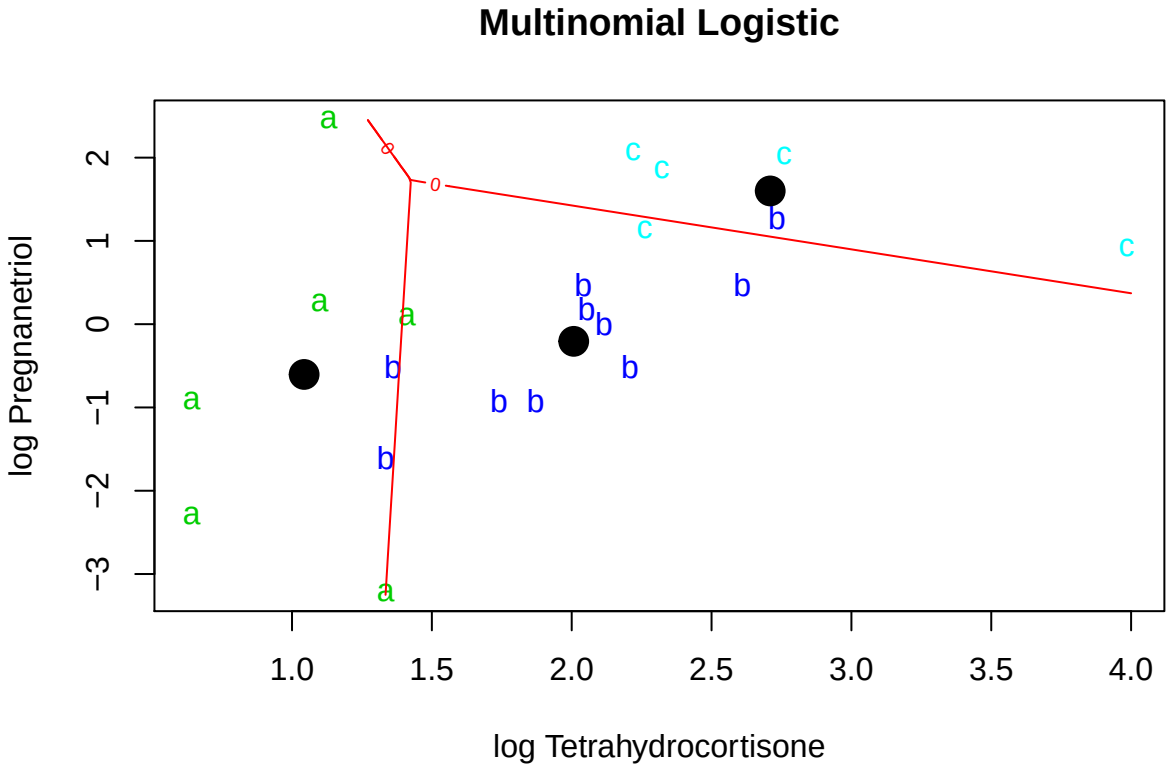
```
# QDA
plot(Cf[,1], Cf[,2], type="n",xlab = "log Tetrahydrocortisone", ylab ="log Pregnanetriol",
     main="QDA")
text(Cf[,1], Cf[,2],labels=Cf$Type,col=as.numeric(Cf$Type)+2)
contour(xp,yp,Z.p1q,levels=0,add=T,col="red")
contour(xp,yp,Z.p3q,levels=0,add=T,col="red")
points(meanA[1], meanA[2], pch=19, cex=2)
points(meanB[1], meanB[2], pch=19, cex=2)
points(meanC[1], meanC[2], pch=19, cex=2)
```




The QDA model only mischaracterizes two points from the dataset, which is a better error rate than the LDA model.

Finally, let's look at multinomial logistic:

```
# Multinomial logistic
plot(Cf[,1], Cf[,2], type="n",xlab = "log Tetrahydrocortisone", ylab ="log Pregnanetriol",
     main = "Multinomial Logistic")
text(Cf[,1], Cf[,2],labels=Cf$Type, col=as.numeric(Cf$Type)+2)
contour(xp,yp,Z.p1m,levels=0,add=T,col="red")
contour(xp,yp,Z.p3m,levels=0,add=T,col="red")
points(meanA[1], meanA[2], pch=19, cex=2) # mean of Type A
points(meanB[1], meanB[2], pch=19, cex=2) # mean of Type B
points(meanC[1], meanC[2], pch=19, cex=2) # mean of Type C
```

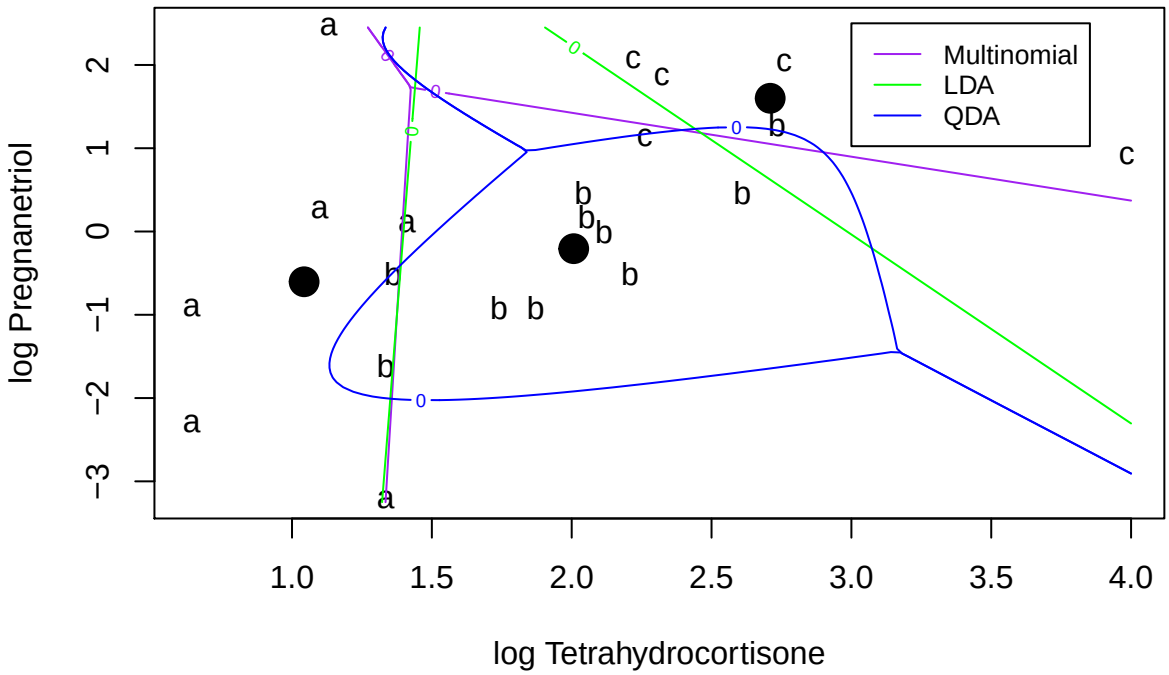



The multinomial logistic model misdiagnoses the same six points as the LDA model.

Mischaracterizing only two points from the dataset, the QDA model has a better error rate than both LDA and QDA. The multinomial and LDA models both mischaracterize the two points that QDA gets wrong.

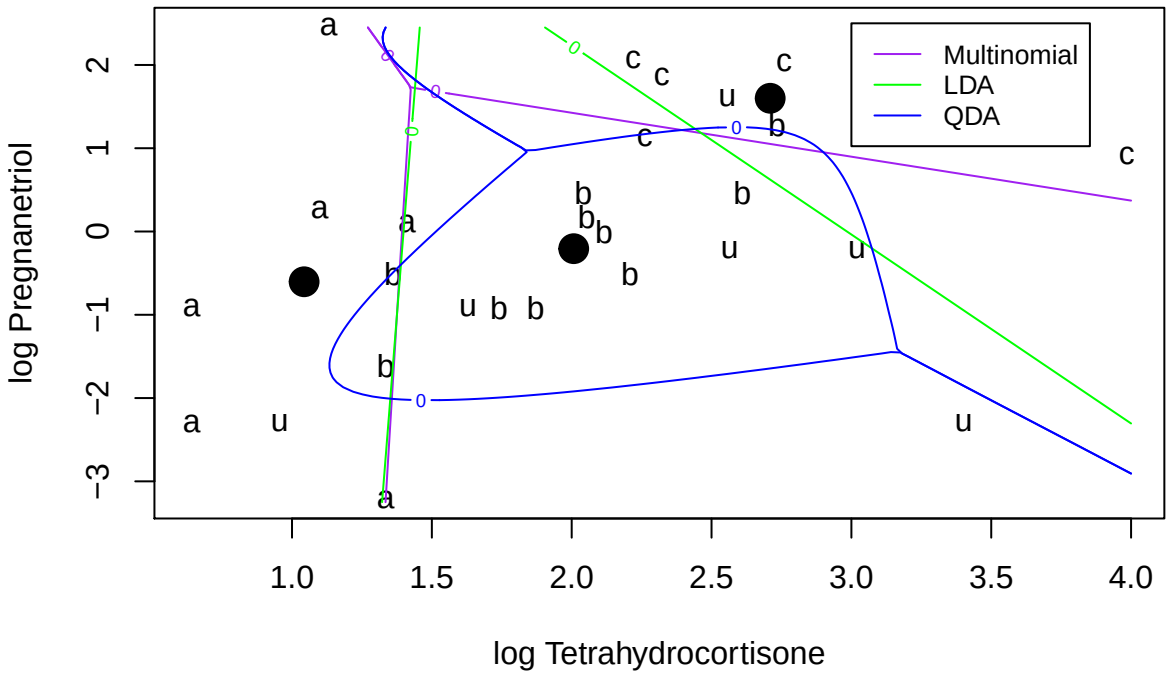
Now I can combine all three models into one plot to compare the boundaries between Type A, Type B, and Type C.

```
# All three combined
plot(Cf[,1], Cf[,2], type="n",xlab = "log Tetrahydrocortisone", ylab ="log Pregnanetriol")
text(Cf[,1], Cf[,2],labels=Cf$Type)
contour(xp,yp,Z.p1m,levels=0,add=T,col="purple") # multinomial
contour(xp,yp,Z.p3m,levels=0,add=T,col="purple") #multinomial
contour(xp,yp,Z.p1l,levels=0,add=T,col="green") #lda
contour(xp,yp,Z.p3l,levels=0,add=T,col="green") #lda
contour(xp,yp,Z.p1q,levels=0,add=T,col="blue") #qda
contour(xp,yp,Z.p3q,levels=0,add=T,col="blue") #qda
legend(3,2.5, legend=c("Multinomial", "LDA", "QDA"),
       col=c("purple", "green", "blue"), lty=1, cex=0.8)
points(meanA[1], meanA[2], pch=19, cex=2)
points(meanB[1], meanB[2], pch=19, cex=2)
points(meanC[1], meanC[2], pch=19, cex=2)
```




Lastly, I'll add the unknown points back into the plot to see how the three different models would characterize them.

```
# All three combined, with the unknown values included
plot(Cf_u_included[,1], Cf_u_included[,2], type="n",xlab = "log Tetrahydrocortisone",
     ylab ="log Pregnanetriol")
text(Cf_u_included[,1], Cf_u_included[,2],labels=Cf_u_included$Type)
contour(xp,yp,Z.p1m,levels=0,add=T,col="purple") # multinomial
contour(xp,yp,Z.p3m,levels=0,add=T,col="purple") #multinomial
contour(xp,yp,Z.p1l,levels=0,add=T,col="green") #lda
contour(xp,yp,Z.p3l,levels=0,add=T,col="green") #lda
contour(xp,yp,Z.p1q,levels=0,add=T,col="blue") #qda
contour(xp,yp,Z.p3q,levels=0,add=T,col="blue") #qda
legend(3,2.5, legend=c("Multinomial", "LDA", "QDA"),
       col=c("purple", "green", "blue"), lty=1, cex=0.8)
points(meanA[1], meanA[2], pch=19, cex=2)
points(meanB[1], meanB[2], pch=19, cex=2)
points(meanC[1], meanC[2], pch=19, cex=2)
```




## Conclusion

The QDA model seems most apt for the data we observed, though the small size of our dataset inhibits our predictive power and our ability to generalize our model preferences to other Cushing's data. The QDA model has the lowest approximate error rate of 0.095 and only misclassifies two patients. It better accounts for the greater spread in log Tetrahydrocortisone levels (as compared to log Pregnanetriol levels) for Type B and restricts the Type B region to patients without extreme levels of either metabolite. All three models seem to accommodate the trend that Type B patients have log levels of both metabolites that are between those of patients classified as either Type A OR Type C. The LDA and Multinomial models have somewhat similar boundaries, but the LDA model allows for some patients who have very high log Pregnanetriol levels to be classified as Type B, whereas the Multinomial predicts this upper region corresponds to Type C patients. The Multinomial model thus grants more representation to Type C than does the LDA model in terms of log Tetrahydrocortisone levels, but less representation in terms of log Pregnanetriol levels.

Both the LDA and the Multinomial model have mediocre approximate error rates 0.286, and notably they both misclassify the same six points in the same ways (including the two points the QDA misclassifies). The QDA model's boundaries are closest to properly classifying these two points, and the LDA model is the least close to doing so. But the QDA model predicts Type A patients will have larger log levels of Tetrahydrocortisone than the data suggest for these patients. In contrast, the other models predict Type A patients will have log Tetrahydorcortisone levels of no larger than 1.5, which seems more appropriate given the data we have.

I re-incorporated the patients with unknown Type to see what the models would predict for their Types. All three models classify the patient with very low levels of both metabolites as Type A, just as all three models classify the patient with high log Pregnanetriol and moderately high log Tetrahydrocortisone levels as Type C. There are three patients with moderate levels of both hormones which all three models classify as Type B, and finally the point to which we drew attention in my initial graph of the dataset is classified as Type B by the Multinomial and LDA models and alternatively as Type A by the QDA. This disagreement matches my observation that this particular point does not seem to fit well into the other classification groups.