NCSU ST 503 Discussion 11

Probem 2.1 Faraway, Julian J. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models CRC Press.

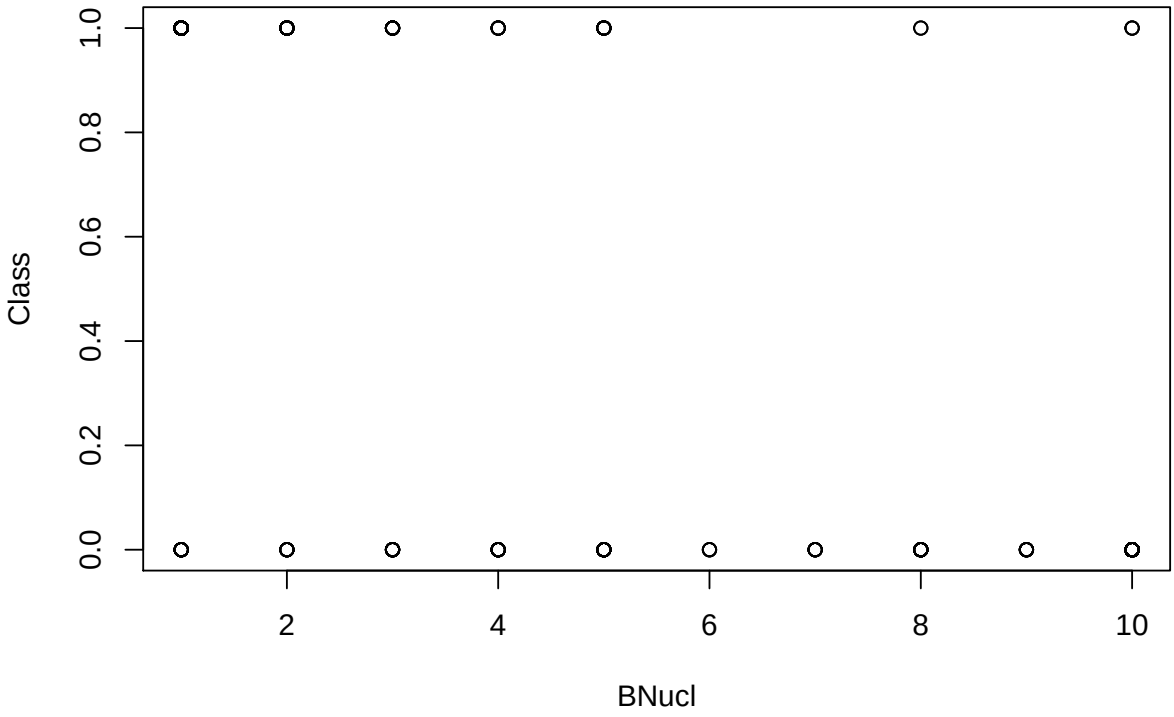
Bruce Campbell

2.1 wbca analysis

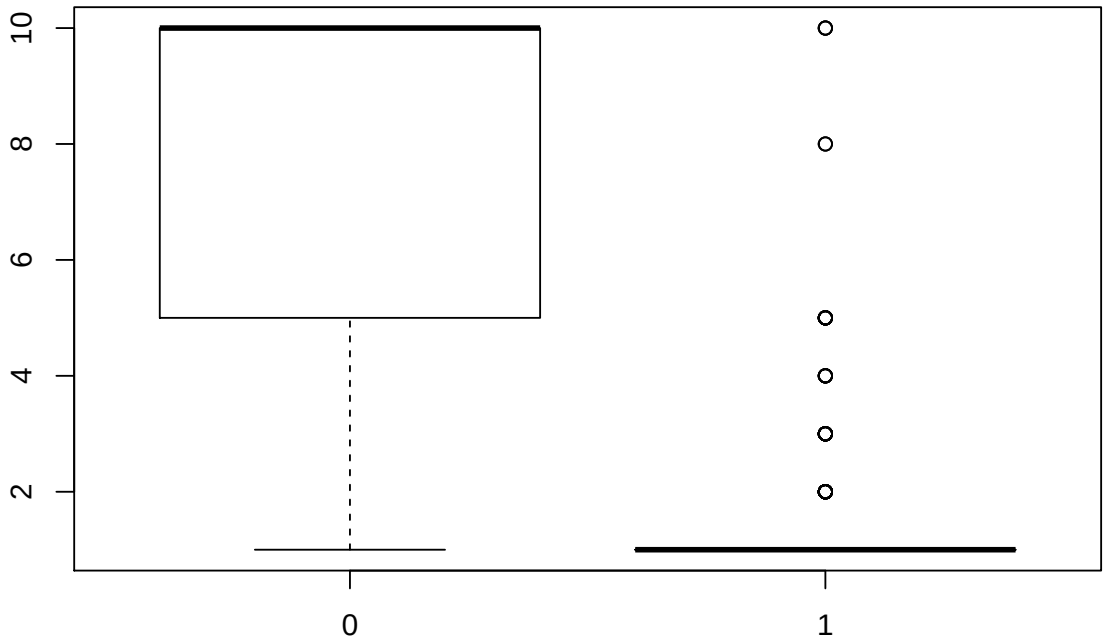
The dataset wbca comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

(a) Plot the relationship between the classification and BNucl. i. Explain why plot(Class ~ BNucl, wbca) does not work well. ii. Create a factor version of the response and produce a version of the first panel of Figure 2.1. Comment on the shape of the boxplots. iii. Produce a version of the second panel of Figure 2.1. What does this plot say about the distribution? iv. Produce a version of the interleaved histogram shown in Figure 2.2 and comment on the distribution.

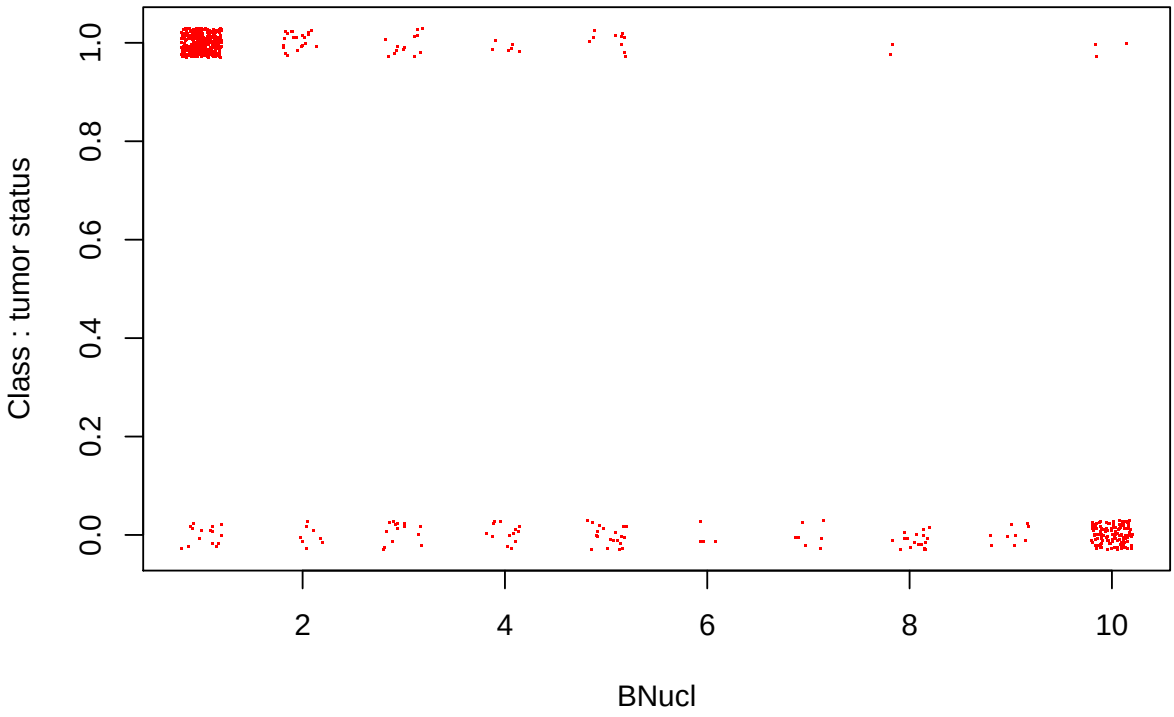
Here we plot $Class \sim BNucl$



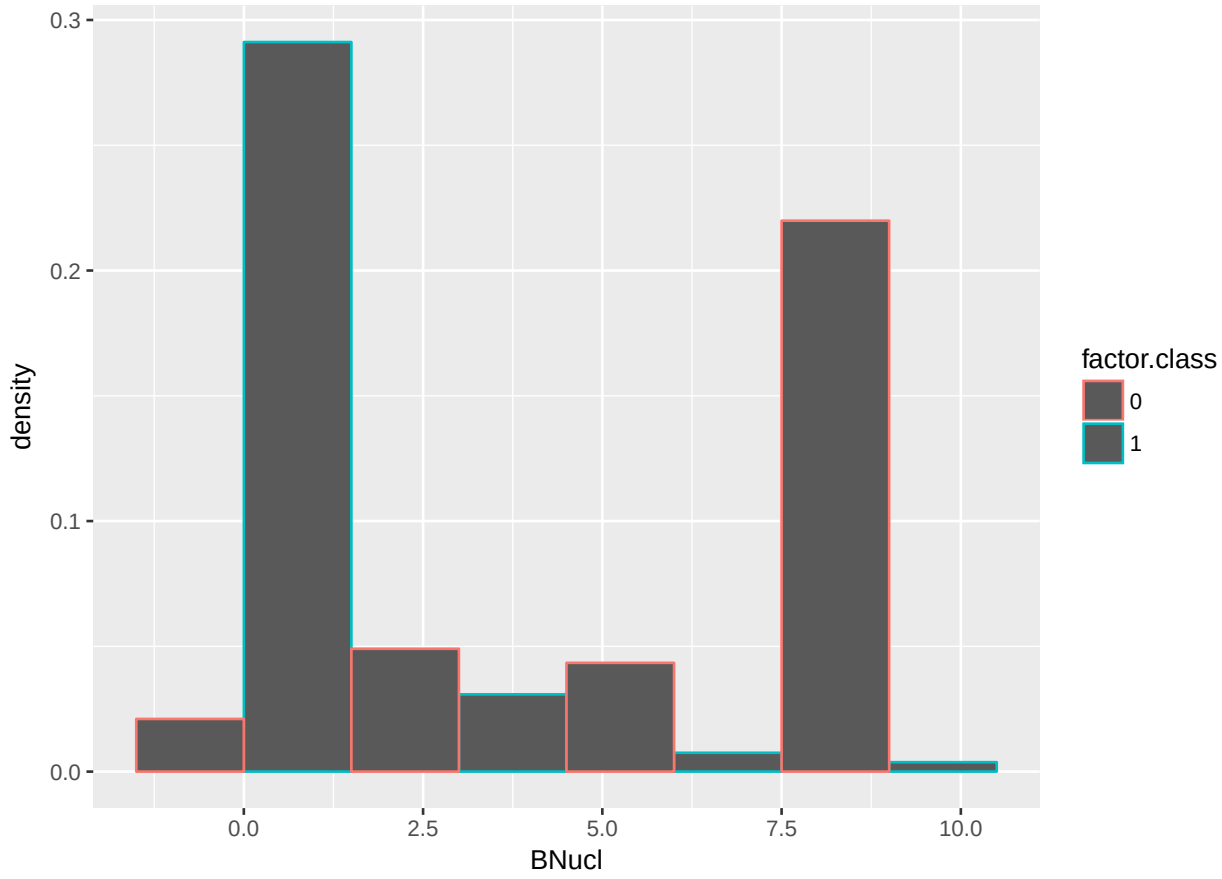
We see that since $BNucl$ is discrete we don't have a sense of how the variable is distributed by class well since the points overlap on the plot. A box plot provides a better visualization of the distribution by class.



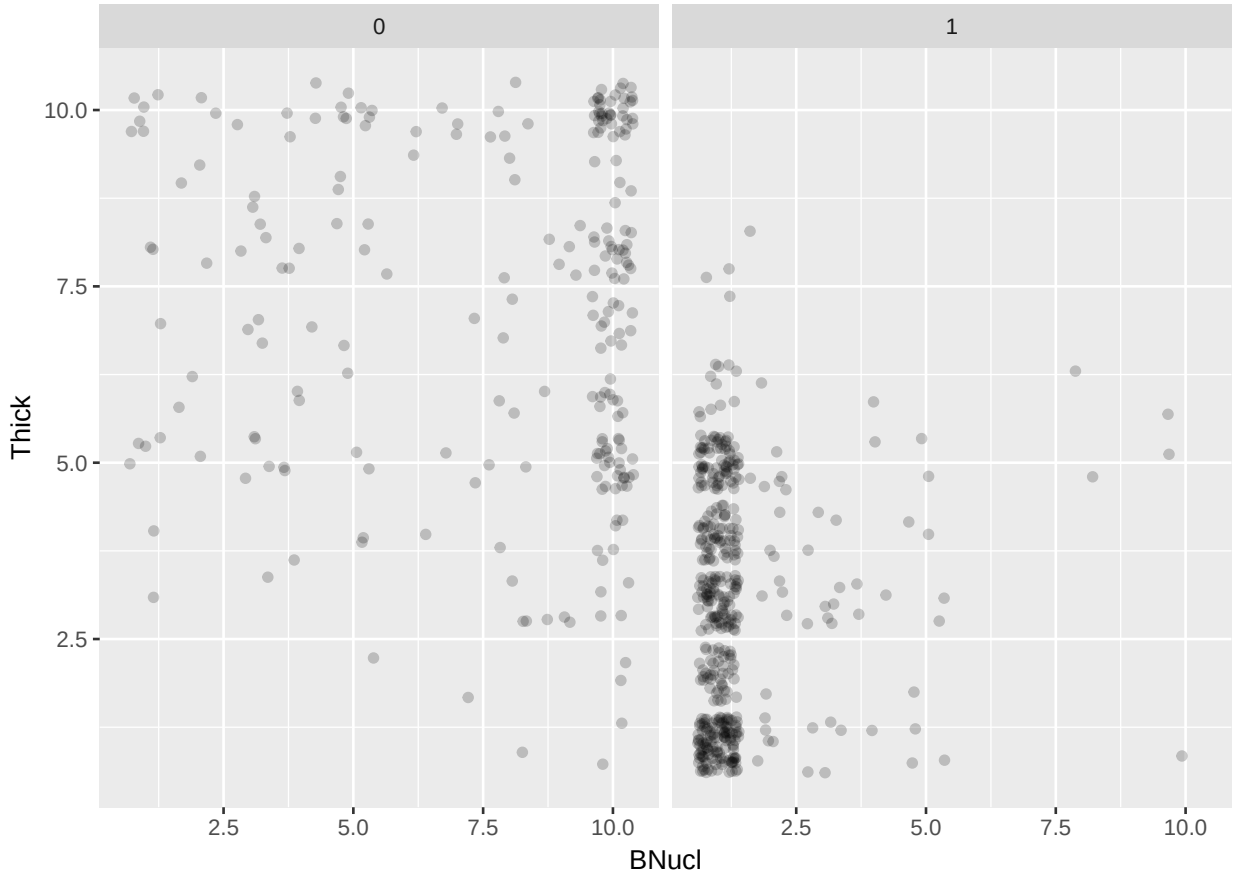
The boxplot show us that the $BNucl$ feature is a viable candidate for predicting cancer status. We can also add noise to the $Class \sim BNucl$ plot to remove the overlap in the points.

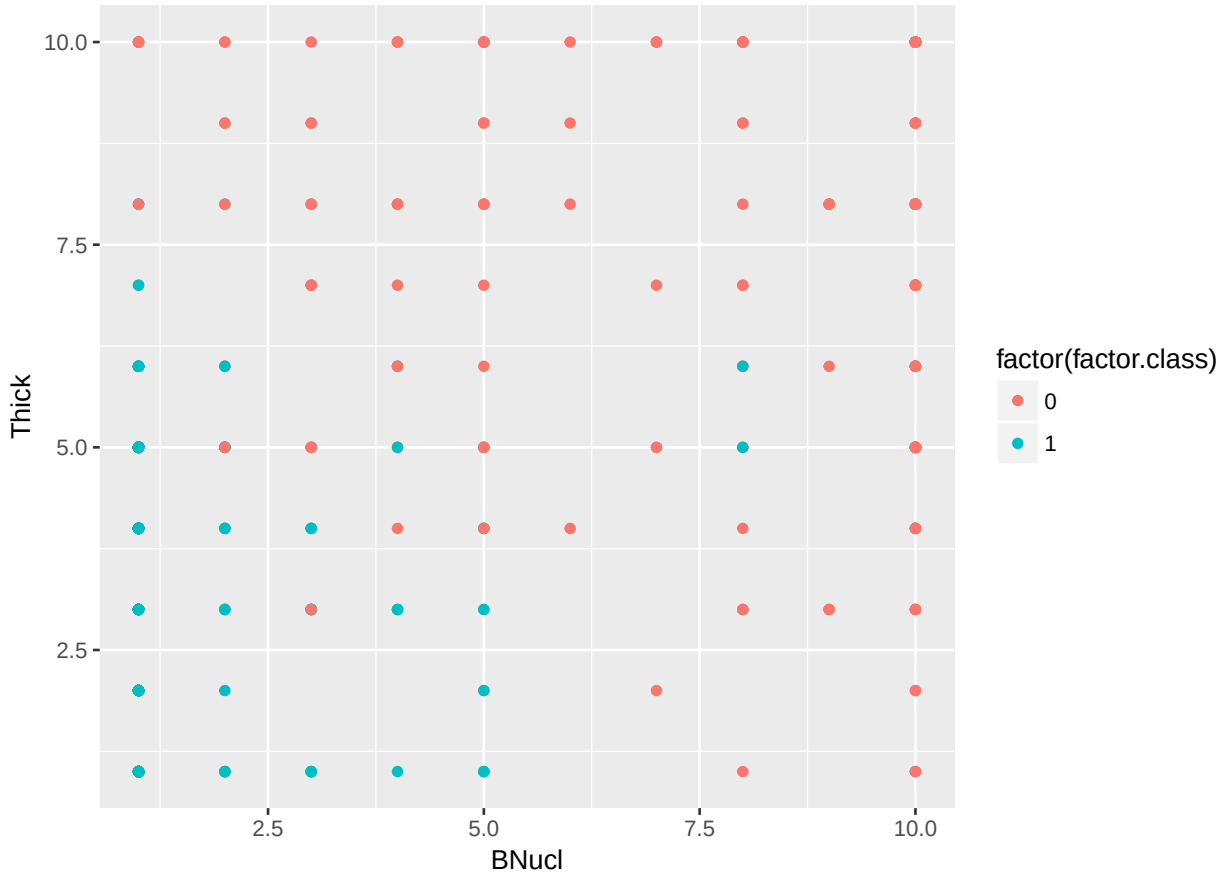


It looks like the $BNucl$ feature may be conditionally (on the class) modeled as a multinomial distribution. Most of the mass for the positive class is located at $BNucl = 1$ while most of the mass for the negative class is located at $BNucl = 10$.


0.3
0.2
0.1
0.0
density
0.0
2.5
5.0
7.5
10.0
BNucl
factor.class
0
1

(b) Produce a version of Figure 2.3 for the predictors BNucl and Thick. Produce an alternative version with only one panel but where the two types are plotted differently. Compare the two plots and describe what they say about the ability to distinguish the two types using these two predictors.





We see that a higher value of $BNucl$ is associated with an elevated value of $Thick$ and that a lower value of $BNucl$ is associated with a lower value of $Thick$. $Thick$ is a good candidate for inclusion in a mode using $BNucl$ to discriminate cancer status.

(c) Fit a binary regression with Class as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.

```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = wbca)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48282  -0.01179   0.04739   0.09678   3.06425
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.16678    1.41491   7.892 2.97e-15 ***
## Adhes       -0.39681    0.13384  -2.965  0.00303 **
## BNucl       -0.41478    0.10230  -4.055 5.02e-05 ***
```

```
## Chrom       -0.56456    0.18728  -3.014  0.00257 **
## Epith       -0.06440    0.16595  -0.388  0.69795
## Mitos       -0.65713    0.36764  -1.787  0.07387 .
## NNucl       -0.28659    0.12620  -2.271  0.02315 *
## Thick       -0.62675    0.15890  -3.944 8.01e-05 ***
## UShap       -0.28011    0.25235  -1.110  0.26699
## USize        0.05718    0.23271   0.246  0.80589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.388  on 680  degrees of freedom
## Residual deviance:  89.464  on 671  degrees of freedom
## AIC: 109.46
##
## Number of Fisher Scoring iterations: 8
```

The deviance is used for hypothesis testing in model comparison. Since our response is Bernoulli we can not use the deviance for evaluating goodness of fit. To use the deviance in this setting we would bin the responses to approximate a binomially distributed response

(e) Suppose that a cancer is classified as benign if p $>$ 0.5 and malignant if p $<$ 0.5. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.

```
##    class.predicted
##     FALSE TRUE
##   0   228   10
##   1     9  434
```

We see that the false positive rate is $10/(228 + 10) = 0.04201681$ and the false negative rate is $9/(434 + 9) = 0.02031603$

(f) Suppose we change the cutoff to 0.9 so that p $<$ 0.9 is classified as malignant and p $>$ 0.9 as benign. Compute the number of errors in this case.

```
##    class.predicted
##     FALSE TRUE
##   0   237    1
##   1    16  427
```

We see that the false positive rate is $1/(237 + 1) = 0.004201681$ and the false negative rate is $16/(427 + 16) = 0.03611738$

(h) It is usually misleading to use the same data to fit a model and test its predictive ability. To investigate this, split the data into two parts - assign every third observation to a test set and the remaining two thirds of the data to a training set. Use the training set to determine the model and the test set to assess its predictive performance. Compare the outcome to the previously obtained results.

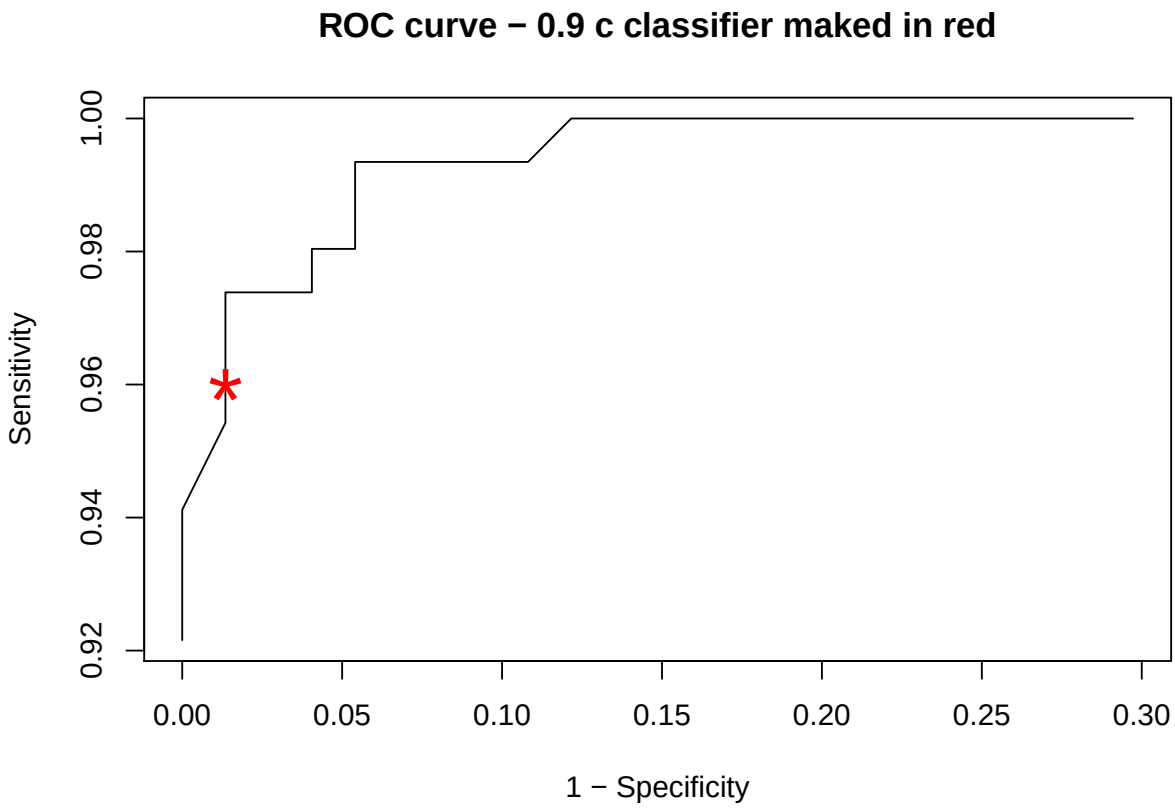
```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = DFTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3520  -0.0123   0.0406   0.0969   3.1771
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.65211    1.84168   6.327  2.5e-10 ***
## Adhes       -0.39057    0.19726  -1.980 0.047708 *
## BNucl       -0.46793    0.13566  -3.449 0.000562 ***
## Chrom       -0.65578    0.23207  -2.826 0.004716 **
## Epith       -0.02961    0.24675  -0.120 0.904469
## Mitos       -0.57934    0.51120  -1.133 0.257094
## NNucl       -0.25630    0.15143  -1.693 0.090543 .
## Thick       -0.80067    0.21174  -3.781 0.000156 ***
## UShap       -0.23802    0.26885  -0.885 0.375988
## USize        0.17773    0.24495   0.726 0.468109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 593.945  on 453  degrees of freedom
## Residual deviance:  62.044  on 444  degrees of freedom
## AIC: 82.044
##
## Number of Fisher Scoring iterations: 9
```

Table 1: Confusion matrix p=0.9

	FALSE	TRUE
0	73	1
1	6	147

Table 2: Confusion matrix p=0.5

	FALSE	TRUE
0	70	4
1	3	150



We see that we have error rate of $(1+6)/(73+1+6+147) = 0.030837$ using the $p = 0.9$ threshold. Using the threshold $p = 0.5$ we have an accuracy $(4+3)/(70+4+3+150) = 0.030837$ In this case we have very good evidence that the classifier will perform well on new data. We note that our total error rates for the 2 models are the same - we may prefer one over the other based on the class conditional error rate. A ROC curve may help us in model tuning.