Ecol 145 Assignment 9
Dahl Winters
4/7/06

Question 1

Fit a logistic regression model that uses all the variables as main effects, i.e., you need not consider the possibility of variable interactions at this point. Think long and hard about the variables **color** and **spine** before you blindly include them in the model.

- Be sure to justify the structural form you chose for the **weight** variable.

```
crabs<-
read.table('http://www.unc.edu/courses/2006spring/ecol/145/001/data/midterm/crabs.txt', header=TRUE, sep='')
```

The binary response variable Y for presences and absences

```
Y<-ifelse(crabs$num.satellites==0, Y<-0, Y<-1)
```

```
table(crabs$color)
 2  3  4  5
12 95 44 22
table(crabs$spine)
  1   2   3
 37  15 121
```

```
color.f<-factor(crabs$color)
spine.f<-factor(crabs$spine)
```

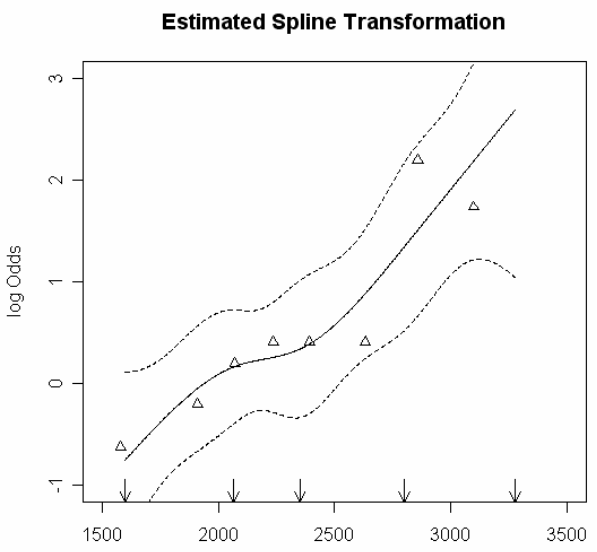
```
contrasts(color.f)
  3 4 5
2 0 0 0
3 1 0 0
4 0 1 0
5 0 0 1
```

```
contrasts(spine.f)
  2 3
1 0 0
2 1 0
3 0 1
```

Color and spine are categorical variables, unlike width and weight. Thus, they must be converted to factors before being incorporated into later models.

The structural form of the width regressor is linear because I found in Assignment 8 that the plot of the logit of the presence/absence of satellite males vs. width was linear. As for the correct form for the weight regressor, I first decided to do a plot of the logit of the presence/absence of the satellite males vs. the weight, and fit a lowess curve to see if that could tell me anything about whether the weight should be inputted as a linear or quadratic term. However, because the lowess curve fit is very sensitive to the number of groups I create (deciles vs. octiles, for example), I later decided to use the rcspline.plot function.

```
library(Design)
rcspline.plot(y=Y,x=crabs$weight,nk=5,m=20)
```



The curve goes upward in a fairly linear manner, which suggests that weight should be inputted into our model as a linear term.

Also, fitting a model with and without a quadratic weight term, I found that the model with a linear term had a lower AIC than the one that had a quadratic term. This is another bit of information that suggests the weight should be included as a linear term.

```
# the full model including weight as a linear term
fullmodel<-glm(Y~width+weight+color.f+spine.f, data=crabs, family=binomial)
AIC(fullmodel)
[1] 201.202

#the full model including weight as a quadratic term
fullmodel2<-glm(Y~width+weight+I(weight^2)+color.f+spine.f, data=crabs,
family=binomial)
AIC(fullmodel2)
[1] 203.1600
```

- Based on the summary output for this "full" model, what do you conclude about the effects of the various variables on the presence/absence of satellite males?

```
summary(fullmodel)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.6151215  3.8442046  -2.241   0.0250 *
width        0.2631279  0.1952986   1.347   0.1779
```

```
weight       0.0008258  0.0007038   1.173   0.2407
color.f1    -0.0514512  0.3912956  -0.131   0.8954
color.f2    -0.1458043  0.1838350  -0.793   0.4277
color.f3    -0.3528526  0.1471485  -2.398   0.0165 *
spine.f2    -0.0959809  0.7033698  -0.136   0.8915
spine.f3     0.4002868  0.5027043   0.796   0.4259
```

None of the variable estimates have significant p-values except for one, color.f3, though its p-value isn't much smaller than 0.05. Thus, this is the only variable that appears to have any effect on the presence/absence of satellite males.

Question 2

In light of your analysis in Assignment 8 is there anything troubling about your answer to Question 1? What do you think may be going on?

In Assignment 8 I fit a model that only had width as a predictor, and found that width was a significant predictor. The Wald test in that summary output gave p-values that were orders of magnitude under 0.05 (and thus significant) than what the model in Question 1 gave for the intercept and width. It appears that including the extra variables of weight, color, and spine is making the model worse instead of better at predicting the presence/absence of satellite males.

Question 3

Refit the model of Question 1 but this time without the **weight** variable. Examine the output from the summary function and answer question 1 again for this new model. What's changed? Explain why this change has occurred.

```
model3<-glm(Y~width+color.f+spine.f, data=crabs, family=binomial)
summary(model3)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.09953    2.97706  -3.728 0.000193 ***
width         0.45624    0.10779   4.233 2.31e-05 ***
color.f3     -0.14340    0.77838  -0.184 0.853830
color.f4     -0.52405    0.84685  -0.619 0.536030
color.f5     -1.66833    0.93285  -1.788 0.073706 .
spine.f2     -0.05782    0.70308  -0.082 0.934453
spine.f3      0.37703    0.50191   0.751 0.452540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 186.61  on 166  degrees of freedom
AIC: 200.61
```

The AIC has decreased slightly (from 201.17 to 200.61), but a bigger difference is that the p-values for both the intercept and the width have become much smaller (and significant) than they were in the model that included weight as a predictor. This tells me that weight should be thrown out of the model, and that the width is a significant predictor of the presence/absence of satellite males.

This model had only width as a predictor, while the model in Question 1 had both width and weight as predictors. These two variables are likely correlated because wider crabs will likely weigh more (unless they shrink in length when they become wider—knowing something about what the crabs look like would help). Without knowing what the crabs look like, I can at least check for a correlation. Sometimes models

having two or more highly correlated variables have a poorer fit to the data than models where those correlations are eliminated, as we saw in the last lab for the meanmin and meanmax variables.

```
cor(crabs$width,crabs$weight)
[1] 0.8868715
```

This is a rather high value (close to 1), so these variables are in fact very correlated. This result suggests we should remove one of the two variables from our model.

Question 4

In light of your answer to Question 3, we will no longer include **weight** among the set of predictors. Using the remaining three variables find a best main effects logistic regression model for these data. Be sure to justify the steps you go through in declaring this model to be best.

I used stepAIC to try different combinations of the remaining three predictors and compare them to find the best model. I set the upper bound model to be the one that considers all possible interactions between width, color.f, and spine.f by using the asterisk notation. The lower bound model is the model without any predictors (just an intercept).

The stepAIC function created models using all possible combinations of predictors and their interactions, and found that the model without spine had the lowest AIC.

```
model3<-glm(Y~width+color.f+spine.f, data=crabs, family=binomial)

library(MASS)

stepAIC(model3, scope=list(upper=~width*color.f*spine.f,lower=~1))
Start:  AIC= 200.61
 Y ~ width + color.f + spine.f

                  Df Deviance    AIC
- spine.f          2   187.46 197.46
<none>                 186.61 200.61
+ width:color.f    3   181.64 201.64
- color.f          3   194.43 202.43
+ color.f:spine.f  6   177.60 203.60
+ width:spine.f    2   186.41 204.41
- width            1   208.83 220.83

Step:  AIC= 197.46
 Y ~ width + color.f

                Df Deviance    AIC
<none>               187.46 197.46
- color.f        3   194.45 198.45
+ width:color.f  3   183.08 199.08
+ spine.f        2   186.61 200.61
- width          1   212.06 220.06

Call:  glm(formula = Y ~ width + color.f, family = binomial, data = crabs)

Coefficients:
(Intercept)        width     color.f3     color.f4     color.f5
  -11.38519      0.46796      0.07242     -0.22380     -1.32992
```

```
Degrees of Freedom: 172 Total (i.e. Null);  168 Residual
Null Deviance:      225.8
Residual Deviance: 187.5        AIC: 197.5
```

Question 5

Now consider a model that includes interactions among the three variables. You have my blessing at this point to use an automated variable selection routine if you wish. What model do you come up with? Interpret the parameter estimates of all the predictors that occur in your final model.

The stepAIC output above already considered all possible two-variable interactions among the three variables, due to what I entered as the upper bound model. It concluded that the best model (with the lowest AIC) was Y~width+color.f, which did not account for any interactions.

```
model4<-glm(Y~width+color.f, data=crabs, family=binomial)
summary(model4)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.38519    2.87346  -3.962 7.43e-05 ***
width         0.46796    0.10554   4.434 9.26e-06 ***
color.f3      0.07242    0.73989   0.098    0.922
color.f4     -0.22380    0.77708  -0.288    0.773
color.f5     -1.32992    0.85252  -1.560    0.119
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46
```

Interpretation of parameter estimates:

Our model is in the form of $Y = \alpha_0 + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, where α_0 = the width, and β_0 through β_3 are the 4 color types (colors 2-5). The coefficients above were estimated for the model on the logit scale. Exponentiating them gives their values on the probability scale:

```
exp(coef(model4)[2:5])
    width  color.f3  color.f4  color.f5
1.5967271 1.0751035 0.7994769 0.2644987
```

The width coefficient estimate α_0 is an odds ratio that measures the effect of increasing the width by 1 on the odds that Y (the presence/absence) = 1. Since α_0 = 1.5967271, this means for every increase in width of that amount there is a unit increase in the odds that the female will have at least one satellite male. The three color estimates, after exponentiating, are estimates of β_1, β_2, and β_3, as described in the equations below. These exponentiated estimates give us the effect, for each of the different colors of female crabs, of increasing the width by 1 on the presence/absence of satellite males.

```
contrasts(color.f)
  3 4 5
2 0 0 0
3 1 0 0
4 0 1 0
5 0 0 1
```

color 2: $\text{logit}(p) = \beta_0 + \alpha_0 * \text{width}$
color 3: $\text{logit}(p) = (\beta_0+\beta_1) + \alpha_0 * \text{width}$
color 4: $\text{logit}(p) = (\beta_0+\beta_2) + \alpha_0 * \text{width}$
color 5: $\text{logit}(p) = (\beta_0+\beta_3) + \alpha_0 * \text{width}$

Question 6

If you compare the model you obtained in Question 5 against various nested simpler models using appropriate significance tests, which model would you conclude is best?

I'll use the LR test outputted in the results of ANOVA model comparisons, comparing the model from Question 5 with models of Y~width, Y~color.f, and Y~1.

```
model4<-glm(Y~width+color.f, data=crabs, family=binomial) #from Question 5
model5<-glm(Y~width, data=crabs, family=binomial)
model6<-glm(Y~color.f, data=crabs, family=binomial)
model7<-glm(Y~1, data=crabs, family=binomial)
```

```
anova(model4,model5,test='Chisq')
Model 1: Y ~ width + color.f
Model 2: Y ~ width
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       168    187.457
2       171    194.453  -3   -6.996     0.072
```

The p-value is just above 0.05—it's not small enough to suggest that color.f should be counted as a significant predictor, but it's close.

```
anova(model4,model6,test='Chisq')
Model 1: Y ~ width + color.f
Model 2: Y ~ color.f
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       168    187.457
2       169    212.061  -1  -24.604 7.041e-07
```

The p-value here is small, so width is a significant predictor and should be included in the model. Thus, model4 is still the best model.

```
anova(model4,model7,test='Chisq')
Model 1: Y ~ width + color.f
Model 2: Y ~ 1
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       168    187.457
2       172    225.759  -4  -38.301  9.71e-08
```

The results of this comparison support those from the one above, that having either of the two predictors is better than having none. Thus, model4 remains the best model of all of these.

Question 7

Graph your final model from Question 5 on a probability scale.

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

Plotting the presence-absence data and model results on the probability scale

color 2: logit(p) = $\beta_0 + \alpha_0$*width
color 3: logit(p) = $(\beta_0+\beta_1) + \alpha_0$*width
color 4: logit(p) = $(\beta_0+\beta_2) + \alpha_0$*width
color 5: logit(p) = $(\beta_0+\beta_3) + \alpha_0$*width

p(color 2) = $1/(1+e^{-(\beta 0 + \alpha 0*\text{width})})$
p(color 3) = $1/(1+e^{-((\beta 0+ \beta 1) + \alpha 0*\text{width})})$
p(color 4) = $1/(1+e^{-((\beta 0+ \beta 2) + \alpha 0*\text{width})})$
p(color 5) = $1/(1+e^{-((\beta 0+ \beta 3) + \alpha 0*\text{width})})$

The three color estimates after exponentiating are estimates of β_1, β_2, and β_3; α_0 is the width coefficient estimate.

```
exp(coef(model4)[2:5])
    width  color.f3  color.f4  color.f5
1.5967271 1.0751035 0.7994769 0.2644987
```

α_0 β_1 β_2 β_3

plotting the points

```
plot(Y~crabs$width,xlab='Width',ylab='Probability of Presence/Absence')
```

highlighting the points in different colors

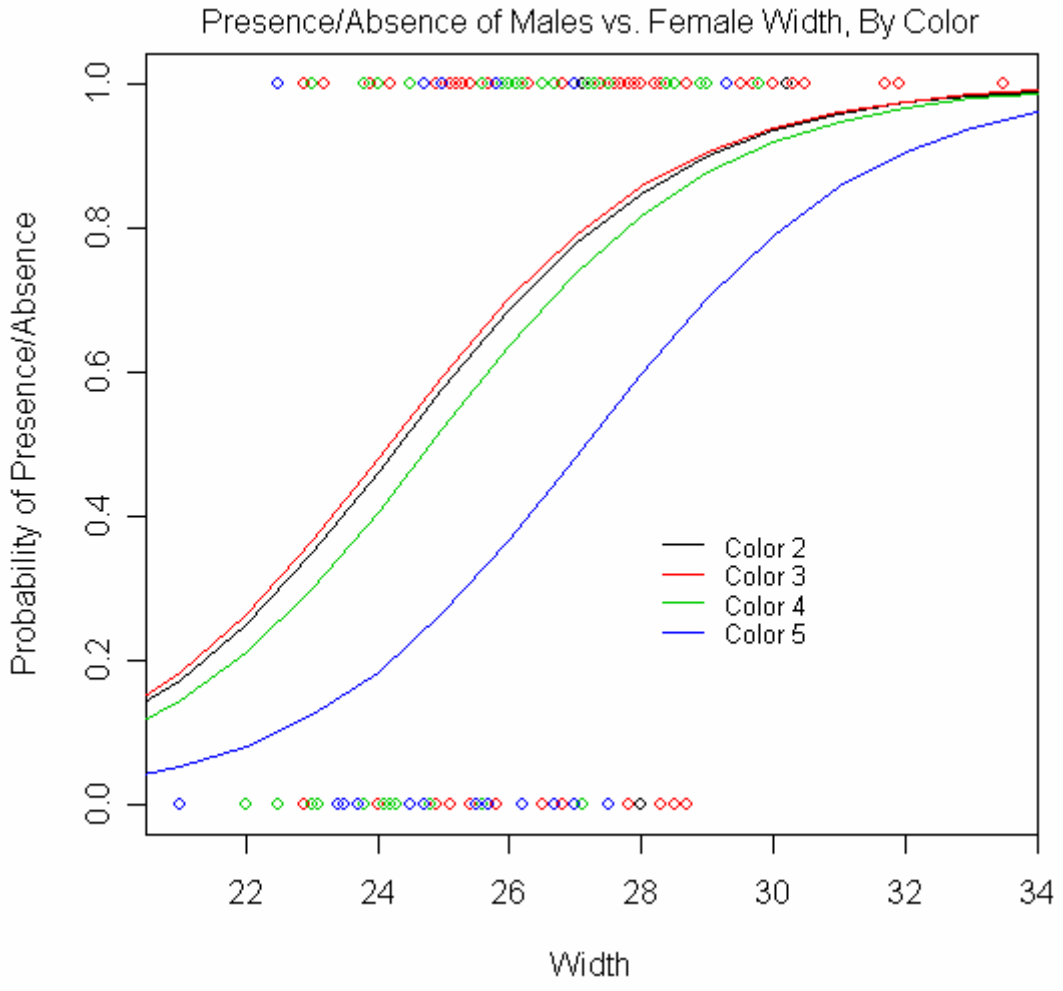
```
points(crabs$width[color.f==2], Y[color.f==2], col=1)
points(crabs$width[color.f==3], Y[color.f==3], col=2)
points(crabs$width[color.f==4], Y[color.f==4], col=3)
points(crabs$width[color.f==5], Y[color.f==5], col=4)
```

lines for colors 2, 3, 4, and 5

```
lines(seq(0,173,1), 1/(1+exp(-coef(model4)[1]-coef(model4)[2]*seq(0,173,1))),
col=1)
lines(seq(0,173,1), 1/(1+exp(-sum(coef(model4)[1],coef(model4)[3]) -
coef(model4)[2]*seq(0,173,1))), col=2)
lines(seq(0,173,1), 1/(1+exp(-sum(coef(model4)[1],coef(model4)[4]) -
coef(model4)[2]*seq(0,173,1))), col=3)
lines(seq(0,173,1), 1/(1+exp(-sum(coef(model4)[1],coef(model4)[5]) -
coef(model4)[2]*seq(0,173,1))), col=4)

mtext("Presence/Absence of Males vs. Female Width, By Color", side=3,
line=.5)

legend(28,.4, c('Color 2', 'Color 3', 'Color 4', 'Color 5'), col=c(1,2,3,4),
lwd=c(1,1,1,1), lty=c(1,1,1,1), cex=c(.8,.8,.8,.8), bty='n')
```



Question 8

Using an appropriate goodness of fit test, test the fit of your final model.

I used Harrell's alternative to the Hosmer-Lemeshow test, which was one of the three goodness-of-fit tests we did in the last assignment:

```
library(Design)
out.harrell<-lrm(Y~width+color.f,data=crabs,x=TRUE,y=TRUE)
residuals.lrm(out.harrell,type='gof')
Sum of squared errors     Expected value|H0                    SD
            31.6457412            31.5393772             0.3184185
                    Z                     P
             0.3340382             0.7383507
```

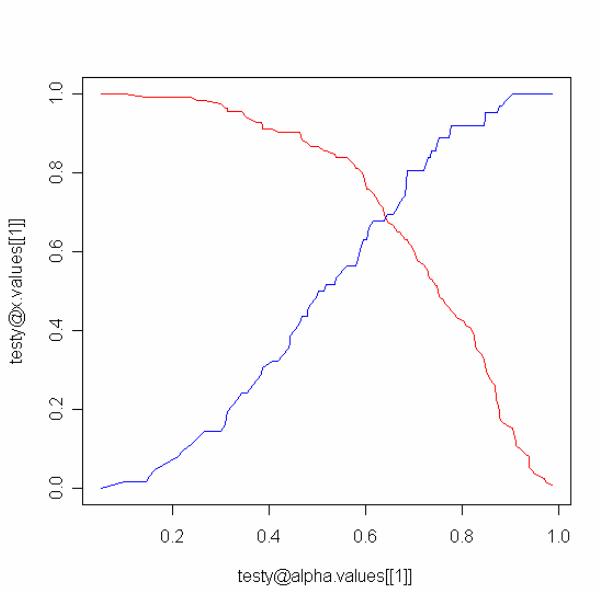
This test gives a p-value of 0.7383507, which is high above 0.05 and thus is significant in indicating the model has a good fit to the data.

Question 9

In this question we explore model calibration using the model obtained in Question 5.

1. What value of *c* should you use if your goal is to maximize both the specificity and sensitivity of your decision rule?

```
library(ROCR)
pred<-prediction(fitted(model4),Y)
testy<-performance(pred,'tpr','tnr')
plot(testy@alpha.values[[1]],testy@x.values[[1]],type='n')
lines(testy@alpha.values[[1]],testy@y.values[[1]],col=2) #the red line
lines(testy@alpha.values[[1]],testy@x.values[[1]],col=4) #the blue line
```



It appears that the value of c that would maximize both the sensitivity and the specificity is roughly 0.65.

To get a more exact value for c, we can look where testy@y.values[[1]] equals testy@x.values[[1]], which is where their difference equals 0.

```
testy@y.values[[1]]-testy@x.values[[1]]
```

Looking at the output, the 58th element of this vector is closest to zero.

```
testy@alpha.values[[1]][58]
[1] 0.6395257
```

The value of c we would want is 0.64, which is very close to my visual estimate.

2. Obtain the value of AUC, area under the curve, for your logistic regression model. Interpret the number you obtain.

```
lrm(Y~width+color.f,data=crabs)
Logistic Regression Model
lrm(formula = Y ~ width + color.f, data = crabs)
Frequencies of Responses
  0   1
 62 111
       Obs  Max Deriv Model L.R.       d.f.          P          C        Dxy
       173      4e-05       38.3          4          0      0.771      0.543
     Gamma      Tau-a         R2      Brier
     0.546      0.251      0.272      0.183

          Coef      S.E.   Wald Z P
```

```
Intercept -11.38519 2.8735 -3.96  0.0001
width       0.46796 0.1055  4.43  0.0000
color.f=3   0.07242 0.7399  0.10  0.9220
color.f=4  -0.22380 0.7771 -0.29  0.7733
color.f=5  -1.32992 0.8525 -1.56  0.1188
```

The AUC is given by C in the lrm output, which is 0.771. According to the scale given to us in class, a value of AUC between 0.7-0.8 is fair. The closer the AUC is to 1, the better the model is in discriminating presences of satellite males from absences, which are a concordant pair of 1s (presences) and 0s (absences). The AUC (also called the concordance index) also represents the fraction of the time that the model yields correct predictions of presences and absences in a pairwise comparison.

3. Carry out a 10-fold cross-validation and report the AUC you would expect to obtain with new data.

```
crabs$Y<-Y
crabs$color.f<-color.f
library(boot)
cost<-function(r, pi=0) mean(abs(r-pi)>0.5)
out<-cv.glm(crabs,model4,cost,K=10)
names(out)
[1] "call"  "K"     "delta" "seed"
out$delta
        1         1
0.2890173 0.2792275
```

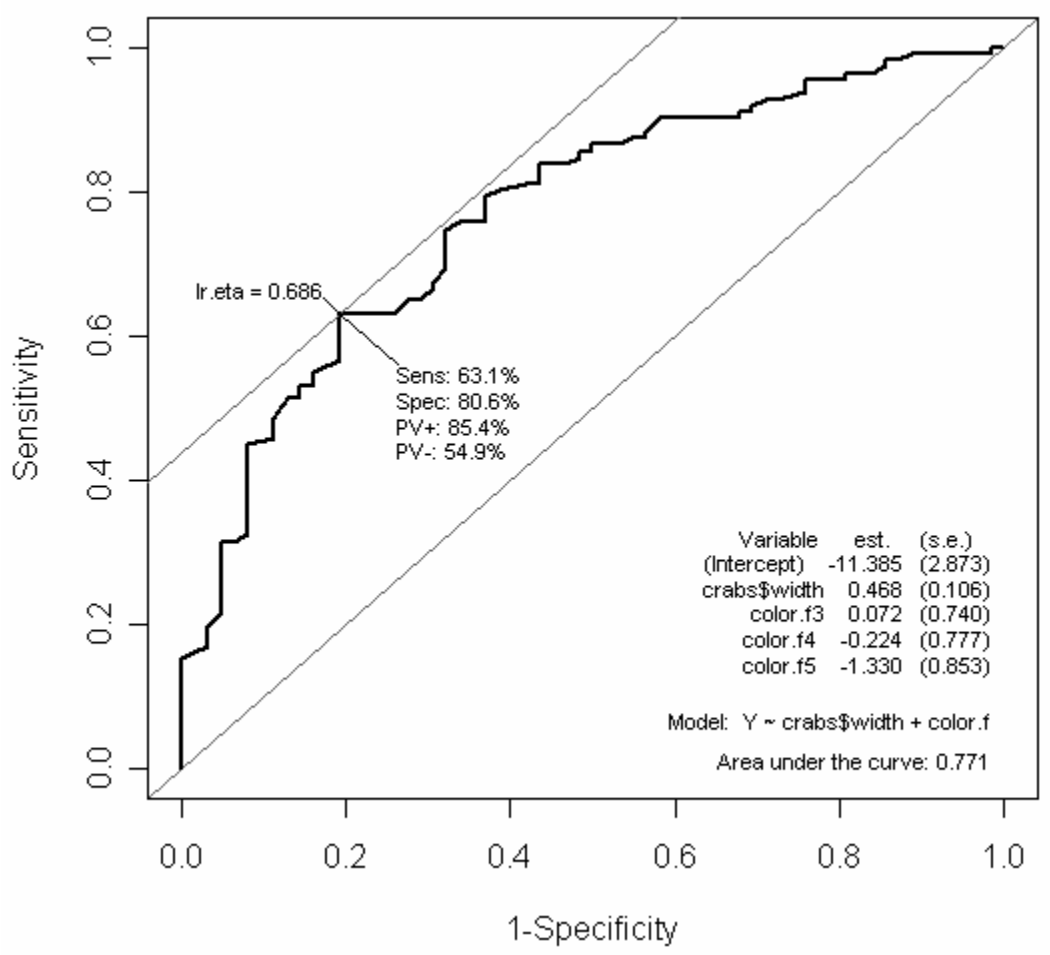
Either one of the above represents the fraction of the time that we will get inaccurate predictions from the model, given new data. I will use the second number because it is a bias-corrected version of the first number. Subtracting it from 1 gives the AUC:

```
1-out$delta[2]
        1
0.7207725
```

The 10-fold cross-validation thus gives a slightly lower AUC than the method in #2 above, which gave a value for the AUC of 0.771.

4. Plot the ROC curve for your logistic regression model.

```
library(Epi)
ROC(form=Y~crabs$width+color.f,plot="ROC")
```

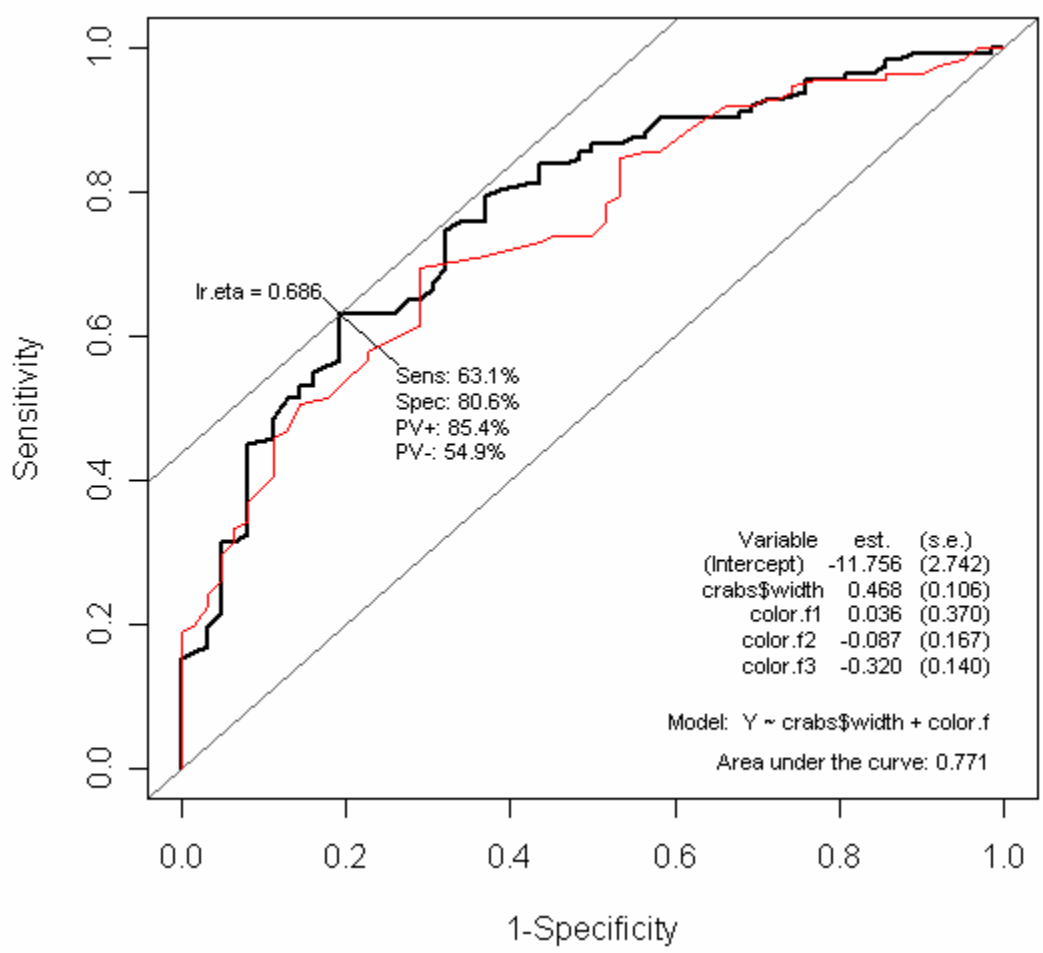


5. Do a second ROC plot but this time include both the ROC curve for the final model obtained in Question 5 and the ROC curve for the final model obtained in Assignment 8. Use different colors for the two curves. What can you conclude from this plot?

The Assignment 8 final model was inputted as model5 back in Question 6:

```
model5<-glm(Y~width, data=crabs, family=binomial)

ROC(form=Y~crabs$width+color.f,plot="ROC") #plots model from Question 5
par(new=TRUE)
library(ROCR)
pred<-prediction(fitted(model5), Y)
perf<-performance(pred,"tpr","fpr")
plot(perf, xlab='', ylab='', axes=FALSE, col=2) #plots the Assignment 8 graph
par(new=FALSE)
```



Just visually inspecting the areas beneath the two curves in the above graph, the curve for the final model in Question 5 (black) has more area beneath it than the curve for the Assignment 8 model (red). The only difference between the two models is the inclusion of the color.f categorical variable as a predictor. Thus, including this variable gives a model with a higher AUC than one without it, which means that this model does a better job more of the time in correctly predicting presences and absences of satellite males.

Question 10

Since the variable **color** represents shades of darkness, it might be treated as an ordinal variable. What evidence do you have from your logistic regression results to suggest that perhaps the log odds of a satellite male being present has an ordinal relationship to the variable color?

The three color estimates, after exponentiating, are estimates of β_1, β_2, and β_3. These adjust the intercepts for each of the 4 colors of crabs according to the equations below, and thus affect the log odds of a satellite male being present. From the logistic regression results below, we see that except for the first case, color.f3, the coefficients for color.f3, f4, and f5 become increasingly negative. This suggests that darker crabs are less likely to have a satellite male present, maybe because darker crabs are less visible.

color 2: logit(p) = $\beta_0 + \alpha_0$*width
color 3: logit(p) = $(\beta_0+\beta_1) + \alpha_0$*width
color 4: logit(p) = $(\beta_0+\beta_2) + \alpha_0$*width
color 5: logit(p) = $(\beta_0+\beta_3) + \alpha_0$*width

```
model4<-glm(Y~width+color.f, data=crabs, family=binomial)
```

```
summary(model4)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.38519    2.87346  -3.962 7.43e-05 ***
width         0.46796    0.10554   4.434 9.26e-06 ***
color.f3      0.07242    0.73989   0.098    0.922
color.f4     -0.22380    0.77708  -0.288    0.773
color.f5     -1.32992    0.85252  -1.560    0.119
```

Question 11

Fit a logistic regression model using **width** and ordinal **color** as predictors. Is there any evidence for linear, quadratic, or cubic trends?

```
color.o<-ordered(crabs$color)
model8<-glm(Y~width+color.o, data=crabs, family=binomial)
summary(model8)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.75552    2.74185  -4.287 1.81e-05 ***
width         0.46796    0.10554   4.434 9.26e-06 ***
color.o.L    -0.95837    0.58032  -1.651   0.0986 .
color.o.Q    -0.58927    0.47472  -1.241   0.2145
color.o.C    -0.09867    0.33738  -0.292   0.7699
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46
```

The significance tests done on the linear, quadratic, and cubic components of the color ordinal variable come back with all non-significant p-values. Thus we can conclude there is no evidence for any trends if the spacing between the color categories is constrained to being equal.

Question 12

A second way of handling ordinal data is to use Helmert contrasts. Based on the output from the summary function, what can you conclude about color and its effect on the presence/absence of satellite males this time? Interpret the color results as best you can.

```
contrasts(color.f)<-'contr.helmert'
contrasts(color.f)
  [,1] [,2] [,3]
2   -1   -1   -1
3    1   -1   -1
4    0    2   -1
5    0    0    3
```

color 2: logit(p) = $\beta_0-\beta_1-\beta_2-\beta_3+\alpha_0$*width
color 3: logit(p) = $\beta_0+\beta_1-\beta_2-\beta_3+\alpha_0$*width
color 4: logit(p) = $\beta_0+2\beta_2-\beta_3+\alpha_0$*width
color 5: logit(p) = $\beta_0+3\beta_3+\alpha_0$*width

```
model9<-glm(Y~width+color.f, data=crabs, family=binomial)
summary(model9)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.75552    2.74185  -4.287 1.81e-05 ***
width         0.46796    0.10554   4.434 9.26e-06 ***
color.f1      0.03621    0.36995   0.098    0.922
color.f2     -0.08667    0.16742  -0.518    0.605
```

```
color.f3     -0.31986    0.13966  -2.290    0.022 *
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46
```

The three color results are the log of β_1, β_2, and β_3, respectively. The coefficient for color.f3, which is $\log(\beta_3)$, is the only significant estimate due to its p-value of 0.022 (below 0.05). Due to the use of the Helmert contrast coding, the significance test for color.f3 tests whether color 5 has a significantly different effect on the presence/absence than colors 2, 3, and 4 averaged together. Being of color 5 would have a significantly different effect if color.f3 is significantly different from zero, which is the case, as indicated by the low p-value. This allows me to conclude that female crabs of color 5 (dark) have a significant effect on the presence/absence of satellite males than females of other color categories.

Question 13

Based on what you observed in Question 12 dichotomize color into two groups. Fit a logistic regression model that includes width and dichotomized color as predictors. How does this model compare to your model of Question 5?

For this question, I want to separate the 4 color categories into two. The first category will be colors 2, 3, and 4, and the second category will be color 5.

```
color.ld<-factor(ifelse(crabs$color<5, color.ld<-1, color.ld<-2))
color.ld
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 2 2 1 1 2 1 1 1 1 1
 [35] 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1
 [69] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1
[103] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1
[137] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
[171] 1 2 1
Levels: 1 2

model0<-glm(Y~width+color.ld, data=crabs, family=binomial)
summary(model0)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6790     2.6925  -4.338 1.44e-05 ***
width         0.4782     0.1041   4.592 4.39e-06 ***
color.ld2    -1.3005     0.5259  -2.473   0.0134 *
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom
AIC: 193.96
```

The AIC of this model is about 4 units lower than that from Question 5 (which was 197.46). Also, the three color coefficients estimated for that model were not significant, but this coefficient has a significant p-value (below 0.05). Both of these suggest that this model than the Question 5 model.