Ecol 145 Assignment 8
Dahl Winters
3/31/06

Question 1

Graphically investigate whether the presence/absence of satellite males seems to be linearly related (on a logit scale) to female width.

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

The binary response variable Y for presences and absences

```
Y<-ifelse(crabs$num.satellites==0, Y<-0, Y<-1)
Y
  [1] 1 0 1 0 1 0 0 0 0 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 0 1 0 0
 [34] 1 1 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1
 [67] 1 0 0 1 1 1 0 1 0 0 1 1 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 1 0 1 1
[100] 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 1 0 1 1 0 1 1 1 1 1 0
[133] 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 0 0 0 1 0 1 0 1 1 1 1 0 1
[166] 0 0 1 1 1 0 0 0
```

Creating deciles of the width

```
quantile(crabs$width, seq(0,1,.1))
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100% 
21.00 23.70 24.50 25.06 25.70 26.10 26.70 27.44 28.20 29.00 33.50 

table(cut(crabs$width, quantile(crabs$width, seq(0,1,.1)), include.lowest=
TRUE))
  [21,23.7] (23.7,24.5] (24.5,25.1] (25.1,25.7] (25.7,26.1] 
         19          18          15          19          16 
(26.1,26.7] (26.7,27.4] (27.4,28.2]   (28.2,29]   (29,33.5] 
         18          16          16          22          14 

width.decs<-cut(crabs$width, quantile(crabs$width, seq(0,1,.1)),
include.lowest=TRUE)
```

Finding the successes and total counts for the empirical logit, to get the y-values for the plot

```
tapply(Y, width.decs, sum)->sums #these are the successes
tapply(Y, width.decs, length)->lengths #these are the ns

sums
  [21,23.7] (23.7,24.5] (24.5,25.1] (25.1,25.7] (25.7,26.1] (26.1,26.7] 
          5           8          10           9          11          11 
(26.7,27.4] (27.4,28.2]   (28.2,29]   (29,33.5] 
         12          12          19          14 
lengths
  [21,23.7] (23.7,24.5] (24.5,25.1] (25.1,25.7] (25.7,26.1] (26.1,26.7] 
         19          18          15          19          16          18 
(26.7,27.4] (27.4,28.2]   (28.2,29]   (29,33.5] 
         16          16          22          14 

logit.p<-log((sums+1/2)/((lengths-sums)+1/2)) #empirical logit
```

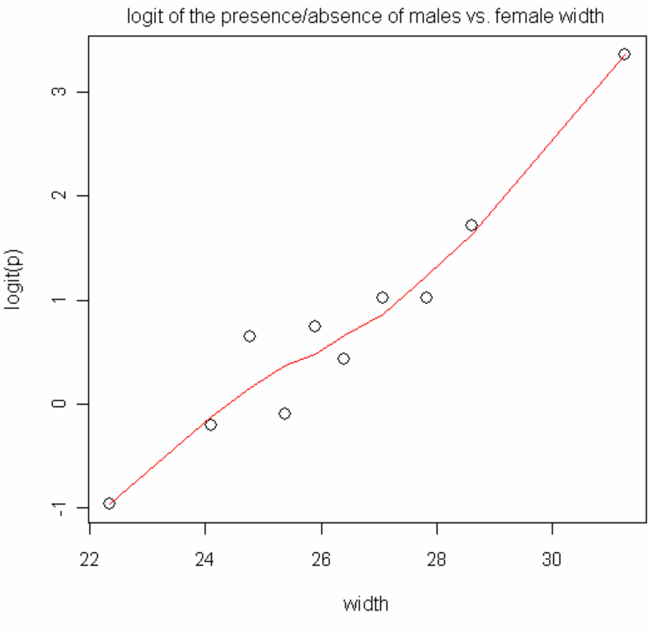
Finding the midpoints of each decile, which will be the x-values in the plot

```
mids<-(quantile(crabs$width, seq(0,1,.1))[1:10] + quantile(crabs$width,
seq(0,1,.1))[2:11])/2
mids
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90% 
22.35 24.10 24.78 25.38 25.90 26.40 27.07 27.82 28.60 31.25 
#This gives the left point of each interval + the right point of each
interval, divided by 2 to get the midpoint of each interval.
```

Plotting the logit of presence/absence vs. width

```
plot(mids, logit.p, xlab='width', ylab='logit(p)', axes=FALSE, cex=1.5)
axis(1,cex.axis=.9)
axis(2,cex.axis=.9)
box()
mtext(“logit of the presence/absence of males vs. female width”, side=3,
line=.5)
lines(lowess(logit.p~mids),col=2)
```

The plot of the logit vs. width does appear to be linear. A lowess curve (red) fitted to the data looks quite linear, which indicates that a line might be a good choice in describing this data.



Question 2

Fit a logistic regression model with Y (as defined above) as the response and width as the predictor using the functional form you decided was appropriate in Question 1.

1. Test whether there is a significant relationship between the presence-absence of males and the width of the female. Do this significance test in two distinct ways and report the results from both. What's the difference between the two tests?

Test 1: This test involves fitting 2 models, one with width as a predictor (model1) and one with no predictors (model2). If width is a significant predictor of the presence-absence of males, then we should expect to see a significant p-value for width and a smaller AIC for model1. When the two models are fit, this is certainly the case. The p-value of width for model1 is 1.02e-06, which is below 0.05 and indicates it is a significant predictor. Also, the AIC for model1 (198.45) is considerably smaller than that for model2 (227.76), which also lends support to the idea that width seems to be a significant predictor because a model without it does not describe the data as well.

```
model1<-glm(Y~width, data=crabs, family=binomial)
summary(model1)
Call:
glm(formula = Y ~ width, family = binomial, data = crabs)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.0458   0.5480   0.9066   1.6941
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45
Number of Fisher Scoring iterations: 4
```

```
model2<-glm(Y~1, data=crabs, family=binomial)
summary(model2)
Call:
glm(formula = Y ~ 1, family = binomial, data = crabs)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.433  -1.433   0.942   0.942   0.942
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5824     0.1585   3.673 0.000239 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 225.76  on 172  degrees of freedom
AIC: 227.76
Number of Fisher Scoring iterations: 4
```

Test 2: This is a likelihood ratio test of the two above models, which is printed in the output when doing an ANOVA comparison of the two models. The resulting p-value is very small, 2.204e-08, which again suggests that width is a significant predictor.

```
anova(model1, model2, test='Chisq')
Analysis of Deviance Table

Model 1: Y ~ width
Model 2: Y ~ 1
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       171    194.453
2       172    225.759  -1  -31.306 2.204e-08
```

2. Interpret in words the coefficient of width that is estimated for the model.

The coefficient of width that was estimated was 0.4972. Because the coefficient was estimated for the model on the logit scale, exponentiating it gives the value of β_1 on the probability scale:

```
exp(coef(model1)[2])
   width
1.644162
```

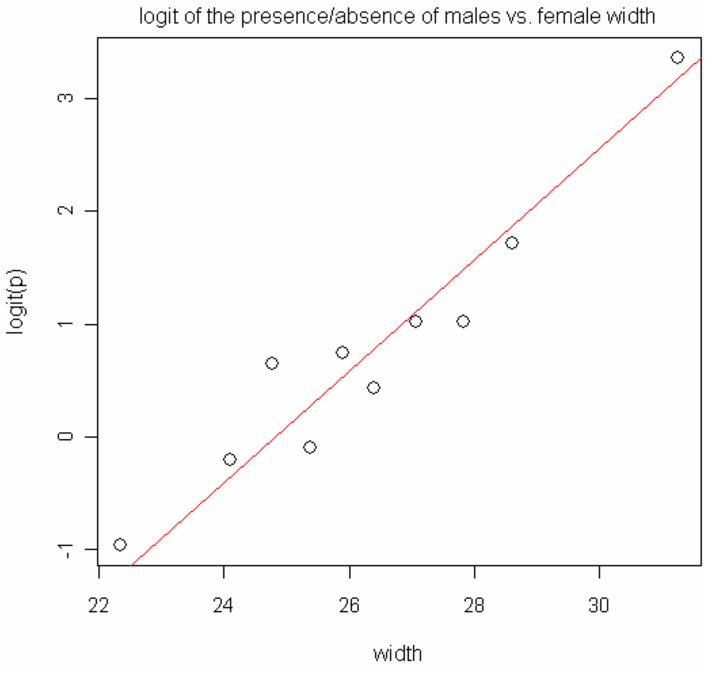
β_1 is the odds ratio, which measures the effect of increasing x_1 (the width) by 1 on the odds that Y (the presence/absence) = 1. Since $\beta_1 > 1$, then the odds of Y = 1 increases as the width increases. Specifically, because $\beta_1 = 1.644162$, this means for every increase in width of 1.644162 there is a unit increase in the odds that the female will have at least one satellite male.

Question 3

Plot the results of the logistic regression model including as much information as possible in your plot.

First I will plot the results of the model on the logit scale to show how well the line fits the data, and then I will plot the results on the probability scale.

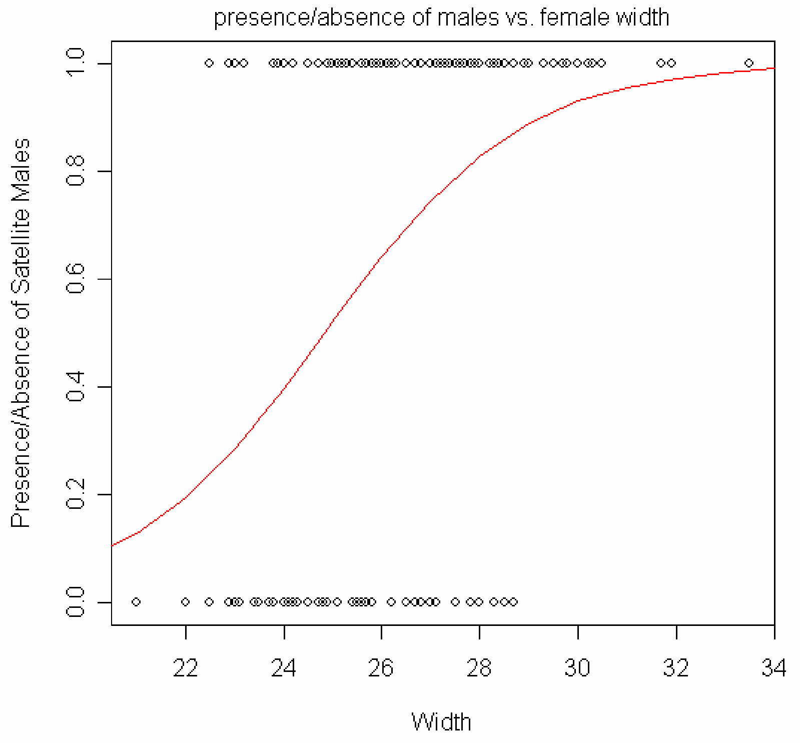
```
coef(model1)
(Intercept)       width
-12.3508177   0.4972306

plot(mids, logit.p, xlab='width', ylab='logit(p)', axes=FALSE, cex=1.5)
axis(1,cex.axis=.9)
axis(2,cex.axis=.9)
box()
mtext("logit of the presence/absence of males vs. female width", side=3,
line=.5)
abline(coef(model1)[1], coef(model1)[2], col=2) #1 = intercept, 2 = slope
```



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

```
plot(Y~crabs$width,xlab='Width',ylab='Presence/Absence of
Satellite Males')
lines(seq(0,173,1), 1/(1+exp(-coef(model1)[1]-
coef(model1)[2]*seq(0,173,1))), col=2)
mtext(“presence/absence of males vs. female width”, side=3, line=.5)
```

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



Question 4

Test your model for lack of fit. Do this in three different ways.

1. By forming groups using the width variable to form categories and then carrying a Pearson chi-square test or a G^2 test.

From question 1:

```
width.decs<-cut(crabs$width, quantile(crabs$width, seq(0,1,.1)),
include.lowest=TRUE)
table(Y,width.decs) #the number of presences/absences in each decile
   width.decs
Y   [21,23.7] (23.7,24.5] (24.5,25.1] (25.1,25.7] (25.7,26.1] (26.1,26.7]
  0        14          10           5          10           5           7
  1         5           8          10           9          11          11
   width.decs
Y   (26.7,27.4] (27.4,28.2] (28.2,29] (29,33.5]
  0           4           4         3         0
  1          12          12        19        14
```

Many groups have fewer than 5 presences/absences, so I will try octiles instead of deciles.

```
width.octs<-cut(crabs$width, quantile(crabs$width, seq(0,1,.125)),
include.lowest=TRUE)
table(Y,width.octs)
   width.octs
Y   [21,23.9] (23.9,24.9] (24.9,25.6] (25.6,26.1] (26.1,26.9] (26.9,27.7]
  0        16          13           6           9           9           4
  1         6          11          13          13          12          18
   width.octs
Y   (27.7,28.7] (28.7,33.5]
  0           5           0
  1          17          21
```

Together, 2 of 8 (25%) of all the categories have fewer than 5 counts, but one of those categories has 4 counts (maybe this is OK) while the other has 0 counts (hopefully this is OK). I tried quintiles but this only raises the number of counts in the last category to 3, while sacrificing more degrees of freedom. So I decided to stick with octiles.

Obtain the number of expected successes (presences of males)

```
np<-tapply(fitted(model1),width.octs,sum)
np
  [21,23.9] (23.9,24.9] (24.9,25.6] (25.6,26.1] (26.1,26.9] (26.9,27.7]
    6.51022    10.93476    10.47002    13.72784    14.47732    17.02836
(27.7,28.7] (28.7,33.5]
   18.55970    19.29179
```

Obtain the total expected presences and absences in each category

```
apply(table(Y,width.octs),2,sum)
  [21,23.9] (23.9,24.9] (24.9,25.6] (25.6,26.1] (26.1,26.9] (26.9,27.7]
         22          24          19          22          21          22
(27.7,28.7] (28.7,33.5]
         22          21
```

Obtain the number of expected failures (absences of males)

```
fails<-apply(table(Y,width.octs),2,sum)-np
```

Calculate Ei and Oi

```
Ei<-rbind(fails,np)
Ei #expected presences (np) and absences (fails)
      [21,23.9] (23.9,24.9] (24.9,25.6] (25.6,26.1] (26.1,26.9] (26.9,27.7]
fails  15.48978    13.06524     8.52998     8.27216    6.522684    4.971644
np      6.51022    10.93476    10.47002    13.72784   14.477316   17.028356
      (27.7,28.7] (28.7,33.5]
fails    3.440303    1.708210
np      18.559697   19.291790
```

The last three categories have counts beneath 5 for the number of absences, but there are 16 categories total, so 3/16*100 = 18.75% of the categories have counts beneath 5. This is less than 20%, so I'll choose not to group the last two categories (the third to last one is close enough to 5) in order to preserve my degrees of freedom.

```
Oi<-table(Y,width.octs)
Oi
   width.octs
Y   [21,23.9] (23.9,24.9] (24.9,25.6] (25.6,26.1] (26.1,26.9] (26.9,27.7]
  0        16          13           6           9           9           4
  1         6          11          13          13          12          18
   width.octs
Y   (27.7,28.7] (28.7,33.5]
  0           5           0
  1          17          21
```

Calculate Pearson

```
sum((Oi-Ei)^2/Ei)
[1] 5.829648
1-pchisq(sum((Oi-Ei)^2/Ei),df=8-2) #8=num of groups, 2=intercept and width
[1] 0.4425412
```

The Pearson test gives a p-value above 0.05, indicating a good fit of the model to the data.

2. By grouping the predicted values using deciles and carrying out the Hosmer-Lemeshow test.

```
p.groups<-cut(fitted(model1),quantile(fitted(model1),seq(0,1,.1)),
include.lowest=TRUE)
table(p.groups)
p.groups
[0.129,0.362] (0.362,0.458] (0.458,0.527] (0.527,0.605] (0.605,0.652]
           19            18            15            19            16
(0.652,0.716] (0.716,0.785] (0.785,0.842] (0.842,0.888] (0.888,0.987]
           18            16            20            18            14
```

Obtain observed counts of presences/absences in each category

```
Oi<-table(Y,p.groups)
Oi
   p.groups
Y   [0.129,0.362] (0.362,0.458] (0.458,0.527] (0.527,0.605] (0.605,0.652]
  0            14            10             5            10             5
  1             5             8            10             9            11
   p.groups
Y   (0.652,0.716] (0.716,0.785] (0.785,0.842] (0.842,0.888] (0.888,0.987]
  0             7             4             4             3             0
  1            11            12            16            15            14
```

The total counts in each category

```
ni<-apply(Oi,2,sum)
ni
[0.129,0.362] (0.362,0.458] (0.458,0.527] (0.527,0.605] (0.605,0.652]
           19            18            15            19            16
(0.652,0.716] (0.716,0.785] (0.785,0.842] (0.842,0.888] (0.888,0.987]
           18            16            20            18            14
```

Obtain expected failures and successes

```
Ei<-rbind(ni-tapply(fitted(model1),p.groups,sum),tapply(fitted(model1),
p.groups,sum))
Ei
     [0.129,0.362] (0.362,0.458] (0.458,0.527] (0.527,0.605] (0.605,0.652]
[1,]     13.610632     10.375685      7.448845      8.017449      5.904548
[2,]      5.389368      7.624315      7.551155     10.982551     10.095452
     (0.652,0.716] (0.716,0.785] (0.785,0.842] (0.842,0.888] (0.888,0.987]
[1,]      5.701006      3.940702      3.733266      2.349822     0.9180457
[2,]     12.298994     12.059298     16.266734     15.650178    13.0819543
```

Carry out H-L test

```
sum((Oi-Ei)^2/Ei)
[1] 4.385541
1-pchisq(sum((Oi-Ei)^2/Ei),df=8)
[1] 0.8207722
```

The Hosmer-Lemeshow test gives a p-value (0.82) that is higher than from the Pearson test (0.44), but since they are both above 0.05, they both suggest that the model has a good fit.

3. By carrying out the alternative to the Hosmer-Lemeshow test contained in Frank Harrell's Design library.

```
library(Design)
out.harrell<-lrm(Y~width,data=crabs,x=TRUE,y=TRUE)
residuals.lrm(out.harrell,type='gof')
Sum of squared errors     Expected value|H0                    SD
           33.3562220            33.0342322             0.3031117
                    Z                     P
            1.0622807             0.2881083
```

This test gives a p-value of 0.288, which is lower than either of the two previous p-values. However, it is still above 0.05, so all 3 tests indicate that the model has a good fit to the data.