Horseshoe crab model selection:

First we begin with a straight forward additive linear model.

```r
> crabs=read.table("http://www.stat.ufl.edu/~dathien/STA6505/crabdata.txt", header=TRUE)
> attach(crabs)

> y=ifelse(satellite>0, 1, 0) # Y = a binary indicator of satellites
> weight=weight/1000 # weight in kilograms rather than grams
> color.f=factor(color);spine.f=factor(spine) # treat color and spine as a factor
> options(contrasts=c("contr.SAS","contr.poly"))
> fit=glm(y ~ weight+width+color.f+spine.f, family=binomial(link=logit))
> summary(fit)
```

Call:
`glm(formula = y ~ weight + width + color.f + spine.f, family = binomial(link = logit))`

Deviance Residuals:
```
                     Min          1Q     Median          3Q         Max
-2.1977       -0.9424       0.4849       0.8491       2.1198
```

Coefficients:
```
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)       -9.27340    3.83780  -2.416  0.01568 *
weight            0.82578    0.70382   1.173  0.24069
width             0.26312    0.19531   1.347  0.17788
color.f1          1.60871    0.93553   1.720  0.08552 .
color.f2          1.50582    0.56669   2.657  0.00788 **
color.f3          1.11981    0.59333   1.887  0.05910 .
spine.f1          -0.40034    0.50266  -0.796  0.42588
spine.f2          -0.49627    0.62926  -0.789  0.43024
```

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: 185.20 on 165 degrees of freedom
AIC: 201.2

Number of Fisher Scoring iterations: 4

A test of $H_0: \beta_1 = \cdots = \beta_7 = 0$, i.e. overall fit yields a difference of deviances of $225.76 - 185.20$ with df=7 and a corresponding p-value < 0.0001. Hence, there is strong evidence that at least one predictor is associated with the presence of satellites. However, looking at the individual Wald tests and the individual LR tests it appears that only color is marginally significant.
> drop1(fit,test="Chisq")

Single term deletions

Model:
y ~ weight + width + color.f + spine.f

Df Deviance AIC   LRT Pr(>Chi)
<none> 185.20 201.20
weight  1 186.61 200.61  1.4099 0.23507
width   1 187.00 201.00  1.7968 0.18010
color.f 3 192.80 202.80  7.5958 0.05515 .
spine.f 2 186.21 198.21  1.0091 0.60377

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Further, exploration indicates that this probably due to multicollinearity between weight and width.

> cor(weight,width)

[1] 0.8868715

Hence, we will drop one of the two terms. We choose weight.

There are multiple methods for model selection but here we will undertake backwards elimination.

> fit.csw=glm(y~color.f*spine.f*width,family=binomial(link=logit))
> summary(fit.csw)

Call:
  glm(formula = y ~ color.f * spine.f * width, family = binomial(link = logit))

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-2.1128  -0.8860   0.4891   0.7637  1.8594

Coefficients: (3 not defined because of singularities)

                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -4.073e+00  7.266e+00  -0.561  0.575
color.f1                6.059e+02  4.790e+05   0.001  0.999
color.f2               -4.996e+00  8.992e+00  -0.556  0.578
color.f3                1.465e+01  9.836e+00   0.490  0.623
spine.f1               1.118e+00  9.035e+04   0.000  1.000
spine.f2               -5.223e+02  4.411e+04  -0.012  0.991
width                  1.352e-01  2.833e-01   0.477  0.633
color.f1:spine.f1      -6.095e+00  4.875e+05  -0.001  0.999
color.f2:spine.f1      -1.038e+00  9.035e+04  -0.000  1.000
color.f3:spine.f1      3.617e+01  9.489e+03   0.004  0.997
color.f1:spine.f2      -6.095e+01  2.410e+04  -0.003  0.998
color.f2:spine.f2      5.364e+02  4.411e+04   0.012  0.990
color.f3:spine.f2     -8.750e+01  1.123e+04  -0.008  0.994
color.f1:width       -2.418e+01  1.857e+04  -0.001  0.999
color.f2:width       2.573e-01  3.477e-01   0.740  0.459
Here we notice that some coefficients cannot be estimated (unidentifiable model) and some have in effect infinite estimates (see large standard errors) due to sparse data. This model is unidentifiable as it has too many predictors for the amount of data points. There are 3 combinations of color and spine that have only one observation each, so 3 coefficients of the 3-way interactions cannot be estimated. The topic of sparse data will be covered in a later section but in effect think of certain combinations as having empty cells and thus 0 for odds ratio and $-\infty$ for log odds ratio.

A guideline is to have $\geq 10$ outcomes of each type per predictor. For the 173 horseshoe crabs, 111 had satellites and 62 did not. Therefore, the model should contain no more than (about) 6 predictors.

Let us remove the 3-way interaction.

```r
> fit.cs.cw.sw = update(fit.cs.w, ~ . - color.f:spine.f:width) # remove 3-way interaction
> summary(fit.cs.cw.sw)
```

**Deviance Residuals:**

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.0785</td>
<td>-0.8860</td>
<td>0.5088</td>
<td>0.8148</td>
<td>1.9250</td>
</tr>
</tbody>
</table>

**Coefficients:**

| Term                  | Estimate  | Std. Error | z value | Pr(>|z|) |
|-----------------------|-----------|------------|---------|---------|
| (Intercept)           | -4.073e+00 | 7.266e+00 | -0.561  | 0.5751  |
| color.f1              | -2.012e+01 | 3.956e+03 | -0.005  | 0.9959  |
| color.f2              | -4.052e+00 | 8.900e+00 | -0.455  | 0.6489  |
| color.f3              | -1.630e+01 | 9.968e+00 | -1.635  | 0.1020  |
| spine.f1              | -1.715e+01 | 3.956e+03 | -0.004  | 0.9965  |
| spine.f2              | -1.581e+01 | 3.956e+03 | -0.004  | 0.9968  |
| width                 | 1.352e-01  | 2.833e-01 | 0.477   | 0.6332  |
| color.f1:spine.f1     | 3.532e+01  | 5.595e+03 | 0.006   | 0.9950  |
| color.f2:spine.f1     | 1.629e+01  | 3.956e+03 | 0.004   | 0.9967  |
A LR test to test for the significance of all 3-way interactions yields a p-value of

\[
> \text{1-pchisq(fit.cs.cw.sw$deviance-fit.csw$deviance,}
+ \text{fit.cs.cw.sw$df.residual-fit.csw$df.residual)}
\]

[1] 0.3570283

Notice that we still have too many predictors though. Next we test for the removal of each 2-way interaction but to save time we will remove them in one go (simultaneous test).

\[
> \text{fit.c.s.w=update(fit,.~.-weight)}
\]

\[
> \text{summary(fit.c.s.w)}
\]

Call:
\[
\text{glm(formula = y ~ width + color.f + spine.f, family = binomial(link = logit))}
\]

Deviance Residuals:
\[
\begin{array}{cccc}
\text{Min} & \text{1Q} & \text{Median} & \text{3Q} & \text{Max} \\
-2.1206 & -0.9724 & 0.5076 & 0.8750 & 2.1158 \\
\end{array}
\]

Coefficients:
\[
\begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & z value & \text{Pr(>|z|)} \\
(\text{Intercept}) & -12.3908 & 2.8193 & -4.395 & 1.11e-05 *** \\
width & 0.4562 & 0.1078 & 4.233 & 2.31e-05 *** \\
color.f1 & 1.6683 & 0.9328 & 1.788 & 0.07371 . \\
color.f2 & 1.5249 & 0.5672 & 2.689 & 0.00718 ** \\
color.f3 & 1.1443 & 0.5933 & 1.929 & 0.05377 . \\
spine.f1 & -0.3770 & 0.5019 & -0.751 & 0.45254 \\
spine.f2 & -0.4348 & 0.6254 & -0.695 & 0.48687 \\
\end{array}
\]

Signif. codes: 0 \(\text{***}\) 0.001 \(\text{**}\) 0.01 \(\text{*}\) 0.05 \(\text{.}\) 0.1 \(\text{ }\) 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 225.76  on 172  degrees of freedom  
Residual deviance: 186.61  on 166  degrees of freedom  
AIC: 200.61

Number of Fisher Scoring iterations: 4

> 1-pchisq(fit.c.s.w$deviance-fit.cs.cw.sw$deviance,  
+ fit.c.s.w$df.residual-fit.cs.cw.sw$df.residual)

[1] 0.2973877

So we drop all 2-way interactions and next we test each predictor left in the model (1-way).

> drop1(fit.c.s.w,test="Chisq")

Single term deletions

Model:

y ~ width + color.f + spine.f

Df Deviance AIC LRT Pr(>Chi)
<none> 186.61 200.61
width 1 208.83 220.83 22.2219 2.429e-06 ***
color.f 3 194.43 202.43 7.8129 0.05004 .
spine.f 2 187.46 197.46 0.8451 0.65536

---

Signif. codes: 0 $***$ 0.001 $**$ 0.01 $*$ 0.05 $.$ 0.1 $.$ 1

Best to remove spine

> fit.c.w=update(fit.c.s.w,.~.-spine.f) # best to remove spine
> summary(fit.c.w)

Call:
  glm(formula = y ~ width + color.f, family = binomial(link = logit))

Deviance Residuals:

Min 1Q Median 3Q Max
-2.1124 -0.9848 0.5243 0.8513 2.1413

Coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -12.7151   2.7617  -4.604  4.14e-06 ***
width        0.4680    0.1055   4.434   9.26e-06 ***
color.f1    1.3299    0.8525   1.560     0.1188
color.f2    1.4023    0.5484   2.557     0.0106 *
color.f3    1.1061    0.5921   1.868     0.0617 .

---

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: 187.46  on 168  degrees of freedom  
AIC: 197.46

Number of Fisher Scoring iterations: 4
We can also try to remove color but it seems that we should leave color in. From the previous horseshoe crap example (part II) we saw that graphically only the dark color is significant.

```r
> dark=ifelse(color>3,1,0) # to indicate dark by "1"
> fit.d.w=glm(y~dark+width,family=binomial(link=logit))
> summary(fit.d.w)
```

Call:
`glm(formula = y ~ dark + width, family = binomial(link = logit))`

Deviance Residuals:
```
  Min       1Q   Median       3Q      Max
-2.0821  -0.9932   0.5274   0.8606   2.1553
```

Coefficients:
```
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6790   2.6925  -4.338 1.44e-05 ***
dark        -1.3005   0.5259  -2.473  0.0134 *
width        0.4782   0.1041   4.592  4.39e-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: 187.96 on 170 degrees of freedom
AIC: 193.96

Number of Fisher Scoring iterations: 4

- Given width, the estimated odds of satellite for non dark crabs is \(\exp(1.3005) = 3.7\) times the estimated odds for dark crabs.
- Given color (dark or not dark), the estimated odds of satellite are multiplied by \(\exp(0.4782) = 1.6\) for each cm increase in width.

```r
> cor(y,fitted(fit.csw));cor(y,fitted(fit.d.w))
```

[1] 0.5258364

[1] 0.4469688