Department of Statistics

COURSE STATS 330

Model Answers for Assignment 4, 2004

Female horseshoe crabs share a nest with a male partner. In some cases, additional males, called satellites, reside nearby. A biologist is interested in what attributes of the female are associated with the presence of satellites.

Some data on female crabs have been collected and are in an Excel spreadsheet **`crab.csv`** in comma-delimited form. This may be downloaded from the course web page. The data are also reproduced overleaf.

The variables in the data set are

colour: colour of the crab (1=light medium, 2=medium, 3=dark medium, 4=dark),
spine: Spine condition (1=both good, 2=one broken, 3-both broken),
width: Width of the carpace (shell) in cm,
weight: weight of the crab (mg),
satelite: presence of satellite crabs (0=absent, 1=present).

The data (in the form of a comma-delimited Excel file) are available on the course web page under the title crab.csv.

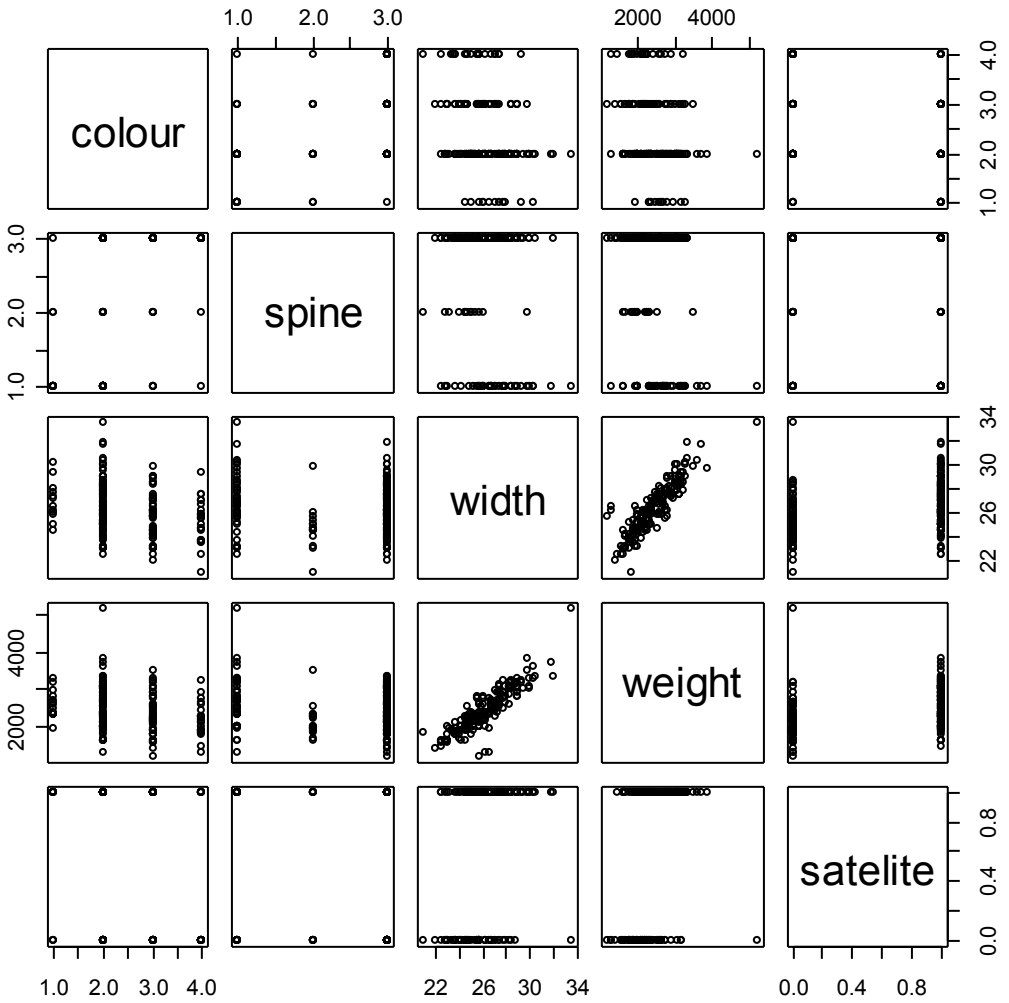
1. Read in the data and check for errors. [5 marks]

Note that the data contains both continuous variables (weight and width) and categorical variables (colour, spine, satellite). The categorical variables are coded numerically, so we will have to convert them in to factors before fitting any models. For the moment, we will keep them as numerical data. We read in the data and do a pairs plot for a first look:

```
temp.df<-read.table(file.choose(), header=T, sep=",")
pairs(temp.df)
```

The pairs plot (shown overleaf) shows up several interesting facts. These are

There is a strong relationship between weight and width, as one might expect.
There seems to be an extreme point in the weight/width plot.
The there seems to be a relationship between weight/width and colour (the lighter the colour, the heavier/wider the crab.
The bigger the crabs, the more likely they are to have a satelite.



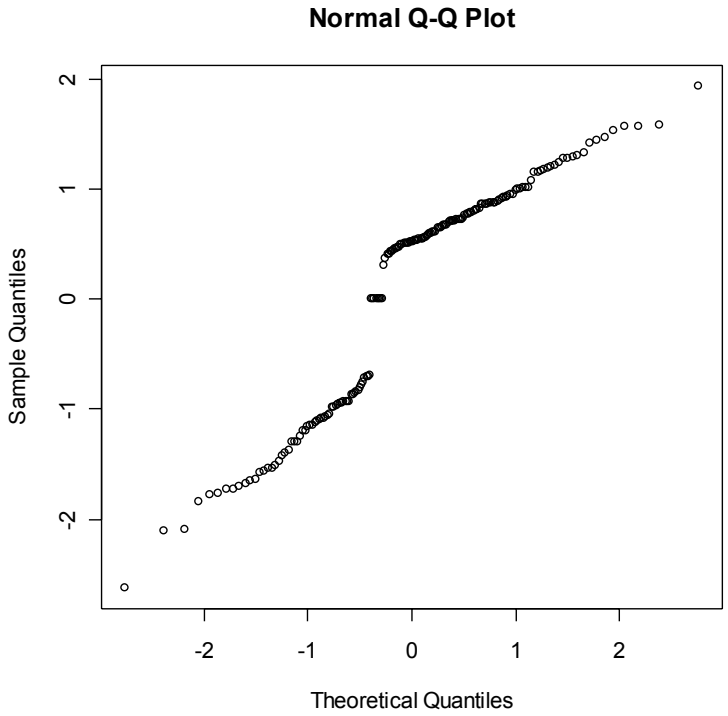
Convert the data frame to factors and fit the model for an outlier check:

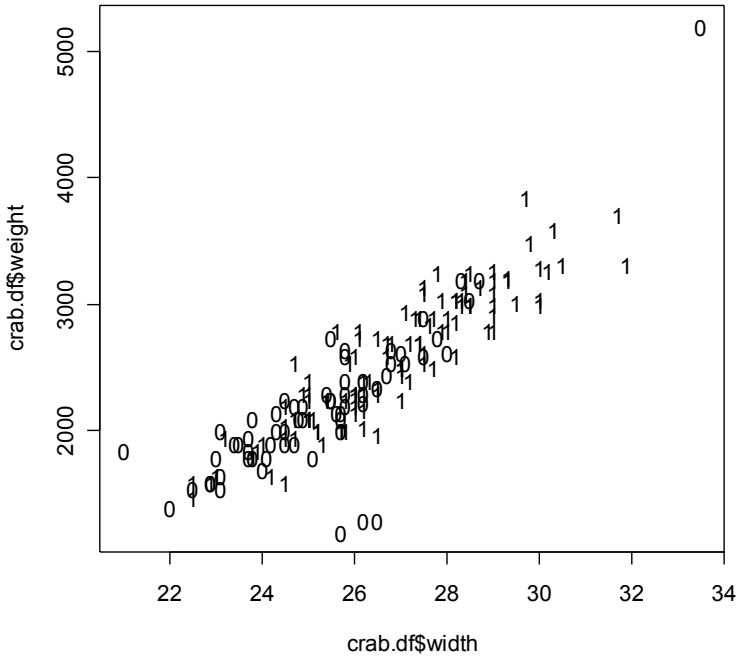
```
> model1<-glm(satellite ~ weight + width + colour*spine,
      family=binomial, data=crab.df)
> qqnorm(residuals(deviance))
```

In the plot shown overleaf, there is one largish deviance residual. Plotting weight versus width, using satellite as the plotting symbol, gives a possible answer:

```
> plot(crab.df$width, crab.df$weight, type="n")
> text(crab.df$width, crab.df$weight, crab.df$satelite)
```

Seems like the extreme point in the upper right of the plot should have been a one rather than a zero! Checking back, we see that this is indeed the case. We change the response value to 1.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2


crab.df$weight
crab.df$width
2000
3000
4000
5000
22
24
26
28
30
32
34

2. Divide the data up into 10 groups according to the weight of the crabs. Draw a suitable plot that illustrates the relationship between the weight of the crab and the probability the crab will have a satellite. Repeat for the widths. What do you conclude? [10 marks: 5 for plots and 5 for conclusions]

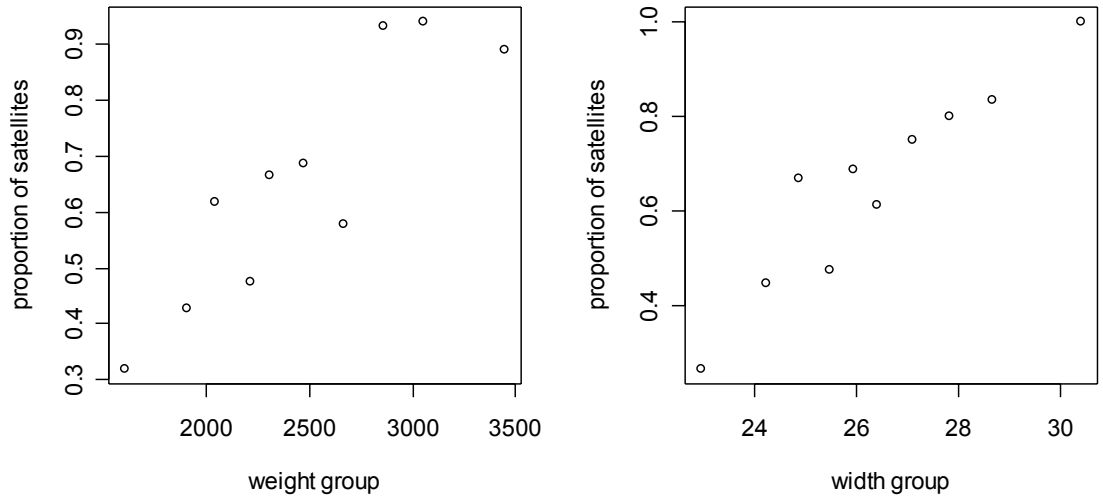
The following code will draw the plots:

```
# First, the weights
par(mfrow=c(1,2))
cut.points<-c(1100, quantile(crab.df$weight,(1:9)/10), 5300)
weight.group<-cut(crab.df$weight,cut.points, labels=F)
av.weight<-tapply(crab.df$weight,weight.group,mean)
weight.prop<-tapply(crab.df$satelite,weight.group,mean)

plot(av.weight, weight.prop, xlab="weight group", ylab="proportion of
satellites")



cut.points<-c(20, quantile(crab.df$width,(1:9)/10), 34)
width.group<-cut(crab.df$width,cut.points, labels=F)
av.width<-tapply(crab.df$width,width.group,mean)
width.prop<-tapply(crab.df$satelite,width.group,mean)
plot(av.width, width.prop, xlab="width group", ylab="proportion of
satellites")
```



We see that there is a strong relationship between size (measured by either width or weight) and the likelihood of having a satellite.

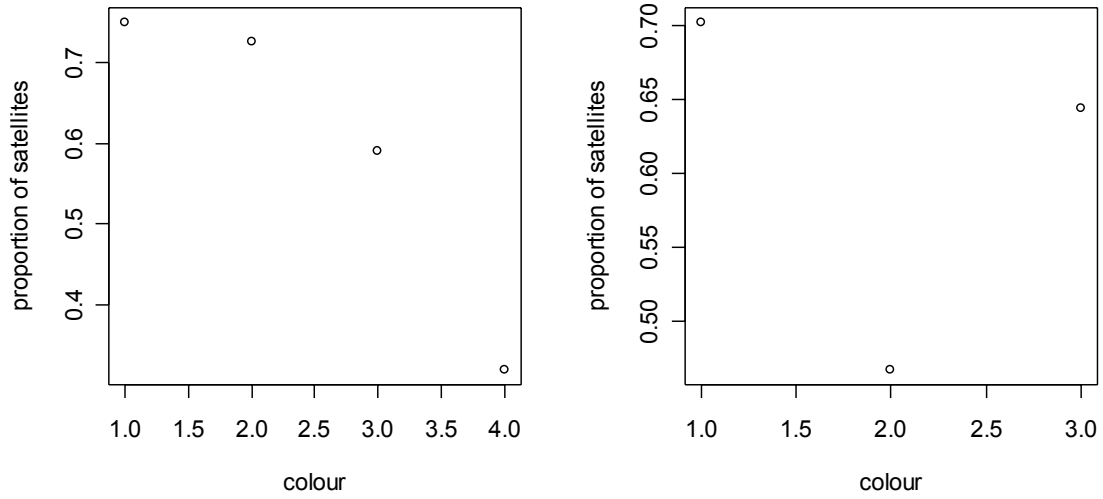
What about colour and spine?

```
par(mfrow=c(1,2))

colour.prop<-tapply(crab.df$satelite,crab.df$colour,mean)
```

```
plot(levels(crab.df$colour) , colour.prop, xlab="colour",
ylab="proportion of satellites")

spine.prop<-tapply(crab.df$satelite,crab.df$spine,mean)
plot(levels(crab.df$spine) , spine.prop, xlab="colour",
ylab="proportion of satellites")
```



There is a strong relationship between colour and proportion of satellites. This may be explained by the relationship between colour and size. Spine is less conclusive.

3. Fit a logistic regression model to these data. Subject your model to the usual diagnostic checks. [10 marks, 5 for fitting and 5 for diagnostic checks]

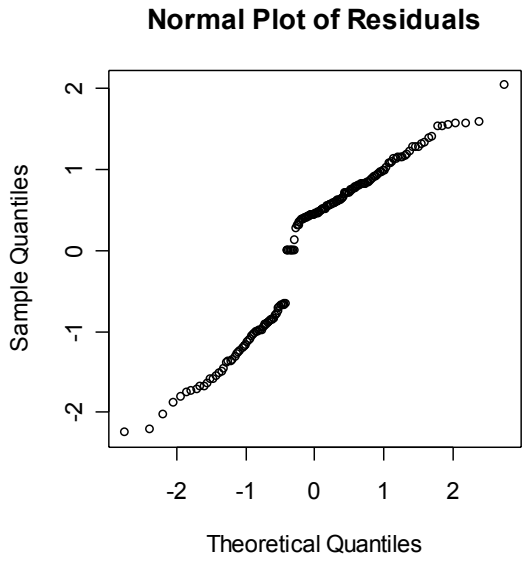
We fit the model satellite ~ width + weight + colour*spine. There is not enough data to fit the more general model satellite ~ width* colour*spine + weight *colour*spine.

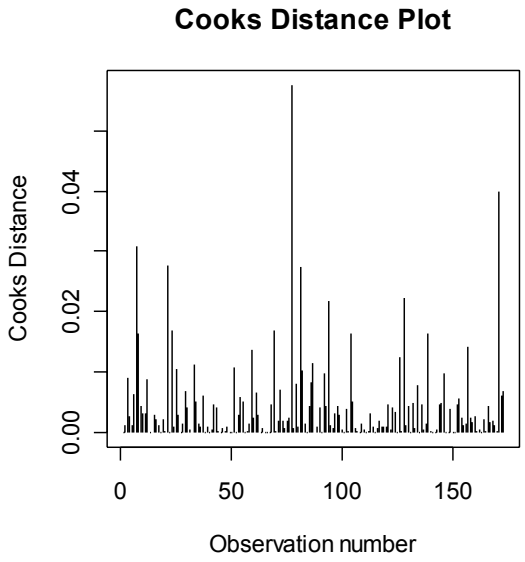
The model is fitted by typing

```
model1<-glm(satelite~weight + width + colour*spine,
       family=binomial, data=crab.df)
```

The plot function will not work on these data for some reason. However, we can draw a normal plot of the residuals and a Cook's distance plot using the code (note use of the function `cooks.distance` and the option `type="h")`

```
par(mfrow=c(1,2))
qqnorm(residuals(model1), main="Normal Plot of Residuals")
plot(1:173, cooks.distance(model1), type="h",
xlab="Observation number", ylab="Cooks Distance",
main="Cooks Distance Plot")
```





These plots don't reveal any problems.

4. What factors are associated with the presence of satellites? [10 marks, must mention that weight, width and colour seem related but spine is not, use a test as well as the graphs in Q2]

Lets do an anova:

```
> anova(model1, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: satelite

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          172    225.759
weight         1   30.021       171    195.737 4.273e-08
width          1    2.845       170    192.892     0.092
colour         3    6.681       167    186.211     0.083
spine          2    1.009       165    185.202     0.604
colour:spine   6    9.608       159    175.594     0.142
```

It seems that once we have added weight, which is highly significant, none of the other variables make a significant contribution to the fit. This is no doubt because of the strong relationships between the explanatory variables. Let's reverse the order of fitting:

```
> model2<-glm(satelite~ spine*colour + width + weight, family=binomial,
data=crab.df)
> anova(model2, test="Chisq")
```

This will allow us to assess the effect of spine and colour.

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: satelite

Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         172    225.759
spine         2    2.526       170    223.232     0.283
colour        3   14.399       167    208.834     0.002
width         1   22.222       166    186.612 2.429e-06
weight        1    1.410       165    185.202     0.235
spine:colour  6    9.608       159    175.594     0.142
```

From this we see that spine seems unrelated to the presence of satellites, but there is a definite relationship between the other variables and the presence of satellites. At this stage we can't say if it is size or colour that is the determining factor, since size and colour are related.

At this point we could try fitting a model

satelite~ colour*width + colour*weight

i.e dropping spine but allowing colour to interact with the continuous variables. However the interactions are not significant in this model.

5. Produce a method for predicting whether a crab will have satellites or not.
[5 marks, must specify a fomula for predicting a crab has satellites, and then predict a "yes" if the probability is > 0.5]

Using the three stepwise methods, we see that the model

satelite~ colour + width

is selected by all three methods, with an AIC of 197.46, so seems a good candidate for a prediction equation.

The model coefficients are

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.38519    2.87346  -3.962 7.43e-05 ***
colour2       0.07242    0.73989   0.098    0.922
colour3      -0.22380    0.77708  -0.288    0.773
colour4      -1.32992    0.85252  -1.560    0.119
width         0.46796    0.10554   4.434 9.26e-06 ***
```

Thus, we predict the probability p of having a satellite by

Log(p/{1-p)) = -11.38519 + 0.46796 width, if the crab is light medium colour,

Log(p/{1-p)) = -11.31278 + 0.46796 width, if the crab is medium colour,
Log(p/{1-p)) = -11.60899 + 0.46796 width, if the crab is dark medium colour,

Log(p/{1-p)) = -12.71511 + 0.46796 width, if the crab is dark colour.

Note that colour is not quite significant (p=0.72) in the above model. The model with width alone has an AIC of 198.45. This model has coefficients

```
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 ***
```

and would predict with

Log(p/{1-p)) = -12.3508 + 0.4972 width .

In view of the usual qualification about p-values after model selection, I would prefer the model including colour.

Finally, to predict if a particular crab has a satellite, use your model to calculate the probability of a satellite for that crab, and predict a "yes" if the probability is more than 0.5.

Mark Scheme

Question 1
Reading in data, noting outlier, correcting outlier: 5 marks

Question 2
Drawing plots: 5 marks. Correctly interpreting plots: 5 marks

Question 3
Fitting model: 5 marks. Drawing and correctly interpreting plots: 5 marks

Question 4
Referring to Q2 graphs and anova for weight etc, use of anova to examine effect of spine: 5 marks. Correct interpretation: 5 marks

Question 5
Selecting a good prediction equation and interpreting correctly : 5 marks
Total marks: 40