Department of Statistics

COURSE STATS 330

Model answers for Assignment 4, 2008

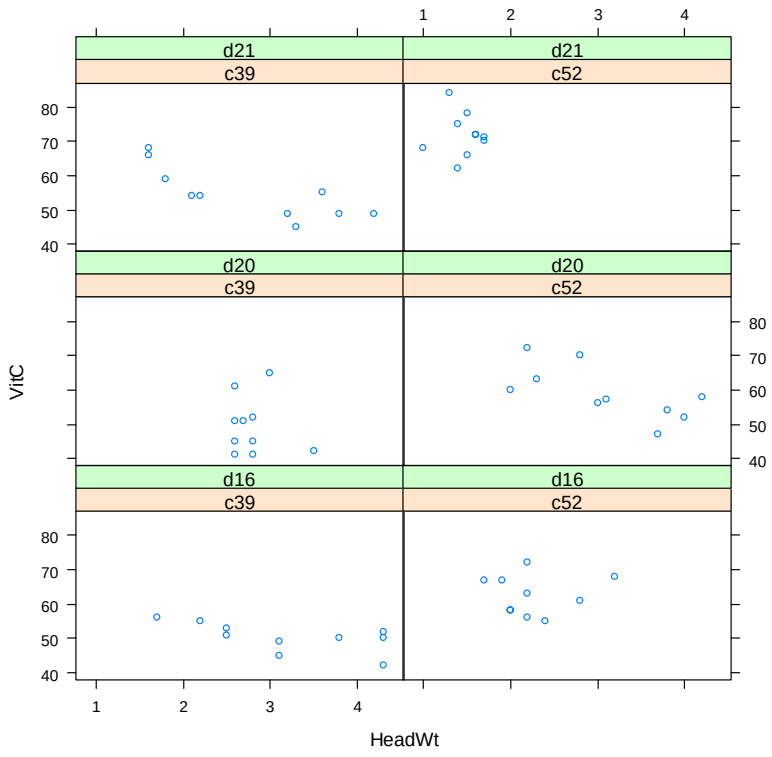
Your task in this assignment is to assess the effect of `Cult, Date` *and* `HeadWt` *on the vitamin content. Specifically,*

1. *Draw a trellis plot and use it to assess graphically which of the factors are related to the response. Report your conclusions.[10 marks]*

The code to draw the plot is

```
library(MASS)
data(cabbages)
library(lattice)
xyplot(VitC~HeadWt|Cult*Date, data=cabbages)
```

This produces



In three of the panels, the range of values of `HeadWt` is too narrow to see much of a pattern

However, in the other three panels, the relationship between `VitC` and `HeadWt` seems linear with approximately the same slopes. This suggests that there is a relationship between `VitC` and `HeadWt`, and that there is not much interaction between `HeadWt, Date` and `Cult`. The intercepts are not the same so some effect of these two factors on the intercepts seems probable.

Note: a plot produced with the plot.design function would have been an alternative, but wouldn't have taken HeadWt into account.

10 marks: 5 marks for a trellis plot
4 marks for saying if the factors are related to the response
1 mark for explaining how you decided if they are related

2. *Fit a suitable model to the data. Are there significant interactions between the explanatory variables? [10 marks]*

Let's fit a full model and then use anova to see if a submodel is OK:

```
> cabbages.lm = lm(VitC~HeadWt*Cult*Date, data=cabbages)
> anova(cabbages.lm)
Analysis of Variance Table

Response: VitC
               Df  Sum Sq Mean Sq F value    Pr(>F)
HeadWt          1 2630.53 2630.53 68.3538 8.704e-11 ***
Cult            1 1145.10 1145.10 29.7550 1.686e-06 ***
Date            2  259.43  129.72  3.3707   0.04268 *
HeadWt:Cult     1    1.53    1.53  0.0398   0.84274
HeadWt:Date     2  130.92   65.46  1.7010   0.19332
Cult:Date       2    1.32    0.66  0.0171   0.98301
HeadWt:Cult:Date  2   24.78   12.39  0.3220   0.72627
Residuals      48 1847.24   38.48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seems like an additive model in the intercepts is OK. i.e. the model

`VitC~HeadWt+Cult+Date`

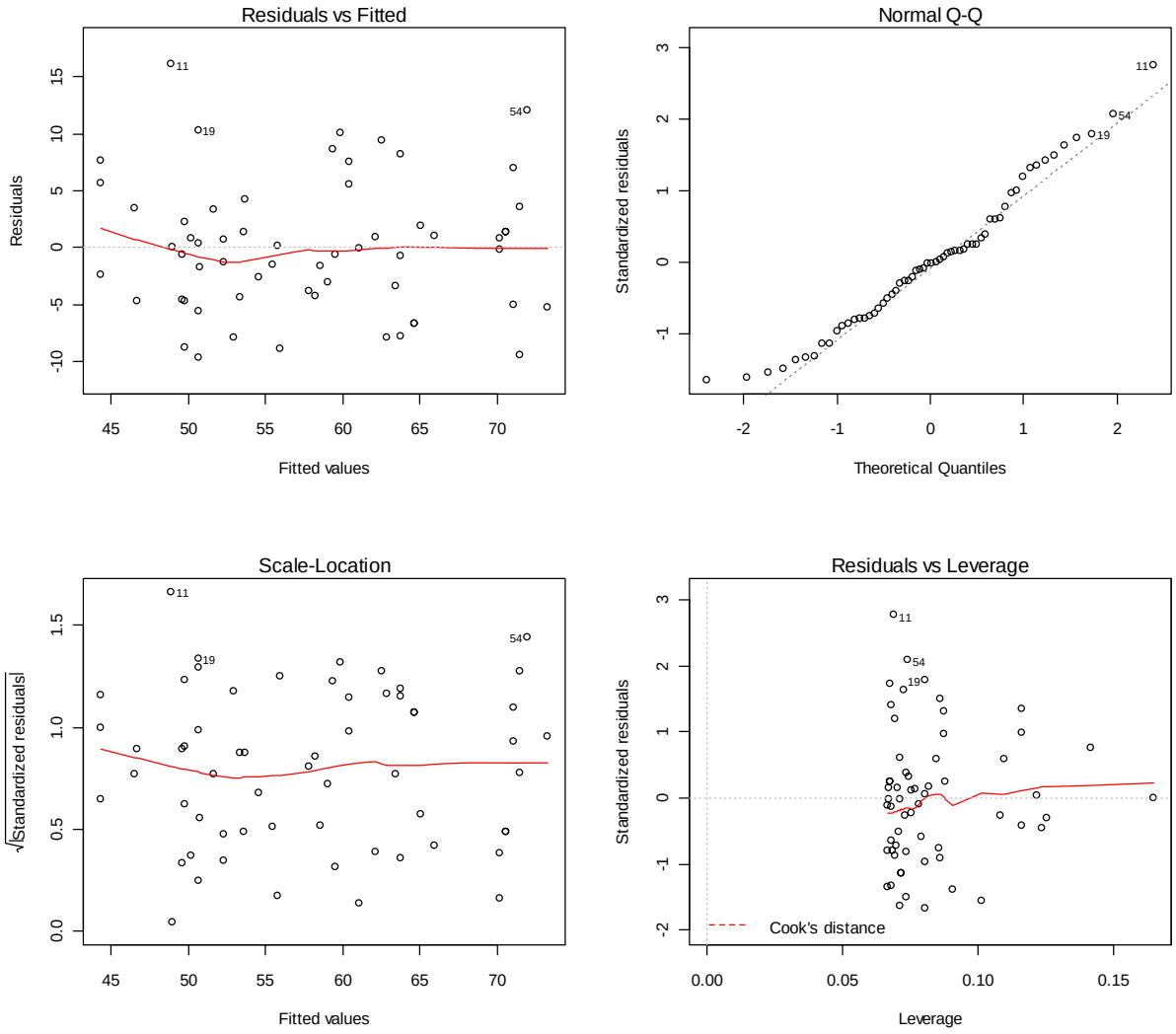
Lets's try stepwise regression:

```
null.lm = lm(VitC~1, data=cabbages)
step(null.lm, formula(cabbages.lm), direction ="both")
```

The output (not shown) also suggests the additive model `VitC~HeadWt+Cult+Date` is OK. Residual plots also confirm this:

```
par(mfrow=c(2,2))
plot(cabbages1.lm)
```

The plot (shown below) indicates no problems. Accordingly we conclude that the explanatory variables do not interact, and we go with the additive model.



10 marks: **3 marks for fitting the full model**
2 marks for using anova to determine significance of interactions
2 marks for doing a stepwise regression
1 mark for conclusion that an additive model is most suitable
2 marks for some diagnostics on final model
[lose 1 mark if remove more than one outlier]

3. *Using the model you fitted in (2), and interpreting the coefficients in the regression summary, discuss how the explanatory variables affect the response. [10 marks]*

The (abbreviated) regression summary is

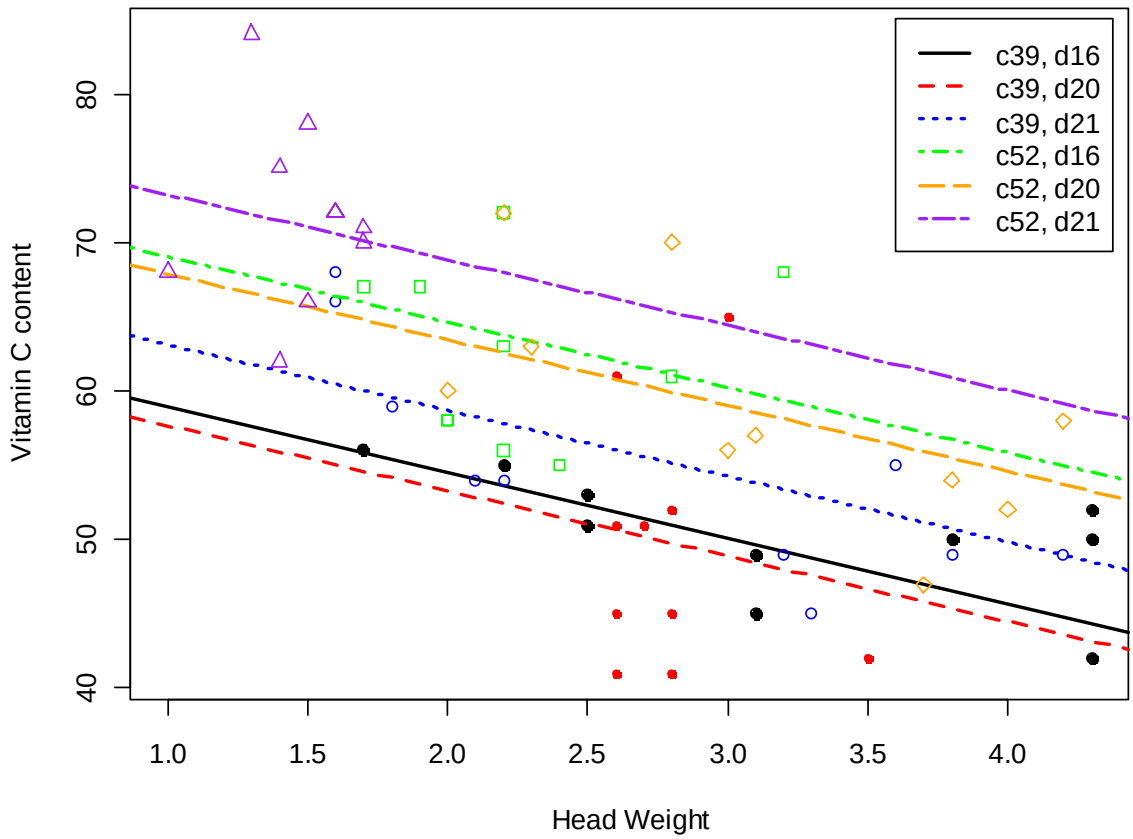
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.334      3.579  17.698  < 2e-16 ***
HeadWt        -4.412      1.062  -4.155 0.000114 ***
Cultc52       10.135      1.695   5.978 1.74e-07 ***
Dated20       -1.213      1.926  -0.630 0.531377
Dated21        4.186      2.018   2.074 0.042747 *
```

From the regression summary, the slope is -4.412. The intercepts are

For C39, d16: 63.334
For C39, d20: 63.334 - 1.213 = 62.121
For C39, d21: 63.334 + 4.186 = 67.520

For C52, d16: 63.334 + 10.135 = 73.469
For C39, d20: 63.334 - 1.213 + 10.135 = 72.256
For C39, d21: 63.334 + 4.186 + 10.135 =77.655

We can summarise the relationships as a series of parallel lines of varying slopes as in the graph below:



The code to produce this graph is shown in the appendix.

We see that C52 generally has higher vitamin C than C39, and that d21 has higher vitamin C than d16 and d20.

10 marks: either 10 marks for a plot plus appropriate comments
or 10 marks for interpreting each of the coefficients
(3 for HeadWt, 3 for c52, 3 for d21, 1 for d20)

4. *For the model you fitted in 2, compute a confidence interval for the mean vitamin C content for each combination of Date and Cult, fixing the value of HeadWt at the mean head weight for the corresponding cell. Hint: use the predict function. [10 marks]*

To answer this part, we can use the predict function. We need to make a new data frame to hold the factor levels and means.

```
attach(cabbages)
```

```
# compute the means
means = as.vector(tapply(HeadWt, list(Cult, Date), mean))
newdata=data.frame(HeadWt = means, Cult = rep(c("c39", "c52"), c(3,3)),
            Date = rep(c("d16", "d20", "d21"), 2))
my.predict = predict(cabbages1.lm, newdata=newdata, interval="confidence")


# Label the rows, note use of paste
rownames(my.predict) = paste(rep(c("c39", "c52"), c(3,3)),rep(c("d16",
"d20", "d21"), 2), sep=",")

> my.predict
             fit      lwr      upr
c39,d16 49.30286 46.16251 52.44322
c39,d20 52.14906 48.35840 55.93972
c39,d21 55.16598 51.93733 58.39462
c52,d16 59.74669 56.28190 63.21148
c52,d20 60.16612 57.03432 63.29793
c52,d21 71.16929 67.97038 74.36820
```

This confirms the information on the graph on page 4: C52 generally has higher mean vitamin C than C39, and that d21 has higher mean vitamin C than d16 and d20.

10 marks: **3 marks for calculating mean HeadWts by any method so long as the method has been explained**
4 marks for using predict with correct syntax
3 marks for obtaining the correct confidence intervals

Question for STATS 763 only.

Write an R program to evaluate the log-likelihood for the CHD data in lecture 20, over a grid of α and β values. Use 50 α-values equally spaced from -5.4, to -5.1, and 50 β-values equally spaced from 0.105 to 0.115. Store the result as a matrix with the rows corresponding to the α–values and the columns corresponding to the β-values, and the matrix entries corresponding to the values of the log-likelihood.

Then plot the log-likelihood as a contour plot. To get a good plot, you may need to manipulate the contour levels (use the levels argument).

Mark the maximum value by eye on the plot, and then verify that the estimates obtained using glm correspond to these maximum values.

The following R program will set up the matrix to hold the log-likelihoods, and calculate the elements. Then we use the R function `contour` to plot the contours.

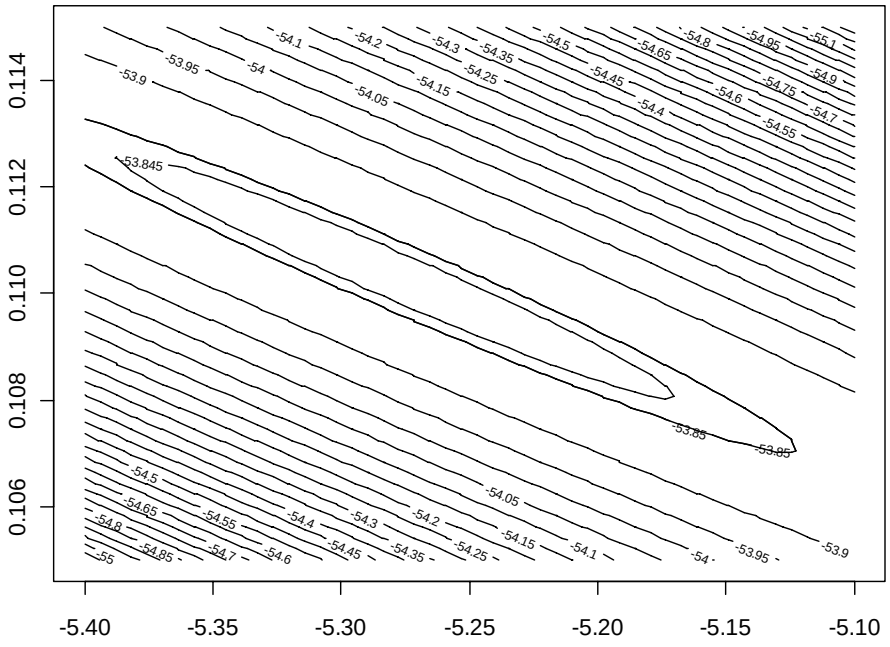
```
chd.df =
read.table("http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/chd.txt",
header=T)
chd.glm = summary(glm(chd~age, data=chd.df, family=binomial))

# define grid of values
n=50
a = seq(-5.4, -5.1, length=n)
b = seq(0.105, 0.115, length=n)
# set up matrix
ll = matrix(0, n,n)
# fill in elements
for(i in 1:n){
for(j in 1:n)ll[i,j]=sum((a[i]+b[j]*chd.df$age)*chd.df$chd -
log(1+exp(a[i]+b[j]*chd.df$age)))
}

Thus, the i-j element of the matrix ll contains the log-likelihood evaluated
at a[i] and b[j].

# draw contours ( set levels after inspecting elements of ll)

contour(a,b,ll,nlevels=30, levels = c(pretty(range(ll),30),-53.85, -53.845))
```

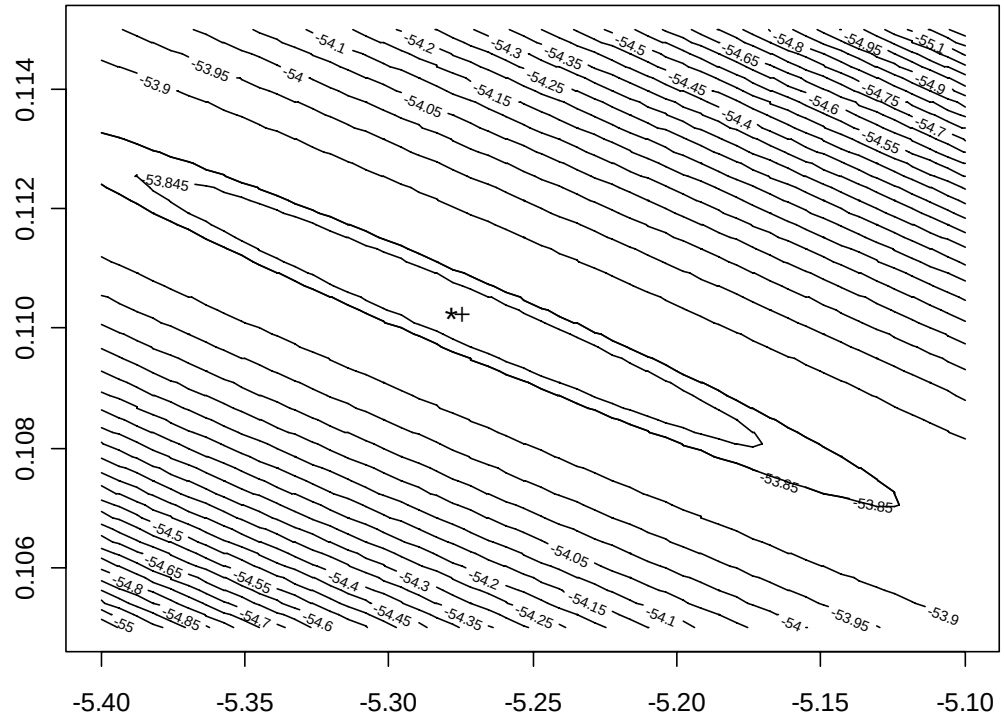


Draw on the maximum (by guessing) by hand: use a "+" . Also mark on the position of the estimates, using a "*":

```
> coefficients(chd.glm)
             Estimate Std. Error   z value     Pr(>|z|)
```

```
(Intercept) -5.2784444 1.13053728 -4.668970 3.027138e-06
age          0.1103208 0.02401837  4.593184 4.365345e-06

points(-5.2784444, 0.1103208  ,pch="*", cex=1.3)
```



Appendix: Code for graph on page 4

```
par(mfrow=c(1,1))
plot(VitC~HeadWt, data = cabbages, type="n", xlab = "Head Weight",
ylab = "Vitamin C content")

mycol = c("black", "red", "blue", "green", "orange","purple")
mycoef = coef(cabbages1.lm)
attach(cabbages)
# c39, d16
use = 1:10
points(HeadWt[use], VitC[use], pch=19, col = mycol[1])
abline(mycoef[1], mycoef[2], lwd = 2, lty=1, col = mycol[1])

# c39, d20
use = use+10
points(HeadWt[use], VitC[use], pch=20, col = mycol[2])
abline(mycoef[1]+ mycoef[4], mycoef[2], lwd = 2,lty=2, col = mycol[2])

# c39, d21
use = use+10
points(HeadWt[use], VitC[use], pch=21, col = mycol[3])
abline(mycoef[1]+ mycoef[5], mycoef[2], lwd = 2,lty=3, col = mycol[3])

# c52, d16
use = use + 10
points(HeadWt[use], VitC[use], pch=22, col = mycol[4])
abline(mycoef[1] + mycoef[3], mycoef[2], lwd = 2,lty=4, col = mycol[4])

# c52, d20
use = use+10
points(HeadWt[use], VitC[use], pch=23, col = mycol[5])
abline(mycoef[1]+ mycoef[4]+ mycoef[3], mycoef[2], lwd = 2,lty=5, col =
mycol[5])


# c52, d21
use = use+10
points(HeadWt[use], VitC[use], pch=24, col = mycol[6])
abline(mycoef[1]+ mycoef[5]+ mycoef[3], mycoef[2],lwd = 2, lty=6, col =
mycol[6])

legend(3.6, 85, legend = c("c39, d16", "c39, d20","c39, d21","c52, d16",
"c52, d20","c52, d21"),
lty=1:6, lwd=2, col = mycol)

detach(cabbages)
```