Department of Statistics

COURSE STATS 330

Model answers for Assignment 4, 2007

1. *Reformat the data into a suitable data frame for the logistic regression. Note that no data is supplied on the web page for this assignment – you have to type in the counts. Print out the data frame. Hint: Use expand.grid for the other factors. [10 marks]*

For a grouped logistic regression, we need a data frame with lines corresponding to the different covariate patterns, plus a variable indicating the number of sucesses, and another variable indicating the number of failures. The following code does the job in for this example:

```
  counts = c(4, 13, 349, 64, 9, 33, 207, 72, 12, 38, 126, 54, 10, 49, 67,
  43, 2, 27, 232, 84, 7, 64, 201, 95, 12, 93, 115, 92, 17, 119, 79, 59,
   8, 47, 166, 91, 6, 74, 120, 110, 17, 148, 92, 100, 6, 198, 42,  73,
   4, 39, 48, 57, 5, 123, 47, 90, 9, 224, 41, 65, 8, 414, 17, 54)
temp.df = data.frame(counts=counts, expand.grid(PE = c("Low","High"),
CP = c("Yes","No"),
IQ = c("L","LM","UM","H"),
SES = c("L","LM","UM","H")))

temp.yes.df = temp.df[temp.df$CP=="Yes",]
temp.no.df = temp.df[temp.df$CP=="No",]
seniors.df = data.frame(temp.yes.df[,c(2,4,5)],
CPyes = temp.yes.df[,1],
CPNo = temp.no.df[,1])
row.names(seniors.df)=1:32

seniors.df
     PE IQ SES CPyes CPNo
1   Low  L   L     4  349
2  High  L   L    13   64
3   Low LM   L     9  207
4  High LM   L    33   72
5   Low UM   L    12  126
6  High UM   L    38   54
7   Low  H   L    10   67
8  High  H   L    49   43
9   Low  L  LM     2  232
10 High  L  LM    27   84
11  Low LM  LM     7  201
12 High LM  LM    64   95
13  Low UM  LM    12  115
14 High UM  LM    93   92
15  Low  H  LM    17   79
16 High  H  LM   119   59
17  Low  L  UM     8  166
18 High  L  UM    47   91
19  Low LM  UM     6  120
20 High LM  UM    74  110
```

```
21  Low UM  UM    17   92
22 High UM  UM   148  100
23  Low  H  UM     6   42
24 High  H  UM   198   73
25  Low  L   H     4   48
26 High  L   H    39   57
27  Low LM   H     5   47
28 High LM   H   123   90
29  Low UM   H     9   41
30 High UM   H   224   65
31  Low  H   H     8   17
32 High  H   H   414   54
```

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

Let's fit a model with no interactions and examine its residual deviance:

```
> seniors.glm = glm(cbind(CPyes,CPNo)~PE + IQ + SES, family=binomial,
data=seniors.df)
> summary(seniors.glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0255     0.1501 -26.819  < 2e-16 ***
PEHigh        2.4554     0.1014  24.215  < 2e-16 ***
IQLM          0.5941     0.1235   4.810 1.51e-06 ***
IQUM          1.3332     0.1194  11.162  < 2e-16 ***
IQH           1.9663     0.1210  16.255  < 2e-16 ***
SESLM         0.3560     0.1232   2.889  0.00386 **
SESUM         0.6624     0.1196   5.537 3.08e-08 ***
SESH          1.4140     0.1210  11.690  < 2e-16 ***
---
Null deviance: 2262.607  on 31  degrees of freedom
Residual deviance:   25.236  on 24  degrees of freedom
AIC: 186.93
```

The p-value for the residual deviance is big:

```
> 1-pchisq(25.236, 24)
[1] 0.3930282
```

so there is no reason to doubt that the model with no interactions fits well.
The residual plots are not too bad, although point 32 is not very well fitted by the model.

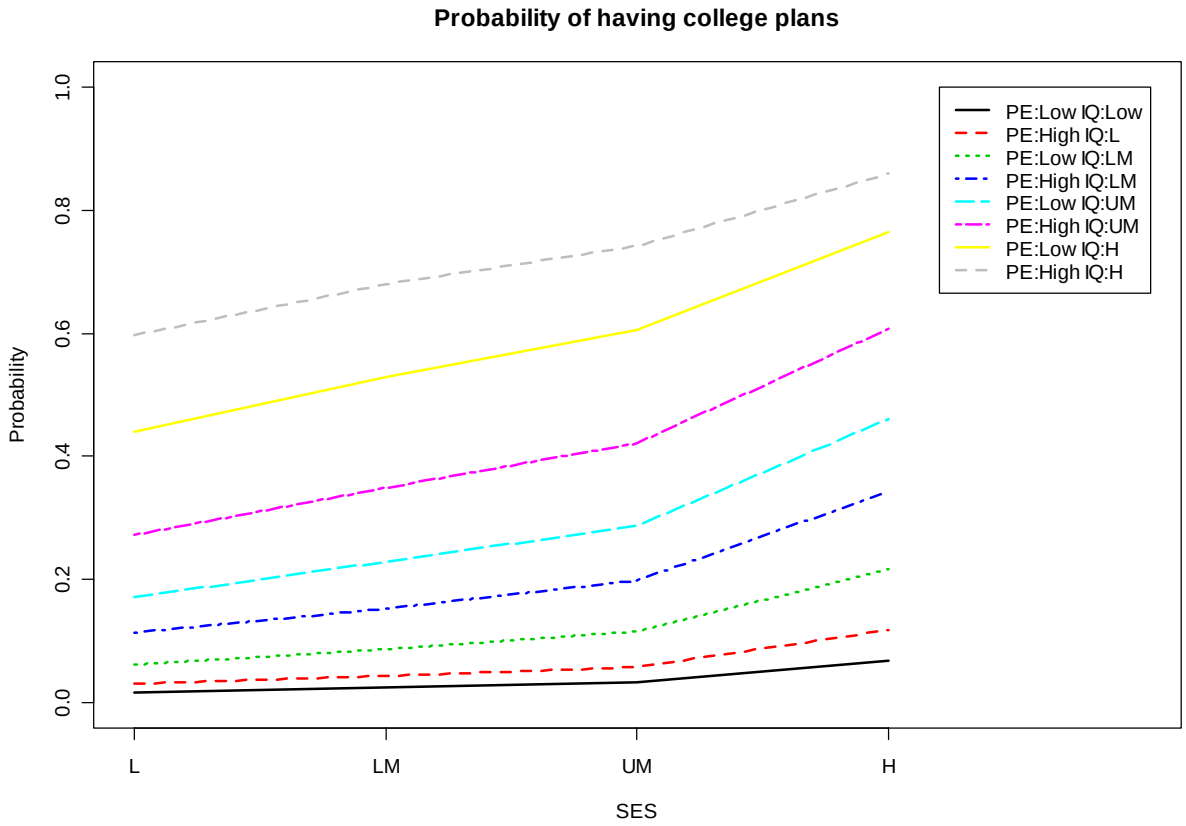
3. *For the model you fitted in 2, calculate the probability a student will have college plans for each combination of the explanatory variables. Plot these probabilities against SES for the eight different combinations of intelligence and parental encouragement. Describe the patterns that you see in the picture. [10 marks]*

The following code does the job – note the use of the function matlines to draw the eight separate lines. This is similar to the rat plot we drew in lecture 2.

```
plotmat = matrix(predict(seniors.glm, type="response"), 4,8,
byrow=T)[,c(1,3,5,7,2,4,6,8)]
```

```
# the last bit is to rearrange the columns
plot(c(1,5), c(0,1), type="n", xlab="SES",ylab="Probability",xaxt="n",
main="Probability of having college plans" )
matlines(plotmat, col=1:8, lty=1:8, lwd=2)
legend(4.2,1,c("PE:Low IQ:Low","PE:High IQ:L","PE:Low IQ:LM","PE:High
IQ:LM","PE:Low IQ:UM","PE:High IQ:UM","PE:Low IQ:H","PE:High IQ:H"),
col=1:8, lty=1:8, lwd=2)
 axis(side=1, at = 1:4, labels = c("L","LM","UM","H"))
```

This produces the picture



From this we can see that the most important factor in having college plans is IQ: the two lowest lines correspond to low IQ and so on. SES and parental encouragement also play a role: for example the chances of a high SES high PE person with lower middle IQ are more than a low SES, low PE, upper middle IQ person.

4. *Calculate a confidence interval for the difference in the log-odds of planning to attend college between students whom have parental encouragement and those who do not. Does this difference depend on IQ and SES? Do you think parental encouragement influences the decision to attend college? [10 marks]*

For the additive (i.e. no interaction model) the effect on the log-odds of changing from low PE to high PE is the same, no matter what the values of SES and IQ. The change is measured by the beta coefficient for PE, which is 2.4554 . We get a confidence interval by typing

```
> confint(seniors.glm)
```

which results in

```
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -4.3251408 -3.7366573
PEHigh       2.2598840  2.6576247
IQLM         0.3534664  0.8378619
IQUM         1.1010826  1.5695291
IQH          1.7314140  2.2057942
SESLM        0.1153969  0.5985049
SESUM        0.4290222  0.8981688
SESH         1.1782668  1.6525734
```

Thus the desired confidence interval is (2.25, 2.66) - there is a significant increase in the log-odds due to parental encouragement. The effect on the probabilities is seen from the graph above, by looking at the pairs of lines (e..g. comparing the lowPE low IQ line with the High PE low IQ line.)