

# Department of Statistics

## COURSE STATS 330

Model answers for Assignment 5, 2006

**Note:** No data was supplied for this assignment. You had to type your own and construct suitable data frames using the **expand.grid** function.

*Q1. Make an appropriate data frame with variables “Father’s occupational status” (**father**, taking values 1:5), “Son’s occupational status” (**son**, taking values 1:5), **country** (taking values “Britain” and “Denmark”) and **count**. [5 marks]*

```
count<-scan( )
50 45 8 18 8 18 17 16 4 2
28 174 84 154 55 24 105 109 59 21
11 78 110 223 96 23 84 289 217 95
14 150 185 714 447 8 49 175 384 198
3 42 72 320 411 6 8 69 201 246
```

```
status.df = data.frame(count, expand.grid(son = factor(1:5),
country = c("Britain", "Denmark"), father = factor(1:5)))
```

*Q2. Suppose we want to ask the question “Is the relationship between Father’s occupational status and Son’s occupational status is different for the two countries?” How could we reformulate this question in terms of testing a sub-model of the saturated model? Recall that relationships in two-way tables are measured in terms of sets of odds ratios. [5 marks]*

If we interpret “relationship” as meaning “the set of odds ratios”, then we need to test if the odds ratios in the British table are the same as those in the Danish table. That is, if the pattern of association in the distribution of fathers status and sons status, given country is the same for both countries. This amounts to testing if the homogeneous association model fits well, or equivalently, if the three-way interactions in the saturated model are zero.

*Q3. Make a 4x4 table of odds ratios for the British data, and another table for the Danish data. Compare the tables. What does this tell you about the question raised in Q2?[10 marks]*

There are lots of ways to do this. We could fit a saturated model to each table, extract the interactions, and then exponentiate the results. Alternatively, we could just work them out directly – this is what we do here.

First, make two matrices containing the appropriate counts:

```
> Both = matrix(count, 5,10,byrow=T)
> British = Both[,1:5]
> Danish = Both[,6:10]
```

Then compute and display the results using the formula

$$OR_{ij} = (x_{ij} x_{11}) / (x_{i1} x_{1j})$$

where  $x_{ij}$  is the count in the  $i,j$  cell. (you may need to look up the function **outer** to see whats happening here)

```
> OR.British = British[-1,-1] *
  British[1,1]/(outer(British[-1,1], British[1,-1]))
> OR.Danish = Danish[-1,-1] *
  Danish[1,1]/(outer(Danish[-1,1], Danish[1,-1]))
> round(OR.British,1)
      [,1] [,2] [,3] [,4]
[1,]  6.9  18.8  15.3  12.3
[2,]  7.9  62.5  56.3  54.5
[3,] 11.9  82.6 141.7 199.6
[4,] 15.6 150.0 296.3 856.2
> round(OR.Danish,1)
      [,1] [,2] [,3] [,4]
[1,]  4.6  5.1 11.1  7.9
[2,]  3.9 14.1 42.5 37.2
[3,]  6.5 24.6 216.0 222.8
[4,]  1.4 12.9 150.8 369.0
```

The OR's are not very similar, so we suspect the British and Danish relationships are different.

*Q4. Fit a saturated model to these data. Use your model to resolve the question posed in Q2. To test if the actual as opposed to the sample OR's are identical, we see if the homogeneous association model fits well.[10 marks]*

```
> anova(glm(count ~ son*father*country, family=poisson,
data=status.df), test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: count
Terms added sequentially (first to last)
```

	Df	Deviance	Resid.	Df	Resid. Dev	P(> Chi )
NULL				49	6789.2	
son	4	2430.0		45	4359.2	0.0
father	4	2324.4		41	2034.8	0.0
country	1	195.3		40	1839.5	2.186e-44
son:father	16	1470.6		24	368.8	1.047e-303
son:country	4	186.3		20	182.5	3.245e-39
father:country	4	138.9		16	43.5	4.787e-29
son:father:country	16	43.5		0	1.865e-14	<b>2.316e-04</b>

The small p-value of **2.316e-04** indicates that the ORs cannot be considered the same.

*Q5. Can you suggest some measures or measures of social mobility based on these tables? Which country has more social mobility?[10 marks]*

A measure of social mobility is the percent of counts in a row that are in the diagonal cell – a high percentage means that most sons kept the status of their fathers.

If we work this out for our data, we get

```
percent.B = numeric(5)
for(i in 1:5) percent.B[i] = British[i,i]/sum(British[i,])
percent.D = numeric(5)
for(i in 1:5) percent.D[i] = Danish[i,i]/sum(Danish[i,])
round(100*cbind(percent.B,percent.D))
```

	percent.B	percent.D
[1,]	39	32
[2,]	35	33
[3,]	21	41
[4,]	47	47
[5,]	48	46

Thus, it seems that the two countries are fairly similar. Britain has less social mobility for sons whose fathers have status 1, and more for fathers for sons whose fathers have status 3.