960:553– Categorical Data Analysis– Fall, 2023

Homework 5 Solutions, 23 Nov

1. Cox and Snell (1980), as example M, present data on size of cauliflowers as a function of nitrogen and potassium levels in four different plots. Cauliflowers are separated into 5 size grades. The file `https://statweb.rutgers.edu/kolassa/Data/cauliflower.dat` has plot number, nitrogen level, and potassium level (coded as ”A” for low and ”B” for high), followed by numbers of cauliflowers at each grade, starting with the best and working towards the worst. Nitrogen is numbered 0, 1, 2, or 3.

a. Model size as a function of nitrogen level, and potassium level. Here and below, use the cumulative logit model. Comment on your results.

The following code will read the data:

```
cf<-read.table("cauliflower.dat")
names(cf)<-c("plot","nit","pot","g0","g1","g2","g3","g4")
# All five grades are represented in a single line.  Spread these out.
bigcauli<-NULL
for(i in 0:4){
  cf$count<-cf[[paste("g",i,sep="")]]
  cf$grade<-i
  bigcauli<-rbind(bigcauli,
    cf[,c("plot","nit","pot","count","grade")])
}
bigcauli$fnit<-factor(bigcauli$nit)
```

Now, fit the model, using polr from package MASS. First, fit using nitrogen and postassium. Nitrogen is coded as an ordered categorical variable. The question doesn't specify that grade should increase or decrease with nitrogen, and so it's best to use it as an unordered categorical explanatory variable. Use it as a factor.

```
library(MASS)
pout1<-polr(factor(grade)~fnit+pot,
  weights=count, data=bigcauli)
summary(pout1)
```

Fit the smaller models:

```
pout2<-polr(factor(grade)~pot,weights=count,
  data=bigcauli)
anova(pout1,pout2)
```

```
pout3<-polr(factor(grade)~fnit,weights=count,
  data=bigcauli)
anova(pout1,pout3)
```

The R output is:

```
Call:
polr(formula = factor(grade) ~ fnit + pot, data = bigcauli, weights = count)
Coefficients:
      Value Std. Error t value
fnit1 -0.8441    0.1941  -4.348
fnit2 -1.4169    0.2002  -7.077
fnit3 -1.5480    0.2016  -7.677
potB   0.2250    0.1387   1.622
Intercepts:
   Value   Std. Error t value
0|1  -6.9454  0.7249   -9.5811
1|2  -3.4943  0.2072  -16.8615
2|3  -0.7023  0.1570   -4.4732
3|4   1.6864  0.1745    9.6663
Residual Deviance: 1668.038
AIC: 1684.038
> pout2<-polr(factor(grade)~pot,weights=count,
+   data=bigcauli)
> anova(pout1,pout2)
Likelihood ratio tests of ordinal regression models
Response: factor(grade)
     Model Resid. df Resid. Dev   Test   Df LR stat.     Pr(Chi)
1       pot      758   1745.019
2 fnit + pot      755   1668.038 1 vs 2    3 76.98121 1.110223e-16
> pout3<-polr(factor(grade)~fnit,weights=count,
+   data=bigcauli)
> anova(pout1,pout3)
Likelihood ratio tests of ordinal regression models
Response: factor(grade)
     Model Resid. df Resid. Dev   Test   Df LR stat.   Pr(Chi)
1      fnit      756   1670.674
2 fnit + pot      755   1668.038 1 vs 2    1 2.636204 0.1044528
```

Apparently potassium is not significant, but nitrogen is. The nitrogen group 2 appears to be most associated with higher grades of cauliflower, with high and low values associated with lower levels.

b. Does plot appear to influence size? Test an appropriate hypothesis. Use the cumulative logit model. Comment on your results.

The following R commands will do the analysis:

```
pout4<-polr(factor(grade)~fnit+pot+factor(plot),
  weights=count,data=bigcauli)
anova(pout1,pout4)
```

R output is:

```
Likelihood ratio tests of ordinal regression models
Response: factor(grade)
                 Model Resid. df Resid. Dev   Test    Df LR stat.
1           fnit + pot     755  1668.038
2 fnit + pot + factor(plot)     752  1635.119 1 vs 2     3 32.91863
     Pr(Chi)
1
2 3.350505e-07
```

Plot is clearly significant.

c. Does the effect of nitrogen depend on plot? Test an appropriate hypothesis, including any variables you used in part (a).

The following R commands to the analysis:

```
pout5<-polr(factor(grade)~fnit+pot+factor(plot)+
  factor(plot)*fnit, weights=count,data=bigcauli)
anova(pout4,pout5)
```

Here are the likelihood ratio test results:

```
Likelihood ratio tests of ordinal regression models
Response: factor(grade)
                                   Model Resid. df Resid. Dev   Test
1                fnit + pot + factor(plot)     752  1635.119
2 fnit + pot + factor(plot) + factor(plot) * fnit     744  1619.071 1 vs 2
   Df LR stat.   Pr(Chi)
1
2    8 16.04772 0.04170211
>
```

The effect of nitrogen depends on plot.

2. Cox and Snell (1981) present, as their example N, ratings of soap pads. Thirty-two raters each rated two soap pads on two successive days, for a total of $32 \times 2 \times 2$ ratings. These soap pads had different properties, which we will ignore for this problem. The file `https://statweb.rutgers.edu/kolassa/Data/soap.dat` contains a cleaned- up version of this data set. The first column of the data set contains a rater number. The second column contains a character string indicating pad characteristis, which as noted above we will ignore. The third and fourth columns represent ratings on days one and two respectively. These ratings are from 1 (worst) to 5 (best). The fifth through eighth columns are the same as the first four columns, for independent observations. In order to obtain independent observations, retain only the first pair of observations for each judge.

a. Make a table of kappa statistics for this data set, with the data grouped three ways: As original, with the most extreme quality categories collapsed (to obtain categories 1 and 2, 3, 4 and 5), and with categories 1,2,3 collapsed, and 4,5 collapsed. Comment on the stability of the kappa statistic.

The R code below does these calculations:

```
soap<-read.table("soap.dat",col.names=c("a","b","day1","day2","e","f","g","h"))
# Create variables to be tabulated for second and third parts of the question.
soap$good1<-(soap$day1>1)+(soap$day1>4)
soap$good2<-(soap$day2>1)+(soap$day2>4)
soap$ok1<-0+(soap$day1>3)
soap$ok2<-0+(soap$day2>3)
# cohen.kappa will have problems if one of the variables is missing a level.
# This occurs in the biggest table, although I think that it doesn't occur in
# the smaller tables.
# alllevs<-sort(unique(c(soap$day1,soap$day2)))
# soap$day1<-factor(soap$day1,levels=alllevs)
# soap$day2<-factor(soap$day2,levels=alllevs)
library(psych)
outtab<-array(NA,c(2,3))
dimnames(outtab)<-list(c("CohenKappa","PolychoricCorr"),paste("Grouping",1:3,sep=""))
outtab[1,1]<-cohen.kappa(tab1<-table(soap$day1,soap$day2))$kappa
outtab[1,2]<-cohen.kappa(tab2<-table(soap$good1,soap$good2))$kappa
outtab[1,3]<-cohen.kappa(tab3<-table(soap$ok1,soap$ok2))$kappa
library(polycor)
outtab[2,1]<-polychor(tab1)
outtab[2,2]<-polychor(tab2)
outtab[2,3]<-polychor(tab3)
print(tab1)
print(outtab)
```

to get

```
               Grouping1 Grouping2 Grouping3
CohenKappa     0.3227513 0.3600000 0.6666667
PolychoricCorr 0.7634048 0.7035254 0.8798005
```

Note that the kappa changes quite a bit moving from the second to the third grouping.

b. Repeat for the polychoric correlation.

See the above results. Polychoric correlation is more stable.