Analysis of the minn38 Dataset Using Log-Linear Models and Proportional Odds Models

Tsu-Hao Fu

I.

1. Analysis

- **Initial model**

The data frame minn38 in MASS gives a dataset with four factors and a numeric column of frequencies. These are the first 6 rows of the data.

```
> head(minn38)
  hs phs fol sex  f
1  L   C  F1   M 87
2  L   C  F2   M 72
3  L   C  F3   M 52
4  L   C  F4   M 88
5  L   C  F5   M 32
6  L   C  F6   M 14
```

I will use surrogate log linear model to proceed the analysis. Our initial model, also our minimal model, may be described as having the conditional probabilities for each of the four phs classes the same for all hs x fol x sex groups. In other words, phs is independent of the other explanatory factors.

```
Call:
glm(formula = f ~ hs * fol * sex + phs, family = poisson, data = minn38)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-12.3618   -2.5793   -0.7388    1.8396   15.0211

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18660.9  on 167  degrees of freedom
Residual deviance:  3010.1  on 123  degrees of freedom
AIC: 4011

Number of Fisher Scoring iterations: 5
```

The high residual deviance clearly indicates that this simple model is inadequate, so the probabilities do appear to vary with the explanatory factors. We now consider adding the interaction terms between response and explanatory factors to the model that allow for some variation of this kind.

- **Stepwise Selection**

After achieving our minimal model, I will choose a candidate model by AIC in a Stepwise Algorithm.

```
mn.step<-step(mn.minimal,scope=list(lower=~.,upper=~.^2))
glm(formula = terms(f ~ hs * fol * sex + phs + fol:phs + hs:phs +
    sex:phs + fol:sex:phs + hs:fol:phs, keep.order = T), family = poisson,
    data = minn38)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.53279  -0.41864   0.00036   0.37007   1.70933

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18660.923  on 167  degrees of freedom
Residual deviance:    52.881  on  42  degrees of freedom
AIC: 1215.8

Number of Fisher Scoring iterations: 4

> 1  - pchisq(deviance(mn.m1),  mn.m1$df.resid)
[1] 0.1211617
```

As we can see the difference between the minimal model and the model by stepwise is that we add five interaction terms phs : (hs+fol+sex), hs:fol:phs and fol:sex:phs. The stepwise procedure greatly reduced the AIC from 4011 to 1215.8. The deviance also indicates that it is satisfactorily fitting model, but we might need check if there are other adjustments to the model are needed.

```
> drop1(mn.m1)
Single term deletions

Model:
f ~ hs * fol * sex + phs + fol:phs + hs:phs + sex:phs + fol:sex:phs +
    hs:fol:phs
           Df Deviance    AIC
<none>          52.881 1215.8
hs:fol:sex  12   78.051 1217.0
fol:sex:phs 18  113.107 1240.0
hs:fol:phs  36  136.355 1227.3
```

Note that the first term here is part of the minimum model and hence it may not be removed. Only terms that contain the response factor, phs, are of any interest to us for this analysis. However, deleting them will not help us

reduce the deviance or AIC. Now we consider adding possible interaction terms.

```
> add1(mn.m1, ~. + hs:sex:phs)
Single term additions

Model:
f ~ hs * fol * sex + phs + fol:phs + hs:phs + sex:phs + fol:sex:phs
    hs:fol:phs
           Df Deviance    AIC
<none>          52.881 1215.8
hs:sex:phs  6   47.745 1222.7
Likelihood ratio test

Model 1: f ~ hs * fol * sex + phs + fol:phs + hs:phs + sex:phs + fol:sex:phs +
    hs:fol:phs
Model 2: f ~ hs * fol * sex + phs + fol:phs + hs:phs + sex:phs + fol:sex:phs +
    hs:fol:phs + hs:sex:phs
  #Df  LogLik Df  Chisq Pr(>Chisq)
1 126 -481.90
2 132 -479.33  6 5.1359     0.5265
```

I added the interaction term hs : sex : phs, so it is my maximum model. However, as it increases the AIC and the likelihood ratio test showed that there is no significant difference between the model by stepwise and the add1 model, I choose not to include it on the grounds of simplicity, although in some circumstances we might view this decision differently. Therefore, my final model will be the model after stepwise selection.

Final model:
f ~ hs * fol * sex + phs + fol:phs + hs:phs + sex:phs + fol:sex:phs + hs:fol:phs

2. Presentation

I present my finial model final fit in the form of estimated cell probabilities.

	hs	fol	sex	prob.C	prob.E	prob.N	prob.O
1	L	F1	F	0.328	0.112	0.054	0.507
2	M	F1	F	0.484	0.079	0.085	0.351
3	U	F1	F	0.689	0.078	0.038	0.195
4	L	F2	F	0.177	0.079	0.083	0.661
5	M	F2	F	0.267	0.118	0.100	0.516
6	U	F2	F	0.422	0.124	0.090	0.365
7	L	F3	F	0.081	0.074	0.052	0.792
8	M	F3	F	0.151	0.131	0.057	0.661
9	U	F3	F	0.231	0.260	0.072	0.436

10	L	F4	F	0.166	0.083	0.061	0.690
11	M	F4	F	0.243	0.119	0.085	0.553
12	U	F4	F	0.408	0.114	0.101	0.377
13	L	F5	F	0.118	0.104	0.033	0.745
14	M	F5	F	0.159	0.137	0.055	0.650
15	U	F5	F	0.320	0.108	0.091	0.481
16	L	F6	F	0.061	0.080	0.012	0.847
17	M	F6	F	0.092	0.112	0.057	0.739
18	U	F6	F	0.225	0.102	0.052	0.621
19	L	F7	F	0.052	0.047	0.020	0.881
20	M	F7	F	0.121	0.088	0.046	0.746
21	U	F7	F	0.210	0.107	0.077	0.607
22	L	F1	M	0.430	0.063	0.009	0.498
23	M	F1	M	0.610	0.043	0.015	0.332
24	U	F1	M	0.788	0.039	0.006	0.167
25	L	F2	M	0.253	0.050	0.025	0.672
26	M	F2	M	0.378	0.073	0.030	0.519
27	U	F2	M	0.559	0.072	0.025	0.344
28	L	F3	M	0.082	0.024	0.018	0.876
29	M	F3	M	0.160	0.045	0.021	0.774
30	U	F3	M	0.282	0.102	0.031	0.586
31	L	F4	M	0.202	0.044	0.022	0.733
32	M	F4	M	0.303	0.065	0.031	0.601
33	U	F4	M	0.501	0.060	0.036	0.402
34	L	F5	M	0.171	0.051	0.011	0.767
35	M	F5	M	0.233	0.068	0.018	0.680
36	U	F5	M	0.444	0.051	0.029	0.476
37	L	F6	M	0.080	0.045	0.007	0.868
38	M	F6	M	0.125	0.065	0.034	0.776
39	U	F6	M	0.291	0.056	0.030	0.623
40	L	F7	M	0.132	0.040	0.015	0.814
41	M	F7	M	0.279	0.067	0.030	0.624
42	U	F7	M	0.431	0.073	0.045	0.452

3. Discussion

The message of the fitted model is now clear. The factor having most effect on the probabilities is high school rank. With an increase in high school rank increasing the probability of enrolling in college, however, middle and low high school rank students have higher probabilities of doing other stuffs. The next important factor is the sex. Males have higher probability to go to

college. Finally, the father's occupational level has a relatively small effect on the post high school status.

II.

1. Analysis

- **Initial model**

Our initial model, also our minimal model, may be described as having the conditional probabilities for each of the twelve phs and hs pairs the same for all fol x sex groups. In other words, phs and hs pair is independent of the other two explanatory factors.

```
glm(formula = f ~ fol * sex + hs + phs, family = poisson, data = minn38)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-12.707   -2.993   -1.203    1.485   18.029

    Null deviance: 18660.9  on 167  degrees of freedom
Residual deviance:  3638.2  on 149  degrees of freedom
AIC: 4587.1
```

The high residual deviance clearly indicates that this simple model is inadequate, so the probabilities do appear to vary with the explanatory factors. We now consider adding the interaction terms between response and explanatory factors to the model that allow for some variation of this kind.

- **Stepwise Selection**

After achieving our minimal model, I will choose a candidate model by AIC in a Stepwise Algorithm.

```
mn.step<-step(mn.minimal,scope=list(lower=~.,upper=~.^2))
glm(formula = f ~ fol + sex + hs + phs + fol:sex + fol:phs +
    hs:phs + sex:hs + sex:phs + fol:hs + fol:sex:phs, family = poisson,
    data = minn38)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0190  -0.5605  -0.0359   0.5308   3.3310


    Null deviance: 18660.92  on 167  degrees of freedom
Residual deviance:   160.17  on  90  degrees of freedom
AIC: 1227.1
```

```
> 1  - pchisq(deviance(mn.m2),  mn.m2$df.resid)
[1] 7.655967e-06
```

As we can see the difference between the minimal model and the model by stepwise is that we add four interaction terms fol:phs, hs:phs, sex:hs, fol:hs and fol:sex:phs . The stepwise procedure greatly reduced the AIC from 4587.1 to1227.1. However, the deviance indicates that it is not a satisfactorily fitting model, so we might need check if there are other adjustments we can do to the model.

```
Single term deletions

Model:
f ~ fol * sex + hs + phs + fol:phs + hs:phs + sex:hs + sex:phs +
    fol:hs + fol:sex:phs
           Df Deviance    AIC
<none>          160.17 1227.1
hs:phs      6  1098.38 2153.3
sex:hs      2   562.68 1625.6
fol:hs     12   195.68 1238.6
fol:sex:phs 18   220.04 1251.0
```

Note that deleting these interaction terms will increase the AIC and the deviance, so we may not remove them. Now consider adding possible interaction terms.

```
Single term additions

Model:
f ~ fol * sex + hs + phs + fol:phs + hs:phs + sex:hs + sex:phs +
    fol:hs + fol:sex:phs
          Df Deviance    AIC
<none>          160.17 1227.1
fol:sex:hs 12   136.35 1227.3
```

```
Likelihood ratio test

Model 1: f ~ fol * sex + hs + phs + fol:phs + hs:phs + sex:hs + sex:phs +
    fol:hs + fol:sex:phs
Model 2: f ~ fol * sex + hs + phs + fol:phs + hs:phs + sex:hs + sex:phs +
    fol:hs + fol:sex:phs + fol:sex:hs
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  78 -535.54
2  90 -523.63 12 23.819    0.02152 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

I added the interaction term fol : sex : hs, so it is my maximum model. As it slightly increases the AIC and the likelihood ratio test showed that there is a significant difference between the model by stepwise and the add1 model, I choose to include it to have a lower deviance model, although in some

circumstances we might view this decision differently. Therefore, my final model will be the model after adding fol : sex : hs.

Final model:
f ~ fol * sex + hs + phs + fol:phs + hs:phs + sex:hs + sex:phs +
fol:hs + fol:sex:phs + fol:sex:hs

2. **Presentation**
   I present my finial model final fit in the form of estimated cell probabilities.

	fol	sex	prob.C-L	prob.C-M	prob.C-U	prob.E-L	prob.E-M	prob.E-U	prob.N-L	prob.N-M	prob.N-U	prob.O-L	prob.O-M	prob.O-U
1	F1	F	0.065	0.187	0.083	0.128	0.145	0.057	0.012	0.211	0.041	0.011	0.048	0.011
2	F2	F	0.004	0.040	0.009	0.001	0.010	0.001	0.043	0.411	0.057	0.063	0.332	0.028
3	F3	F	0.149	0.235	0.008	0.197	0.107	0.013	0.025	0.209	0.008	0.016	0.028	0.006
4	F4	F	0.015	0.056	0.003	0.003	0.009	0.002	0.086	0.366	0.092	0.085	0.173	0.110
5	F5	F	0.362	0.056	0.028	0.280	0.095	0.029	0.048	0.026	0.025	0.017	0.021	0.012
6	F6	F	0.024	0.019	0.009	0.003	0.009	0.004	0.086	0.199	0.164	0.049	0.303	0.131
7	F7	F	0.056	0.228	0.062	0.113	0.258	0.038	0.025	0.094	0.043	0.023	0.049	0.012
8	F1	M	0.011	0.041	0.008	0.005	0.014	0.002	0.109	0.253	0.092	0.165	0.257	0.043
9	F2	M	0.134	0.266	0.008	0.182	0.177	0.021	0.053	0.086	0.006	0.033	0.026	0.006
10	F3	M	0.037	0.048	0.003	0.011	0.009	0.002	0.206	0.187	0.087	0.208	0.111	0.089
11	F4	M	0.351	0.022	0.034	0.280	0.038	0.060	0.110	0.016	0.024	0.040	0.010	0.014
12	F5	M	0.060	0.007	0.012	0.011	0.003	0.005	0.210	0.121	0.175	0.125	0.151	0.120
13	F6	M	0.121	0.096	0.084	0.140	0.114	0.086	0.156	0.061	0.046	0.053	0.027	0.016
14	F7	M	0.017	0.010	0.007	0.007	0.002	0.002	0.312	0.097	0.062	0.378	0.081	0.025

3. **Discussion**
   Our main interest lies in the conditional probabilities of the pair of the two response factors given the two explanatory factors. However, either sex or the father's occupational level has a relatively small effect on the pair of post high school status and high school rank. It is hard to interpret or get some useful knowledge by just observing the cell probabilities.

III.

By taking hs as response and fol and sex as predictors, I fitted two proportional odds models, the interaction model and the additive model, to the data.

```
Call:
polr(formula = hs ~ sex * fol, data = minn38, weights = f, Hess = TRUE)

Coefficients:
               Value Std. Error t value
sexM       -0.501032    0.08785 -5.7031
folF2       0.008808    0.08313  0.1060
folF3      -0.442266    0.07125 -6.2069
folF4      -0.296528    0.07931 -3.7386
folF5      -0.463325    0.09901 -4.6798
folF6      -0.709666    0.09978 -7.1122
folF7      -0.477875    0.10476 -4.5618
sexM:folF2 -0.317580    0.11879 -2.6735
sexM:folF3 -0.130369    0.10421 -1.2510
sexM:folF4 -0.289398    0.11374 -2.5443
sexM:folF5 -0.173647    0.14637 -1.1864
sexM:folF6 -0.022173    0.14738 -0.1505
sexM:folF7 -0.201774    0.15603 -1.2932

Intercepts:
    Value    Std. Error t value
LIM  -1.6777   0.0637   -26.3550
MIU   0.0768   0.0619     1.2393

Residual Deviance: 29901.69
AIC: 29931.69
```

```
Call:
polr(formula = hs ~ sex + fol, data = minn38, weights = f, Hess = TRUE)

Coefficients:
       Value Std. Error t value
sexM  -0.6738    0.03197 -21.078
folF2 -0.1469    0.05940  -2.473
folF3 -0.5109    0.05210  -9.806
folF4 -0.4374    0.05696  -7.679
folF5 -0.5492    0.07298  -7.526
folF6 -0.7294    0.07364  -9.905
folF7 -0.5759    0.07765  -7.416

Intercepts:
    Value    Std. Error t value
LIM  -1.7626   0.0495   -35.6116
MIU  -0.0094   0.0470    -0.2004

Residual Deviance: 29913.68
AIC: 29931.68
```

```
Likelihood ratio test

Model 1: hs ~ sex * fol
Model 2: hs ~ sex + fol
  #Df LogLik Df Chisq Pr(>Chisq)
1  15 -14951
2   9 -14957 -6 11.99     0.0622 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above plots, our initial observation is that while the interaction model offers slight improvements in reducing residual deviance and AIC. Also, the likelihood ratio test showed that there is no significant difference between the two models. Therefore, the additive model was ultimately selected as the preferred proportional odds model due to its greater interpretability and similar performance. To evaluate the adequacy of this chosen model, one can simply review the AIC and residual deviance values provided in the model's summary. The resulting residual deviance is 29913.68 and the AIC is 29931.68, both of which
suggest that the model is insufficient in explaining the variability in high school ranking. Since the two models has no significant difference, we can conclude that sex and fol are independent of each other.