Review for STA 2201: April, 2012

A. Log-linear models and categorical data analysis

Wei Lin found an outstanding online resource on this topic: *Generalized Linear Models* by Germán Rodriguez (see especially Ch.5), and a link is posted on our web page. §5.1 is relevant for HW3, and the more complex models in the later sections will not be covered on the final test.

Also, with the help of Agresti (*Categorical Data Analysis*), §9.5, I finally managed to get on top of the coal-miners problem from HW2.

Table 1: Set 11 from Cox & Snell (1981). Numbers of coalminers responding to breathlessness and wheeze according to age group.

Breathlessness		Yes		No		Total
Wheeze		Yes	No	Yes	No	
	20–24	9	7	95	1841	1952
	25–29	23	9	105	1654	1791
	30–34	54	19	177	1863	2113
	35–39	121	48	257	2357	2783
Age	40–44	169	54	273	1778	2274
Group	45–49	269	88	324	1712	2393
	50–54	404	117	245	1324	2090
	55–59	406	152	225	967	1750
	60–64	372	106	132	526	1136
Total		1827	600	1833	14022	18282

Recall that the problem was to assess whether or not the interaction between breathlessness and wheeze changed with age. We can't do this in a log linear model if age is treated as a factor variable, because the model is saturated.

```
> hw4
   count breath wheeze age
1      9      1      1  -4
2      7      1      0  -4
3     95      0      1  -4
4   1841      0      0  -4
5     23      1      1  -3
6      9      1      0  -3
7    105      0      1  -3
...
```

```
> agresti3 = glm(count ~ breath*wheeze*factor(age),data=hw4,family = poisson)
> anova(agresti3)
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)


                          Df Deviance Resid. Df Resid. Dev
NULL                                         35    25889.5
breath                     1  11026.2        34    14863.3
wheeze                     1   7037.5        33     7825.7
factor(age)                8    886.6        25     6939.1
breath:wheeze              1   4237.1        24     2701.9
breath:factor(age)         8   2342.4        16      359.5
wheeze:factor(age)         8    332.9         8       26.7
breath:wheeze:factor(age)  8     26.7         0        0.0
```

The biggest unsaturated model is `breath*wheeze + breath*factor(age) + wheeze*factor(age)`, which models an interaction between `breath` and `wheeze`, and an interaction between `breath` and `factor(age)`, and an interaction between `wheeze` and `factor(age)`, but doesn't get at the question we are interested in.

Agresti suggests the following:

$$\log \mu_{ijk} = (BW, AB, AW) + z\delta I(i = j = 1),$$

where $z = 1, 2, \ldots, 9$ is a linear function of age. This is shorthand for

$$\log \mu_{ijk} = \mu + \beta_i + \omega_j + (\beta\omega)_{ij} + \alpha_k + (\alpha\beta)_{ik} + (\alpha\omega)_{jk} + z\delta I(i = j = 1).$$

The additional term adds a component δ to μ_{111}, 2δ to μ_{112}, etc, and 9δ to μ_{119}. The μ_{11k} term is the log-odds ratio for the kth table. Below we fit the models without and with this term:

```
> agresti = glm(count ~ breath*wheeze + breath*factor(age) + factor(age)*wheeze, family
> anova(agresti)
Analysis of Deviance Table

Model: poisson, link: log
```

```
Response: count

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev
NULL                                  35    25889.5
breath              1  11026.2        34    14863.3
wheeze              1   7037.5        33     7825.7
factor(age)         8    886.6        25     6939.1
breath:wheeze       1   4237.1        24     2701.9
breath:factor(age)  8   2342.4        16      359.5
wheeze:factor(age)  8    332.9         8       26.7

> rstandard(agresti, type="pearson")
         1          2          3          4          5          6 
 0.7477297 -0.7477297 -0.7477297  0.7477297  2.1993530 -2.1993531 
         7          8          9         10         11         12 
-2.1993491  2.1993491  2.0985325 -2.0985326 -2.0985324  2.0985324 
        13         14         15         16         17         18 
 1.7704822 -1.7704822 -1.7704822  1.7704822  1.1310498 -1.1310498 
        19         20         21         22         23         24 
-1.1310498  1.1310498 -0.4220680  0.4220680  0.4220680 -0.4220680 
        25         26         27         28         29         30 
 0.8142649 -0.8142649 -0.8142649  0.8142649 -3.6491197  3.6491197 
        31         32         33         34         35         36 
 3.6491197 -3.6491197 -1.4428896  1.4428896  1.4428896 -1.4428896 
```

The residual deviance is 26.7 on 8 degrees of freedom, suggesting the model doesn't fit very well. I added the Pearson residuals; you can see that they are identical for each set of four observations, i.e. for each age, so the column of residuals could be added to the data table (see next page).

Table 2: Set 11 from Cox & Snell (1981). Numbers of coalminers responding to breathlessness and wheeze according to age group.

Breathlessness		Yes		No		Total	
Wheeze		Yes	No	Yes	No		Std. Pearson
							Residual (1st cell)
	20–24	9	7	95	1841	1952	0.75
	25–29	23	9	105	1654	1791	2.20
	30–34	54	19	177	1863	2113	2.10
	35–39	121	48	257	2357	2783	1.77
Age	40–44	169	54	273	1778	2274	1.13
Group	45–49	269	88	324	1712	2393	-0.42
	50–54	404	117	245	1324	2090	0.81
	55–59	406	152	225	967	1750	-3.65
	60–64	372	106	132	526	1136	-1.44
Total		1827	600	1833	14022	18282	

We can see that the residuals seem to be decreasing with age, which Agresti uses to suggest his linear model. I created an additional variable `indic` for the linear model, as `rep(1:9, each=4)*breath*wheeze`:

```
> hw4
   count breath wheeze age indic
1      9      1      1  -4     1
2      7      1      0  -4     0
3     95      0      1  -4     0
4   1841      0      0  -4     0
5     23      1      1  -3     2
6      9      1      0  -3     0
...

> agresti2 = glm(count ~ breath*wheeze + breath*factor(age) + factor(age)*wheeze
       + indic , family = poisson, data = hw4)
> anova(agresti2)
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)
```

```
                   Df Deviance Resid. Df Resid. Dev
NULL                                  35    25889.5
breath              1  11026.2        34    14863.3
wheeze              1   7037.5        33     7825.7
factor(age)         8    886.6        25     6939.1
indic               1   5956.1        24      982.9
breath:wheeze       1      6.6        23      976.3
breath:factor(age)  8    618.7        15      357.5
wheeze:factor(age)  8    350.7         7        6.8

> summary(agresti2)

...
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           7.51872    0.02328 322.990  < 2e-16 ***
breath               -5.76300    0.27913 -20.646  < 2e-16 ***
wheeze               -2.97769    0.10395 -28.644  < 2e-16 ***
factor(age)-3        -0.10885    0.03383  -3.217  0.00129 **
factor(age)-2         0.01041    0.03279   0.317  0.75089
factor(age)-1         0.24659    0.03101   7.953 1.82e-15 ***
factor(age)0         -0.03531    0.03308  -1.067  0.28575
factor(age)1         -0.06994    0.03330  -2.100  0.03571 *
factor(age)2         -0.33839    0.03568  -9.484  < 2e-16 ***
factor(age)3         -0.63138    0.03896 -16.205  < 2e-16 ***
factor(age)4         -1.26584    0.04863 -26.030  < 2e-16 ***
indic                -0.13063    0.02949  -4.430 9.43e-06 ***

...
```

and we see that the linear coefficient for age is estimated to be -0.131, with an estimated standard error of 0.029. The residual deviance is 6.80 on 7 degrees of freedom. The Pearson residuals show less association with age, although there might be a suggestion of a nonlinear effect as well. We could try

```
> agresti4 = glm(count ~ breath*wheeze + breath*factor(age) + factor(age)*wheeze
              + indic + I(indic^2) , family = poisson, data = hw4)
```

but the estimated coefficient for the quadratic term is only 0.005 with an estimated standard error of 0.013.

This example is also analyzed in McCullagh & Nelder (*Generalized Linear Models*), but I have been unable to reproduce their results. I think Agresti may have had the same problem, because he simply writes “McCullagh and Nelder (1989, Sec. 6.6) showed other analyses.”

Another approach to the question, used by many in HW2, is to choose one of `breath` or `wheeze` as a response, and treat the other as a covariate, and fit a logistic regression. This will tell us whether or not the effect of `wheeze` on `breath` changes with `age`, but seems a little unsatisfactory in not treating the two responses symmetrically. With categorical data cross-classified by several factors, it is not always clear which category or categories to use as the response(s). Cox & Snell (*Applied Statistics*), Example W, consider cross-classified data with 4 categories; they treat 3 of the categories as responses and one as an explanatory variable. Venables & Ripley (*Modern Applied Statistics with S*), analyze the same data but treating one category as a response and having three explanatory variables.

In the social sciences there is an enormous literature on cross-classified data, choosing response variables, analysing dependencies, and so on; much of the work comes under the general heading of “graphical models”, which doesn’t mean graphics, it means models where graphs are used to understand complex dependencies. There will be a workshop on graphical models at the Fields Institute, April 16-18.

B. Some Review Questions

1. Samples of the same material are sent to four laboratories for chemical analysis as part of a study to determine whether laboratories give the same results. The results for laboratories A–D, with the row means, are:

A	58.7	61.4	70.9	59.1	58.2	59.66
B	62.7	64.5	63.1	59.2	60.3	61.96
C	55.9	56.1	57.3	55.2	58.1	56.52
D	60.7	60.3	60.9	61.4	62.3	61.12

The summary and analysis of variance table are given below, partially completed.

```
> testlm2=lm(y~lab)
Warning message:
In model.matrix.default(mt, mf, contrasts) :
  variable ’lab’ converted to a factor

> summary(testlm2)
Call:
lm(formula = y ~ lab)
Residuals:
   Min     1Q Median     3Q    Max
-2.760 -0.855 -0.320  1.150  2.540
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.6600     0.6556  91.007  < 2e-16 ***
labB          ______     ______   2.481  0.02460 *
labC         -3.1400     0.9271  -3.387  0.00376 **
labD          1.4600     0.9271   1.575  0.13486
...
> anova(testlm2)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
lab        3 85.926 _______  13.329 0.0001282 ***
Residuals __ 34.380 _______
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) Complete the missing entries, indicated by ______ lines.

(b) Write out the linear model used to fit this data, explaining all notation and giving needed assumptions.

(c) Explain how to assess the consistency of the data with the null hypothesis that there is no difference among the laboratories.

2. Assume the following model for a one-way analysis of variance:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}; j = 1, \ldots, n_i; i = 1, \ldots, k, \tag{1}$$

and assume as usual that the ϵ_{ij} are independent with mean 0, constant variance σ^2.

(a) Show that the log-likelihood function for $\theta = (\mu, \alpha_1, \ldots, \alpha_k, \sigma^2)$ depends on the sample $\underline{y}$ through the sufficient statistics $(\bar{y}_{1.}, \ldots, \bar{y}_{k.}, \sum_{ij}(y_{ij} - \bar{y}_{i.})^2)$, and thus the parameterization of the model give in (1) has one redundant component.

(b) Show that under the restriction $\alpha_1 = 0$, the maximum likelihood estimate of θ is

$$\hat{\theta}^{(1)} = (\bar{y}_{1.}, 0, \bar{y}_{2.} - \bar{y}_{1.}, \ldots, \bar{y}_{k.} - \bar{y}_{1.}),$$

and that under the restriction $\sum n_i\alpha_i = 0$, the maximum likelihood estimate of θ is

$$\hat{\theta}^{(2)} = (\bar{y}_{..}, \bar{y}_{1.} - \bar{y}_{..}, \ldots, \bar{y}_{k.} - \bar{y}_{..}),$$

where $\bar{y}_{i.} = (1/n_i)\sum_j y_{ij}$, and $\bar{y}_{..} = (1/N)\sum_{ij} y_{ij}$, $N = \sum_i n_i$.

(c) Show that

$$\tilde{\sigma}^2 = \frac{\sum(y_{ij} - \bar{y}_{i.})^2}{N - K}$$

is an unbiased estimate of σ^2.

(d) Give an expression for the variance of $\bar{y}_{i.} - \bar{y}_{i'.}$, and argue that if we want to minimize this variance for all pairs i, i', for fixed $n = mk$, say, that we should choose all n_i equal (to m).

3. Given a sample consisting of 15 hospitals in the Toronto area and 15 in the Montreal area, each with 10 patients that potentially could have heart attacks in 2009, three treatments (A, B, C) were randomly assigned to this sample and the number of patients encountering heart attacks were observed. Each treatment group consists of 10 hospitals. Also included as a potential explanatory variable is the average age of the patients in each of the three hospitals. It is of interest to test whether the treatment effect varies in the same pattern for hospitals in Toronto and Montreal. Consider a Binomial model with the logistic link function.

   (a) Write down the full model and the reduced model involved in the above test, explicitly specifying the model components, predictors and parameters using suitable dummy variables. Then express the null hypothesis in terms of regression parameters.

   (b) The residual deviances of the full model and the reduced models are estimated as 57 and 75, respectively The average numbers of heart attack patients in the hospitals within each combination of treatment and area are also provided below. Describe how you will perform this test, i.e., specifying the value of the test statistic and its distribution under the null hypothesis.

	Trt. A	Trt. B	Trt. C
Toronto	3.2	2.8	1.9
Montreal	4.5	3.3	0

4. Questions from Davison Ch. 8: Problems 8.6, 8.10, 8.13 (a,b,c)

5. Questions from Davison Ch. 9: Problems 9.3 (c), 9.6

6. Questions from Davison Ch. 10: Exercises 10.3.6, 10.4.3, 10.5.3, 10.6.2, 10.7.1

7. Questions from Davison Ch. 10: Problems 10.2, 10.8, 10.9, 10.12