Stat 539 Homework 5

Kenny Flagg

March 9, 2017

1. *Agresti Exercise 6.1 (p. 223-4). The multivariate generalization of the exponential dispersion family is*

$$f(\mathbf{y}_i;\boldsymbol{\theta}_i,\phi)=\exp\left\{\left[\mathbf{y}_i^T\boldsymbol{\theta}_i-b\left(\boldsymbol{\theta}_i\right)\right]/a(\phi)+c\left(\mathbf{y}_i,\phi\right)\right\},$$

where $\boldsymbol{\theta}_i$ is the natural parameter (a $(c-1)\times 1$ vector).

Show that the multinomial variate $\mathbf{y}=(y_1,\ldots,y_{c-1})^T$ (with $y_j=1$ if outcome j occurred and 0 otherwise) for a single trial with parameters $(\pi_1,\ldots,\pi_{c-1})$ has distribution in the $(c-1)$-parameter exponential dispersion family, with baseline-category logits as natural parameters.

The probability mass function is

$$\begin{aligned}
f(\mathbf{y};\boldsymbol{\theta},\phi) &= \left(\prod_{i=1}^{c-1}\pi_i^{y_i}\right)\left(1-\sum_{j=1}^{c-1}\pi_j\right)^{1-\sum_{i=1}^{c-1}y_i}\\
&= \exp\left\{\sum_{i=1}^{c-1}y_i\log\left(\pi_i\right)+\left(1-\sum_{i=1}^{c-1}y_i\right)\log\left(1-\sum_{j=1}^{c-1}\pi_j\right)\right\}\\
&= \exp\left\{\sum_{i=1}^{c-1}y_i\log\left(\pi_i\right)+\log\left(1-\sum_{j=1}^{c-1}\pi_j\right)-\sum_{i=1}^{c-1}y_i\log\left(1-\sum_{j=1}^{c-1}\pi_j\right)\right\}\\
&= \exp\left\{\sum_{i=1}^{c-1}y_i\left[\log\left(\pi_i\right)-\log\left(1-\sum_{j=1}^{c-1}\pi_j\right)\right]+\log\left(1-\sum_{j=1}^{c-1}\pi_j\right)\right\}\\
&= \exp\left\{\sum_{i=1}^{c-1}y_i\log\left(\frac{\pi_i}{1-\sum_{j=1}^{c-1}\pi_j}\right)+\log\left(1-\sum_{j=1}^{c-1}\pi_j\right)\right\}\\
&= \exp\left\{\sum_{i=1}^{c-1}y_i\log\left(\frac{\pi_i}{1-\sum_{j=1}^{c-1}\pi_j}\right)+\frac{c-1}{c-1}\log\left(1-\sum_{j=1}^{c-1}\pi_j\right)\right.\\
&\qquad\left.-\frac{1}{c-1}\sum_{i=1}^{c-1}\log\left(\pi_i\right)+\frac{1}{c-1}\sum_{i=1}^{c-1}\log\left(\pi_i\right)\right\}\\
&= \exp\left\{\left[\sum_{i=1}^{c-1}y_i\log\left(\frac{\pi_i}{1-\sum_{j=1}^{c-1}\pi_j}\right)-\frac{1}{c-1}\sum_{i=1}^{c-1}\log\left(\frac{\pi_i}{1-\sum_{j=1}^{c-1}\pi_j}\right)\right]+\frac{1}{c-1}\sum_{i=1}^{c-1}\log\left(\pi_i\right)\right\}
\end{aligned}$$

$$= \exp\left\{\left[\mathbf{y}^T\boldsymbol{\theta} - \frac{\mathbf{1}^T\boldsymbol{\theta}}{c-1}\right]/1 + \frac{1}{c-1}\sum_{i=1}^{c-1}\log(\pi_i)\right\}$$

so this is a member of the exponential dispersion family with natural parameter

$$\boldsymbol{\theta} = \left(\log\left(\frac{\pi_1}{1-\sum_{j=1}^{c-1}}\right) \quad \cdots \quad \log\left(\frac{\pi_{c-1}}{1-\sum_{j=1}^{c-1}}\right)\right)^T,$$

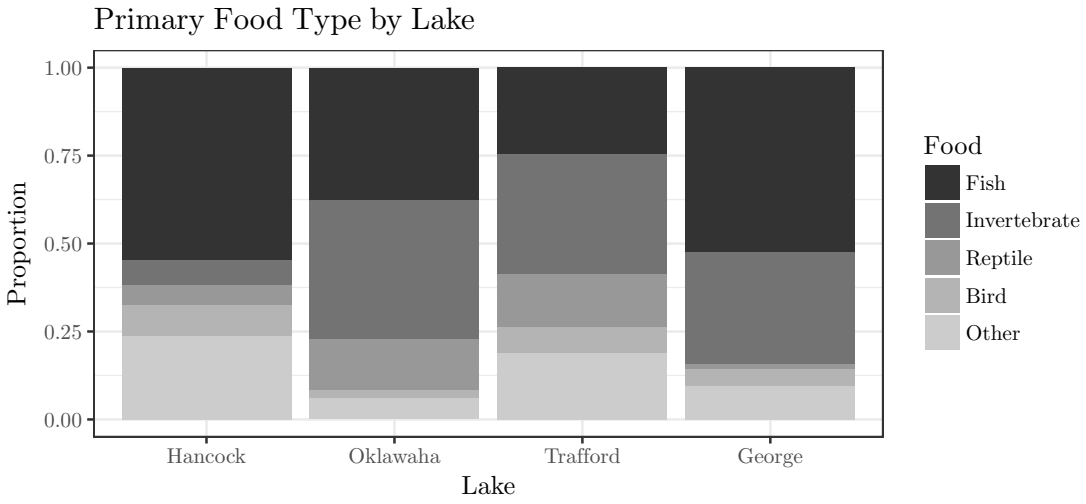
dispersion parameter

$$\phi = \sum_{i=1}^{c-1}\log(\pi_i),$$

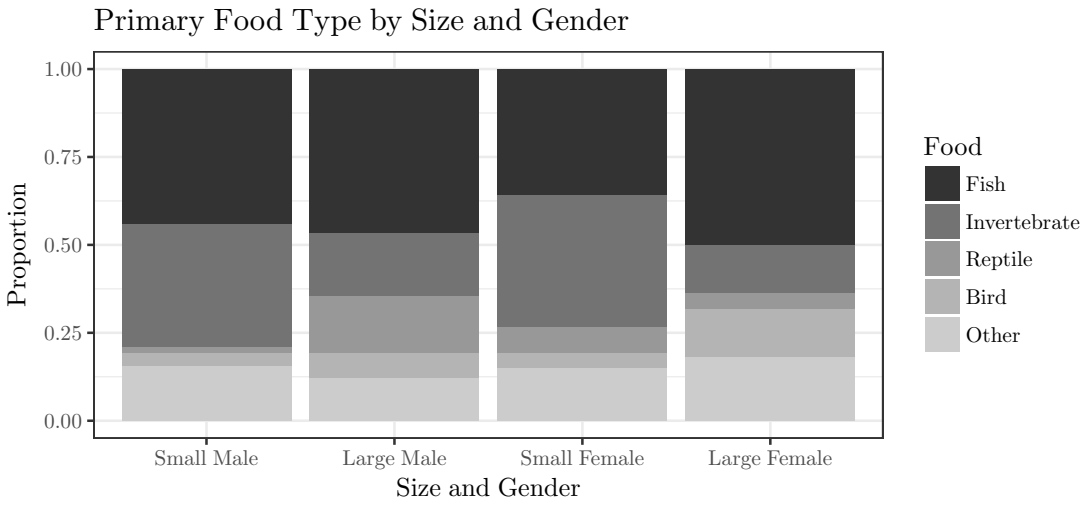
and $a(\phi) = 1$, $b(\boldsymbol{\theta}) = \frac{\mathbf{1}^T\boldsymbol{\theta}}{c-1}$, and $c(\mathbf{y}, \phi) = \frac{\phi}{c-1}$.

2. *The data in* `Alligators.csv` *(on course webpage) is from a study of factors influencing the primary food choice of alligators. The study captured 219 alligators in four Florida lakes. The nominal response variable is the primary food type, in volume, found in an alligator's stomach: F = fish, I = invertebrate, R = reptile, B = bird, O = other. (The other category consisted of amphibian, mammal, plant material, stones or other debris, or no food or dominant type.) The study also classified the alligators according to the lake captured (1 = Hancock, 2 = Oklawaha, 3 = Trafford, 4 = George), gender (1 = male, 2 = female), and size (1 = small (≤2.3 meters long), 2 = large (>2.3 meters long)).*

   (a) *Produce two plots that illuminate the relationship between one or more of the explanatory variables and primary food type. (Since these are all categorical variables, you may need to be creative!) Write a few sentences describing what each plot shows you about these data.*



Across all lakes, fish and invertebrates are the most common primary food types, except at Hancock Lake where fish was most common and "other" was the second most common.



Fish and invertebrates are the most common primary food types among small alligators of both genders. Among large alligators, a smaller proportion had invertebrates as the primary food type; for large males this was made up for by larger proportions with fish or reptiles as the primary food type compared to small males, while large females had bigger proportions with fish or birds as their primary food type than small females did.

(b) *Fit the baseline-category logit model for alligator food choice based on an indicator variable for size ($s = 1$ if small, $s = 0$ if large) and indicator variables for each lake except Lake George (L_H, L_O, L_T, L_G). Use fish as the baseline category. Write the equation of the fitted model. Choose two of the estimated coefficients and write a sentence interpreting each of the chosen coefficients.*

The estimated model is

$$\log\left(\frac{\pi_B}{\pi_F}\right) = -2.093 - 0.631s + 0.695L_H - 0.653L_O + 1.088L_T,$$
$$\log\left(\frac{\pi_I}{\pi_F}\right) = -1.549 + 1.458s - 1.658L_H + 0.937L_O + 1.122L_T,$$
$$\log\left(\frac{\pi_O}{\pi_F}\right) = -1.904 + 0.332s + 0.826L_H + 0.006L_O + 1.516L_T,$$
$$\log\left(\frac{\pi_R}{\pi_F}\right) = -3.315 - 0.351s + 1.243L_H + 2.459L_O + 2.935L_T.$$

$\exp(-0.631) = 0.532$: At any lake, the conditional odds of birds as the primary food type versus fish as the primary food type are estimated to be 46.8% lower for small alligators than for large alligators.

$\exp(0.826) = 2.285$: For alligators of a given size, the conditional odds of "other" as the primary food type versus fish as the primary food type are estimated to be 128.5% higher at Lake Hancock than at Lake George.

(c) *Calculate and interpret a 95% confidence interval for the effect of size on the conditional odds π_I/π_R adjusting for lake, where π_I is the probability an alligator's primary food type is invertebrate, and π_R is the probability an alligator's primary food type is reptile.*

I used the `LinContr.mfit` function to estimate $\exp(\beta_{I,s} - \beta_{R,s})$, where $\beta_{I,s}$ and $\beta_{R,s}$ are the coefficients of the size term in the models for $\text{logit}\,(\pi_I/\pi_F)$ and $\text{logit}\,(\pi_R/\pi_F)$ respectively.

We are 95% confident that, at given lake, the conditional odds of invertebrates as the primary food type versus reptiles as the primary food type are between 87.4% and 1,891% higher for small alligators than for large alligators.

	RRR.est	SE.est	zStat	pVal	CI95.lo	CI95.hi
Test of H_0: 1 * I:s + -1 * R:s = 0	6.107	0.603	3.001	0.003	1.874	19.906

(d) *What is the estimated probability that a small alligator in Lake Oklawaha has invertebrates as the primary food choice?*

I used R to compute

$$\hat{\pi}_I|\{s=1, L_H=0, L_O=1, L_T=0\} = \frac{e^{\eta_I}}{1+e^{\eta_B}+e^{\eta_I}+e^{\eta_O}+e^{\eta_R}} = 0.602$$

(see code submission).

There is an estimated 60.2% chance that a small alligator in Lake Oklawaha has invertebrates as its primary food type.

(e) *An alternative fitting approach for the baseline-category logit model fits binary logistic models separately for the $c-1$ pairings of responses. The estimates have larger standard errors than the maximum likelihood estimates for simultaneous fitting of the $c-1$ logits, but Begg and Gray (1984) showed that the efficiency loss is minor when the response category having highest prevalence is the baseline. Illustrate, by showing that the fit using categories fish and invertebrate alone is*

$$\log\left(\frac{\hat{\pi}_I}{\hat{\pi}_F}\right) = -1.69 + 1.66s - 1.78L_H + 1.05L_O + 1.22L_T$$

with standard error values (0.43, 0.62, 0.49, 0.52) for the effects. Compare with the model in part (b).

The coefficient estimates are somewhat different; between the two models none of the estimates even agree to the first digit after the decimal but none differ by more than about 0.2. The standard errors for the binary logistic regression model are all a little bigger than the standard errors for the multinomial logistic regression model.

	Binary Est.	Binary SE	Multinom. Est.	Multinom. SE
(Intercept)	-1.694	0.450	-1.549	0.425
s	1.660	0.426	1.458	0.396
L_H	-1.779	0.619	-1.658	0.613
L_O	1.052	0.495	0.937	0.472
L_T	1.218	0.517	1.122	0.491

3. *For the alligator food choice data in the previous problem, at one of the lakes the alligators' actual length (rather than a size indicator variable) was measured in meters. Download the data from this lake here:*

   `http://www.stat.ufl.edu/~aa/glm/data/Alligators3.dat`.

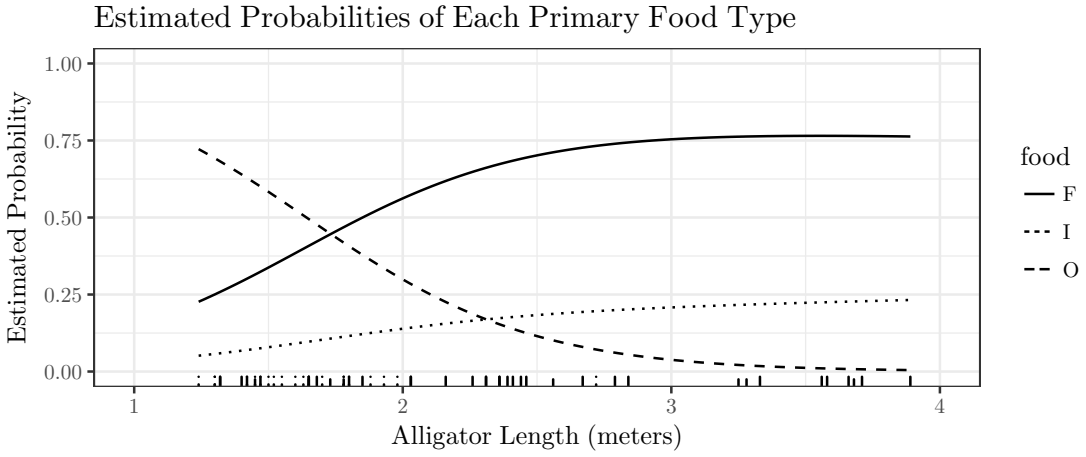
   Read the data into R and fit the baseline-category logit model for alligator food choice based on length. (Note that only fish, invertebrates, and other food choice categories were recorded at this lake).

   (a) *Choose one of the "length" coefficients and write a sentence interpreting this coefficient.*

   For alligators at Lake George of a given length, the conditional odds of invertebrates as the primary food type versus fish as the primary food type are estimated to be 90.5% lower than the conditional odds of invertebrates as the primary food type versus fish as the primary food type for alligators at Lake George that are one meter shorter.

Comparison	Term	RRR	SE	zStat	pVal	CI95.lo	CI95.hi
Level I vs. Level F	(Intercept)	59.128	1.469	2.778	0.005	3.324	1051.720
	length	0.095	0.803	-2.932	0.003	0.020	0.458
Level O vs. Level F	(Intercept)	0.198	1.307	-1.237	0.216	0.015	2.572
	length	1.116	0.517	0.213	0.831	0.405	3.076

   (b) *Produce a single well-labeled plot of the estimated multinomial probabilities (y-axis) versus length (x-axis) by food choice category (line type and/or color with legend). Turn in the R code you used to fit the model and create the plot. Write a few sentences describing the features of the plot.*



   For the shortest alligators, we estimate about a 0.75 probability of invertebrates as the primary food type, but this probability quickly decreases to near 0 as length increases. The shortest alligators have about a 0.25 estimated probability of fish as the primary food and this probability increases with length; fish is estimated to be the most likely primary food for alligators over about 1.75m, and alligators over 3m have an estimated 0.75 probability that of fish as the primary food. The estimated probability of other food types is relatively low for all alligators but increases with length.

4. *Agresti Exercise 6.20 (p. 226-7) The following R output shows output from fitting a cumulative logit model to data from the US 2008 General Social Survey. For subject i let y_i = belief in existence of heaven (1 = yes, 2 = unsure, 3 = no), x_{i1} = gender (1 = female, 0 = male) and x_{i2} = race (1 = black, 0 = white). State the model fitted here, and interpret the race and gender effects. Test goodness-of-fit and construct confidence intervals for the effects.*

```
> cbind(race, gender, y1, y2, y3)

     race gender  y1  y2 y3
[1,]    1      1  88  16  2
[2,]    1      0  54   7  5
[3,]    0      1 397 141 24
[4,]    0      0 235 189 39

> summary(vglm(cbind(y1,y2,y3)~gender+race,family=cumulative(parallel=T)))

              Estimate  Std. Error  z value
(Intercept):1   0.0763      0.0896   0.8515
(Intercept):2   2.3224      0.1352  17.1749
gender          0.7696      0.1225   6.2808
race            1.0165      0.2106   4.8266
Residual deviance: 9.2542 on 4 degrees of freedom
Log-likelihood: -23.3814 on 4 degrees of freedom
```

Models:

The theoretical model is

$$\text{logit}\left(P\left(Y_i \leq j\right)\right) = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2}$$

for $j = 1, 2$, where $x_{i1} = 1$ if respondent i is female and 0 if respondent i is male, and $x_{i2} = 1$ if respondent i is black and 0 if respondent i is white.

The estimated model is

$$\begin{aligned}\text{logit}\left(P\left(Y_i \leq 1\right)\right) &= 0.0763 + 0.7696 x_{i1} + 1.0165 x_{i2},\\ \text{logit}\left(P\left(Y_i \leq 2\right)\right) &= 2.3224 + 0.7696 x_{i1} + 1.0165 x_{i2}.\end{aligned}$$

Interpretations:

$\exp(0.7696) = 215.9$: For individuals of a given race, the odds of being in any specified belief category or below are 115.9% higher for women than for men.

$\exp(1.0165) = 276.3$: For individuals of a given gender, the odds of being in any specified belief category or below are 176.3% higher for black people than for white people.

Confidence Intervals:

$\exp(0.7696 \pm 1.96 \times 0.1225) = (1.698, 2.745)$: For individuals of a given race, we are 95% confident that the true odds of being in any specified belief category or below are between 69.8% and 174.5% higher for women than for men.

$\exp(1.0165 \pm 1.96 \times 0.2106) = (1.829, 4.175)$: For individuals of a given gender, we are 95% confident that the true odds of being in any specified belief category or below are between 82.9% and 317.5% higher for black people than for white people.

Goodness of Fit:

The deviance statistic tests H_0: the model describes the data adequately against H_a: a more general model is needed. Under H_0, the deviance follows a χ^2_4 distribution. The observed deviance is $D = 9.2542$ with a p-value of 0.055, giving moderate evidence against an adequate fit.

5. *Agresti Exercise 6.21 (p. 227). Refer to the previous exercise. Consider the model*

$$\log(\pi_{ij}/\pi_{i3}) = \alpha_j + \beta_j^G x_{i1} + \beta_j^R x_{i2}, \qquad j = 1, 2.$$

(a) *Fit the model and report prediction equations for* $\log(\pi_{i1}/\pi_{i3})$, $\log(\pi_{i2}/\pi_{i3})$, *and* $\log(\pi_{i1}/\pi_{i2})$.

I used the `vglm` function and the `multinomial` family to fit the model (see code submission). The estimated model is

$$\log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) = 1.7943 + 1.0339x_{i1} + 0.6727x_{i2},$$
$$\log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) = 1.5309 + 0.3087x_{i1} - 0.4757x_{i2}.$$

From this, we get

$$\log\left(\frac{\pi_{i1}}{\pi_{i2}}\right) = \log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) - \log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) = 0.2634 + 0.7252x_{i1} + 1.1484x_{i2}.$$

(b) *Using the "yes" and "no" response categories, interpret the conditional gender effect using a 95% confidence interval for an odds ratio.*

For individuals of a given race who responded "yes" or "no" we are 95% confident that the true odds of a yes response are between 69.4% and 366.9% higher for females than for males.

	Estimate	OR	2.5 %	97.5 %
(Intercept):1	1.794	6.015	4.332	8.353
(Intercept):2	1.531	4.622	3.302	6.471
gender:1	1.034	2.812	1.694	4.669
gender:2	0.309	1.362	0.803	2.310
race:1	0.673	1.960	0.875	4.389
race:2	-0.476	0.621	0.256	1.511

(c) *Conduct a likelihood-ratio test of the hypothesis that opinion is independent of gender, given race. Interpret.*

We are testing H_0: $\beta_1^G = \beta_2^G = 0$ versus H_a: $\beta_j^G \neq 0$ for $j = 1$ or $j = 2$. The null model has log-likelihood $\ell_0 = -42.158$ and 4 residual degrees of fredom, and the alternative model has log-likelihood $\ell_a = -21.792$ and 2 residual degrees of fredom. Under H_0, the LRT statistic $-2(\ell_0 - \ell_a)$ has a χ^2 distribution with 2 degrees of freedom. The observed LRT statistic is 40.732 with a p-value of < 0.0001, providing very strong evidence of an association between belief and gender after controlling for race.