HOMEWORK 8 with SOLUTIONS

Directions. This homework pertains to materials on multinomial logistic regression in Lesson 8. The assignment should be typed, with your name on the document, and with properly labeled computer output. I suggest you attach your SAS/R input code at the end of your file clearly indicating the problem it corresponds to. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the HW8 Dropbox on ANGEL.

1. **45 pts** Consider the *alligator food choice* example from the lecture notes and the baseline logit model.

   a. Fit a model to describe effects of *size* and *lake* on primary food choice; i.e., main effects only. You can use the SAS or R code provided in the notes. Report the overall fit of the model, and the prediction equations.

      Solution: G^2 for the overall model fit is 52.48 with p-value 0.147, while X^2 is 15.043 with a p-value of 0.239, indicating that the overall the model fits the data moderately well. Since value/DF is larger than 1 for both statistics, there is evidence of over-dispersion which was adjusted for with the scale=pearson option in SAS.

      Let
      $X_1 = 1$ for Lake Oklawaha, 0 otherwise
      $X_2 = 1$ for Lake Trafford, 0 otherwise
      $X_3 = 1$ for Lake George, 0 otherwise
      $X_4 = 1$ forsize=large,0forsize=small

      And let
      π_1 = prob. of fish,
      π_2 = prob. of birds,
      π_3 = prob. of invertebrates,
      π_4 = prob. of other,
      π_5 = prob. of reptiles,

      Therefore the prediction equations with Fish as the baseline category are:

      Birds vs Fish

      $$\log(\frac{\hat{\pi}_2}{\hat{\pi}_1}) = -2.0286 - 1.3483X_1 + 0.3926X_2 - 0.6951X_3 + 0.6307X_4.$$

      Invertebrates vs Fish

      $$\log(\frac{\hat{\pi}_3}{\hat{\pi}_1}) = -1.7492 + 2.5956X_1 + 2.7803X_2 + 1.6583X_3 - 1.4582X_4.$$

Other vs Fish

$$\log(\frac{\hat{\pi}_4}{\hat{\pi}_1}) = -0.7465 - 0.8205X_1 + 0.6902X_2 - 0.8262X_3 - 0.3316X_4.$$

Reptiles vs Fish

$$\log(\frac{\hat{\pi}_5}{\hat{\pi}_1}) = -2.4230 + 1.2161X_1 + 1.6925X_2 - 1.2422X_3 + 0.3513X_4.$$

b. Using the fit of this model, estimate the probability that the primary food choice is "fish", for each length (that is the size) in Lake Oklawaha. Interpret the effect of *size*.

Solution:

Since $\pi_i = \frac{\exp(\alpha_j+\beta_j x)}{\sum_h \exp(\alpha_h+\beta_h x)}$ for $j = 1, \ldots, J$, and for Lake Oklawaha $X_1 = 1$, $X_2 = 0$ and $X_3 = 0$, so the probability that the primary food choice is fish for size=large in Lake Oklawaha is:

$$1/\{1+e^{-2.0286-1.3483+0.6307}+e^{-1.7492+2.5956-1.4582}+e^{-0.7465-0.8205-0.3316}+e^{-2.4230+1.2161+0.3513}\}$$

$$= 0.458.$$

Similarly, the probability that the primary food choice is fish for size=small in Lake Oklawaha is 0.2582.
The effect of size is highly significant. It indicates that large alligators in Lake Oklawaha are significantly more likely to choose fish as their primary food than the small alligators in Lake Oklawaha.

c. Fit a cumulative logit model with main effects for *size* and *lake* on primary food choice. Let the order in the response be: "Fish" "Invertebrate" "Reptile" "Bird" "Other". Comment on the overall fit of the model and write few sentences on what do you conclude about the food preference. Is the effect of size different from part (b).

Solution: Below is the SAS/R output of the main-effects cumulative logit model is. With $J = 5$ the model has four a_j intercepts.

The G^2 is 105.99 which indicates that the mode does not fit. Also, even size is not significant, its effect estimate is 0.206, indicating that the effect is not different from part (b). Thus, for a given lake, for small aligators the estimated odds that primary food choice is fish is $e^{(-0.205)}$ times the odds for large alligator. In other words, for a given lake, large alligators would be more willing to consum fish for primary food choice.

R code/Output:

```
## Problem 1 (a) and (b)
## Gender={f=female, m=male}
## Size={<2.3=small, >2.3=large}
## Lake={george, hancock,oklawaha,trafford}
## Food is given in 5 column {Fish, Invertebrate,Reptile,Bird,Other}

> library(VGAM)
>
> # Baseline Categories Logit model for nominal response
>
> gator = read.table("gator.txt",header=T)
> gator$Size = factor(gator$Size,levels=levels(gator$Size)[2:1])
> totaln=sum(gator[1:16,5:9]) ## total sample size
> rown=c(1:16)
> for (i in 1:16) {
+  rown[i]=sum(gator[i,5:9])
+  rown
+  }

> ##set the ref levels so that R output matches the SAS code
> ##sets Hancock as the baseline level
> contrasts(gator$Lake)=contr.treatment(levels(gator$Lake),base=2)
> contrasts(gator$Lake)
         george oklawaha trafford
george        1        0        0
hancock       0        0        0
oklawaha      0        1        0
trafford      0        0        1
>
> ##sets "small" as the refernce level
> contrasts(gator$Size)=contr.treatment(levels(gator$Size),base=2)
> contrasts(gator$Size)
     >2.3
>2.3    1
<2.3    0
>
> ##sets male as the reference level
> contrasts(gator$Gender)=contr.treatment(levels(gator$Gender),base=2)
> contrasts(gator$Gender)
  f
f 1
m 0
```

```
>
> ##Fit all basline logit models with Fish as the basline level
> ## By default VGLM will use the last level as the baseline level
 for creating the logits
> ## to set Fish as the baseline level,
 specify it last in vglm call below
>
> # with Lake + Size
> fit4 = vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Lake+Size,
 data=gator, family=multinomial)
> deviance(fit4); df.residual(fit4); pchisq(deviance(fit4),
df.residual(fit4),lower.tail=F) # GOF
[1] 52.47849
[1] 44
[1] 0.1783715
> summary(fit4)

Call:
vglm(formula = cbind(Bird, Invertebrate, Reptile, Other, Fish) ~
    Lake + Size, family = multinomial, data = gator)

Pearson Residuals:
                         Min       1Q    Median      3Q    Max
log(mu[,1]/mu[,5]) -0.98734 -0.50822 -0.114360 0.23730 3.9940
log(mu[,2]/mu[,5]) -1.37163 -0.43794 -0.024796 0.24357 1.9945
log(mu[,3]/mu[,5]) -0.82981 -0.58503 -0.230867 0.22255 2.2373
log(mu[,4]/mu[,5]) -1.58731 -0.31886 -0.015898 1.03299 1.4135

Coefficients:
               Estimate Std. Error  z value
(Intercept):1  -2.02862    0.55805 -3.63518
(Intercept):2  -1.74917    0.53918 -3.24411
(Intercept):3  -2.42302    0.64361 -3.76473
(Intercept):4  -0.74653    0.35198 -2.12092
Lakegeorge:1   -0.69512    0.78126 -0.88974
Lakegeorge:2    1.65836    0.61288  2.70586
Lakegeorge:3   -1.24278    1.18542 -1.04839
Lakegeorge:4   -0.82620    0.55754 -1.48186
Lakeoklawaha:1 -1.34833    1.16334 -1.15901
Lakeoklawaha:2  2.59558    0.65971  3.93443
Lakeoklawaha:3  1.21610    0.78601  1.54717
Lakeoklawaha:4 -0.82054    0.72962 -1.12461
```

```
Laketrafford:1  0.39265    0.78177  0.50226
Laketrafford:2  2.78034    0.67122  4.14220
Laketrafford:3  1.69248    0.78045  2.16860
Laketrafford:4  0.69017    0.55967  1.23317
Size>2.3:1      0.63066    0.64247  0.98161
Size>2.3:2     -1.45820    0.39594 -3.68285
Size>2.3:3      0.35126    0.58003  0.60559
Size>2.3:4     -0.33155    0.44825 -0.73965

Number of linear predictors:  4

Names of linear predictors: log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]),
log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])

Dispersion Parameter for multinomial family:   1

Residual deviance: 52.47849 on 44 degrees of freedom

Log-likelihood: -74.42948 on 44 degrees of freedom

Number of iterations: 5
>
> t(coef(fit4,matrix=T)) # Model coefficients
                   (Intercept) Lakegeorge Lakeoklawaha Laketrafford
log(mu[,1]/mu[,5])  -2.0286189 -0.6951176   -1.3483253    0.3926492
log(mu[,2]/mu[,5])  -1.7491726  1.6583586    2.5955779    2.7803434
log(mu[,3]/mu[,5])  -2.4230189 -1.2427766    1.2160953    1.6924767
log(mu[,4]/mu[,5])  -0.7465252 -0.8261962   -0.8205431    0.6901725

Size>2.3
0.6306597
-1.4582046
0.3512628
-0.3315503
>
> junk = expand.grid(Size=levels(gator$Size),Lake=levels(gator$Lake))
> (predictions = cbind(junk,predict(fit4,type="response",newdata=junk))
  Size     Lake        Bird Invertebrate    Reptile      Other
1 >2.3   george 0.081067366   0.13967784 0.02389871 0.09791362
2 <2.3   george 0.029671169   0.41285579 0.01156654 0.09380245
3 >2.3  hancock 0.140896287   0.02307168 0.07182461 0.19400964
4 <2.3  hancock 0.070401523   0.09309880 0.04745657 0.25373963
```

```
5 >2.3 oklawaha 0.029416087   0.24864518 0.19483742 0.06866280
6 <2.3 oklawaha 0.008817482   0.60189675 0.07722761 0.05387208
7 >2.3 trafford 0.108225068   0.19296122 0.20239954 0.20066164
8 <2.3 trafford 0.035894709   0.51683852 0.08876722 0.17420051
> Fish

0.6574425
0.4521040
0.5701978
 0.5353035
 0.4584385
0.2581861
0.2957525
0.1842990



> # (c)
> # Cumulative Logits/ Proportional Odds
>
> fit4 = vglm(cbind(Fish,Invertebrate,Reptile,Bird,Other)~Lake+Size,
 data=gator, family=cumulative(parallel=T))
> summary(fit4)

Call:
vglm(formula = cbind(Fish, Invertebrate, Reptile, Bird, Other) ~
    Lake + Size, family = cumulative(parallel = T), data = gator)

Pearson Residuals:
                     Min       1Q    Median      3Q    Max
logit(P[Y<=1]) -2.9722 -0.91597 -0.089327 1.22235 2.2025
logit(P[Y<=2]) -2.0910 -0.40097  0.088250 1.02571 1.8123
logit(P[Y<=3]) -2.6552 -1.09132 -0.145618 0.56378 2.8413
logit(P[Y<=4]) -2.6523 -0.59098 -0.153119 0.53399 1.7417

Coefficients:
               Estimate Std. Error  z value
(Intercept):1 -0.306778    0.26552 -1.15537
(Intercept):2  0.916333    0.27274  3.35973
(Intercept):3  1.394521    0.28426  4.90584
(Intercept):4  1.812245    0.29975  6.04595
Lakegeorge     0.406941    0.34634  1.17497
Lakeoklawaha  -0.061391    0.37096 -0.16549
```

```
Laketrafford  -0.711830    0.35767 -1.99020
Size>2.3       0.205918    0.26064  0.79005

Number of linear predictors:  4

Names of linear predictors:
logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), logit(P[Y<=4])

Dispersion Parameter for cumulative family:   1

Residual deviance: 105.9868 on 56 degrees of freedom

Log-likelihood: -101.1836 on 56 degrees of freedom

Number of iterations: 7
>
```

SAS Code/Output proc logist data=gator; freq count; class lake size sex / order=data param=ref ref=first; model food(ref='fish') = lake size sex / link=glogit aggregate scale=none; run;

/*null model*/ proc logist data=gator; freq count; class lake size sex / order=data param=ref ref=first; model food(ref='fish') = / link=glogit aggregate scale=none; run;

/*adjusted null model*/ proc logist data=gator; freq count; class lake size sex / order=data param=ref ref=first; model food(ref='fish') = / link=glogit aggregate=(lake size sex) scale=none; run;

/*null with scale parameter*/

proc logist data=gator; freq count; class lake size sex / order=data param=ref ref=first; model food(ref='fish') = / link=glogit aggregate=(lake size sex) scale=1.148; run;

/*final model with no scale parameter*/

proc logist data=gator; freq count; class lake size sex / order=data param=ref ref=first; model food(ref='fish') = lake size /link=glogit aggregate=(lake size sex) scale=none; run;

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	52.4785	44	1.1927	0.1784
Pearson	58.0140	44	1.3185	0.0765

Number of unique profiles: 16

Parameter		food	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept		bird	1	-2.0286	0.5581	13.2144	0.0003
Intercept		invert	1	-1.7492	0.5392	10.5242	0.0012
Intercept		other	1	-0.7465	0.3520	4.4983	0.0339
Intercept		reptile	1	-2.4230	0.6436	14.1732	0.0002
lake	Oklawaha	bird	1	-1.3483	1.1635	1.3429	0.2465
lake	Oklawaha	invert	1	2.5956	0.6597	15.4797	<.0001
lake	Oklawaha	other	1	-0.8205	0.7296	1.2647	0.2608
lake	Oklawaha	reptile	1	1.2161	0.7860	2.3937	0.1218
lake	Trafford	bird	1	0.3926	0.7818	0.2523	0.6155
lake	Trafford	invert	1	2.7803	0.6712	17.1578	<.0001
lake	Trafford	other	1	0.6902	0.5597	1.5207	0.2175
lake	Trafford	reptile	1	1.6925	0.7804	4.7028	0.0301
lake	George	bird	1	-0.6951	0.7813	0.7916	0.3736
lake	George	invert	1	1.6583	0.6129	7.3216	0.0068
lake	George	other	1	-0.8262	0.5575	2.1959	0.1384
lake	George	reptile	1	-1.2422	1.1852	1.0985	0.2946
size	large	bird	1	0.6307	0.6425	0.9635	0.3263
size	large	invert	1	-1.4582	0.3959	13.5634	0.0002
size	large	other	1	-0.3316	0.4483	0.5471	0.4595
size	large	reptile	1	0.3513	0.5800	0.3668	0.5448

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	105.9868	56	1.8926	<.0001
Pearson	99.3540	56	1.7742	0.0003

Number of unique profiles: 16

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
size	1	0.6240	0.4296
lake	3	10.2381	0.0166

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	fish	1	-0.3070	0.2655	1.3372	0.2475
Intercept	invert	1	0.9161	0.2727	11.2820	0.0008
Intercept	reptile	1	1.3943	0.2842	24.0610	<.0001
Intercept	bird	1	1.8120	0.2997	36.5476	<.0001
size	large	1	0.2059	0.2606	0.6240	0.4296
lake	Oklawaha	1	-0.0611	0.3710	0.0271	0.8692
lake	Trafford	1	-0.7115	0.3577	3.9577	0.0467
lake	George	1	0.4072	0.3463	1.3823	0.2397

2. **30 pts** For a recent General Social Survey, a prediction equation

$$logit[P(Y \leq j)] = \hat{\alpha}_j - 0.54x_1 + 0.60x_2 + 1.19x_3,$$

relates Y = job satisfaction (four ordered categories in increasing order with 1=the least satisfied) to the subject's report of the following variables:

- $x_1 =$ earnings compared with others with similar position (four ordered categories; 1=much less, up to 4=much more),
- $x_2 =$ freedom to make decisions about how to do job (four ordered categories; 1=very true, up to 4=not at all true), and
- $x_3 =$ work environment allows productivity (four ordered categories; 1=strongly agree, up to 4=strongly disagree).

(a) How many different $logit$ models are we simultaneously fitting, and what is different and what is the same across these models with respect to the estimated model coefficients?

Solution: Since the response Y has $J = 4$ categories we are fitting $J - 1 = 3$ models simultaneously. Each cumulative model has its own intercept a_j while the slops are (-0.54, 0.60, and 1.19) the same across these models.

(b) Summarize each partial effect (e.g., β's) by indicating whether subjects tend to be more satisfied, or less satisfied, as (i) x_1, (ii) x_2, (iii) x_3, increases.

Solution:
i) The subjects tend to be more satisfied as x1 increases.
ii) The subjects tend to be less satisfied as x2 increases.

iii) The subjects tend to be less satisfied as x3 increases.

(c) Report the settings for x_1, x_2, and x_3 at which a subject is most likely to have highest job satisfaction.

Solution:
A subject is most likely to have the highest job satisfaction when $x_1 = 4, x_2 = 1$, and $x_3 = 1$. In words, this would mean that the subject thinks that they get paid much more than others in a similar position, that they feel it is very true that they have freedom to make decisions about how to do their job, and that they strongly agree that their work environment allows productivity.

3. **25 pts** For an $I \times J$ contingency table with ordinal Y and scores $(x_i = i)$ for x, consider the following model:

$$logit[P(Y \leq j|X = x_i] = \alpha_j + \beta x_i$$

(a) Show that $logit[P(Y \leq j|X = x_{i+1}] - logit[P(Y \leq j|X = x_i] = \beta$.

Solution: $logitP(Y \leq j/X = x_i) = \alpha_j + \beta x_i$. Hence, for rows $x_i = i$ and $x_{i+1} = i + 1$ the difference in logits is:

$$logitP(Y \leq j/X = x_{i+1}) - logitP(Y \leq j/X = x_i) = \alpha_j + \beta(i+1) - \alpha_j - \beta i = \beta$$

(b) Consider a 2×2 table. Show that the difference in log its from part (a) is a log cumulative odds ratio for the 2×2 table consisting of rows i and $i + 1$ and the binary response having cut point following category j. We will learn later that then this model is also a log-linear model called the *uniform association model*. (Hint: see online notes Sec 8.4, page 182 of Agresti (2007) or Agresti(2013) pg. 302.)

Solution: For every category j and for rows i and $i + 1$

$$logitP(Y \leq j/X = x_{i+1}) - logitP(Y \leq j/X = x_i) = \beta \Rightarrow$$

$$log\frac{P(Y \leq j/X = x_{i+1})}{1 - P(Y \leq j/X = x_{i+1})} - log(\frac{P(Y \leq j/X = x_i)}{1 - P(Y \leq j/X = x_i)}) = \beta \text{ for } j = 1, ..., J-1 \Rightarrow$$

$$log\frac{\frac{P(Y \leq j/X = x_{i+1})}{1-P(Y \leq j/X = x_{i+1})}}{\frac{P(Y \leq j/X = x_i)}{1-P(Y \leq j/X = x_i)}} = \beta \text{ for } j = 1, ..., J-1 \Rightarrow$$

which is the log odds ratio for a 2x2 contingency table with rows i and $i+1$ for a binary response Y with success $Y \leq j$ and failure $Y > j$ and is the same for each j.