STAT436

<u>Directions</u>: All students are to answer both exercises below showing all relevant work. As always, clearly give all needed assumptions and models (give all relevant details); conclusions and justifications are to be given in clear detailed English. Please type up your solutions or write very neatly.

- 1. The EAGLE dataset from problem 7.3 on pp. 236-7 of the Venables & Ripley 1999 text (originally from Knight and Skagen 1988) is analyzed in the attached Appendix. The data relate to the foraging behavior of wintering bald eagles in Washington State, and concern 160 attempts by one (so-called "pirating") Bald Eagle to steal a chum salmon from another (so-called "feeding") Bald Eagle. In the computer programs, the variable "Pirate\_size" represents the size of the pirating eagle, the variable "Pirate\_age" represents the age of the pirating eagle, and the variable "Feeder\_size" represents the size of the feeding eagle, and corresponding dummy variables are given in the table below.
  - (a) Identify the model being fit in the following analysis, clearly defining all terms (variables and parameters). What is a "success" and what are the needed assumptions here?
  - (b) Using the SAS or Minitab computer output below, report on the factors that explain the success of the pirating attempt.
  - (c) Give the predictive formula for the probability of success ( $\pi$ ). [Recall that a "predictive formula" means that we substitute the parameter estimates in place of the parameters.] Write the predictive formula (equation) in terms of  $\hat{\pi}$ .
  - (d) Using the predictive formula, predict the success probability under each of the eight conditions (given on the eight rows of the following data table).
  - (e) Using the computer output, give and clearly interpret the respective odds ratios.
- 2. Stokes, Davis & Koch (*Categorical Data Analysis using the SAS System*, 2<sup>nd</sup> Edition) discuss a study on coronary artery (CA) disease, in which the response is "CA = 1" if CA disease is present and the response is "CA = 0" if it is absent. In this experiment, possible explanatory variables are:
  - AGE (treated as a continuous variable)
  - SEX (which takes the value of 0 for females and 1 for males)
  - ECG (an ordinal variable, with values of 0, 1 and 2, where ECG = 0 is scored if the corresponding ST segment depression is less than 0.1, ECG = 1 is scored if the corresponding ST segment depression lies between 0.1 and 0.2, and ECG = 2 is scored if the corresponding ST segment depression is greater than 0.2)

A logistic model was fit to these data (predicting whether the disease is present – which is termed a "success") including all interaction terms. Since all interaction terms were non-significant, they were dropped, giving the "main-effects logistic regression" reported in the Appendix.

- (a) Clearly write down the model being fit here and give all assumptions/requirements.
- (b) Use the results to comment on the fit of the model (i.e., lack of fit)
- (c) Identify which effects are significant here for predicting CA disease. For example, are males or females more likely to develop CA disease? Why? Is the difference significant? Answer similar questions for the AGE and ECG variables.
- (d) Give and clearly interpret each of the odds ratios.

# Homework 7 Appendix

## Data for First Exercise

Pirate_size	Pirate_age	Feeder_size	Dummy_	Dummy_	Dummy_	Success	Attempt	Success
			pirate_large	pirate_adult	Feeder_small			Rate
Large	Adult	Large	1	1	0	17	24	0.708333
Large	Adult	Small	1	1	1	29	29	1.000000
Large	Immature	Large	1	0	0	17	27	0.629629
Large	Immature	Small	1	0	1	20	20	1.000000
Small	Adult	Large	0	1	0	1	12	0.083333
Small	Adult	Small	0	1	1	15	16	0.937500
Small	Immature	Large	0	0	0	0	28	0.000000
Small	Immature	Small	0	0	1	1	4	0.250000

## **Minitab Output for First Exercise**

Binary Logistic Regression: success, attempt versus dummy\_pirate\_large, dummy\_pirate\_adult, dummy\_feeded\_small Link Function: Logit Response Information Variable Value Count success Event 100 Non-event 60 attempt Total 160 Logistic Regression Table Predictor SE Coef P Odds Ratio Coef Z 1.06584 -4.05 0.000 -4.31267 Constant dummy\_pirate\_large 4.55696 1.05440 4.32 0.000 95.29 dummy pirate adult 1.09730 0.546464 2.01 0.045 3.00 dummy\_feeder\_small 4.93314 1.11923 4.41 0.000 138.81 Predictor Upper Constant dummy pirate large 752.60 dummy pirate adult 8.74 dummy feeder small 1244.87 Log-Likelihood = -45.195Test that all slopes are zero: G = 121.311, DF = 3, P-Value = 0.000 Goodness-of-Fit Tests Method Chi-Square DF Ρ Pearson 6.56679 4 0.161 Deviance 6.95626 4 0.138 4 0.306 Hosmer-Lemeshow 4.82595 Measures of Association: (Between the Response Variable and Predicted Probabilities) Number Percent Summary Measures Pairs Concordant 5449 90.8 Somers' D 0.87 Discordant 233 3.9 Goodman-Kruskal Gamma 0.92 Ties 318 5.3 Kendall's Tau-a 0.41 100.0 Total 6000

#### SAS Program and Output for First Exercise

```
The LOGISTIC Procedure
data eagles;
                                                                  Model Information
  do pirsize='Large','Small';
                                                     Data Set
                                                                                   WORK.EAGLES
    pirlarge=(pirsize='Large');
                                                     Response Variable (Events)
  do pirage='Adult
                         ','Immature';
                                                                                   success
                                                     Response Variable (Trials)
                                                                                   attempt
    piradult=(pirage='Adult
                                     ');
                                                     Model
                                                                                   binary logit
  do fesize='Large','Small';
                                                     Optimization Technique
                                                                                   Fisher's scoring
     fesmall=(fesize='Small');
                                                     Number of Observations Read
                                                                                          8
     input success attempt @@;
                                                     Number of Observations Used
                                                                                          8
  output; end; end; end;
                                                     Sum of Frequencies Read
                                                                                        160
  datalines;
                                                     Sum of Frequencies Used
                                                                                        160
17 24 29 29 17 27 20 20
                                                                    Response Profile
1 12 15 16 0 28 1 4
                                                     Ordered
                                                                 Binarv
                                                                                 Total
                                                       Value
                                                                 Outcome
                                                                             Frequency
proc logistic;
                                                                 Event
                                                                                   100
  model success/attempt=pirlarge
                                                           1
                                                           2
                                                                 Nonevent
                                                                                    60
            piradult fesmall;
                                                    Model Convergence Status
run;
                                            Convergence criterion (GCONV=1E-8) satisfied.
                                                             Model Fit Statistics
                                                                             Intercept
                                                               Intercept
                                                                                   and
                                                 Criterion
                                                                             Covariates
                                                                    Only
                                                 AIC
                                                                 213.700
                                                                                98.389
                                                 SC
                                                                 216.775
                                                                               110.690
                                                 -2 Log L
                                                                 211.700
                                                                                90.389
                                                  Testing Global Null Hypothesis: BETA=0
                                            Test
                                                                                 DF
                                                                 Chi-Square
                                                                                        Pr > ChiSq
                                            Likelihood Ratio
                                                                   121.3111
                                                                                  3
                                                                                            <.0001
                                            Score
                                                                    89.8664
                                                                                  3
                                                                                            <.0001
                                            Wald
                                                                   24.9749
                                                                                  3
                                                                                            <.0001
                                                    Analysis of Maximum Likelihood Estimates
                                                                       Standard
                                                                                       Wald
                                            Parameter
                                                        DF
                                                             Estimate
                                                                         Error
                                                                                 Chi-Square
                                                                                              Pr > ChiSq
                                            Intercept
                                                              -4.3125
                                                                         1.0658
                                                                                                  <.0001
                                                         1
                                                                                    16.3712
                                            pirlarge
                                                         1
                                                               4.5568
                                                                         1.0544
                                                                                     18.6773
                                                                                                  <.0001
                                            piradult
                                                         1
                                                               1.0973
                                                                        0.5465
                                                                                     4.0320
                                                                                                  0.0446
                                            fesmall
                                                         1
                                                               4.9330
                                                                        1.1192
                                                                                    19.4260
                                                                                                  <.0001
                                                            Odds Ratio Estimates
                                                                 Point
                                                                               95% Wald
                                                  Effect
                                                              Estimate
                                                                           Confidence Limits
                                                  pirlarge
                                                                95.279
                                                                           12.064
                                                                                      752.486
                                                  piradult
                                                                 2,996
                                                                            1.027
                                                                                        8.744
                                                  fesmall
                                                               138.793
                                                                           15.477
                                                                                     >999.999
                                            Association of Predicted Probabilities and Observed Responses
                                            Percent Concordant
                                                                  90.8
                                                                          Somers' D
                                                                                       0.869
                                            Percent Discordant
                                                                   3.9
                                                                           Gamma
                                                                                       0.918
                                            Percent Tied
                                                                   5.3
                                                                           Tau-a
                                                                                       0.410
                                                                   6000
                                                                                       0.935
                                            Pairs
                                                                           С
```

#### SAS Program for Second Exercise

```
data coronary;
  input sex ecg age ca @@;
  datalines;
0 0 28 0 1 0 42 1 0 1 46 0 1 1 45 0 0 0 34 0 1 0 44 1 0 1 48 1 1 1 45 1
0 0 38 0 1 0 45 0 0 1 49 0 1 1 45 1 0 0 41 1 1 0 46 0 0 1 49 0 1 1 46 1
0 0 44 0 1 0 48 0 0 1 52 0 1 1 48 1 0 0 45 1 1 0 50 0 0 1 53 1 1 1 57 1
0 0 46 0 1 0 52 1 0 1 54 1 1 1 57 1 0 0 47 0 1 0 52 1 0 1 55 0 1 1 59 1
0 0 50 0 1 0 54 0 0 1 57 1 1 1 60 1 0 0 51 0 1 0 55 0 0 2 46 1 1 1 63 1
0 0 51 0 1 0 59 1 0 2 48 0 1 2 35 0 0 0 53 0 1 0 59 1 0 2 57 1 1 2 37 1
0 0 55 1 1 1 32 0 0 2 60 1 1 2 43 1 0 0 59 0 1 1 37 0 1 0 30 0 1 2 47 1
0 0 60 1 1 1 38 1 1 0 34 0 1 2 48 1 0 1 32 1 1 1 38 1 1 0 36 1 1 2 49 0
0 1 33 0 1 1 42 1 1 0 38 1 1 2 58 1 0 1 35 0 1 1 43 0 1 0 39 0 1 2 59 1
0 1 39 0 1 1 43 1 1 0 42 0 1 2 60 1 0 1 40 0 1 1 44 1
;
proc logistic descending;
  model ca=sex ecg age / lackfit;
run;
```

## **SAS Output for Second Exercise**

The LO	GISTIC Procedu	Jre			
м	odel Informati	on			
Data Set		WORK.CORONARY			
Response Variab	1e	са			
Number of Respo	nse Levels	2			
Model		binary logit			
Optimization Te	chnique	Fisher's scorin	g		
Number of Ob	Number of Observations Read				
Number of Ob	servations Use	d 78			
	Response Profi	le			
Ordered		Total			
Value	ca	Frequency			
1	1	41			
2	0	37			
Probab	ility modeled	is ca=1.			
Mode	1 Convergence	Status			
Convergence cri	terion (GCONV=	1E-8) satisfied.			
Мо	del Fit Statis	tics			
		Intercept			
	Intercept	and			
Criterion	Only	Covariates			
AIC	109.926	94.811			
SC	112.282	104.238			
-2 Log L	107.926	86.811			

	Testin	g Global N	ull Hypot	thesis: E	BETA=0			
Test		Chi	-Square	DF	Pr	> ChiSq		
Likeli	ihood Rat:	io :	21.1145	3		<.0001		
Score			18.5624	3		0.0003		
Wald			14.4410	3		0.0024		
	Analysis of Maximum Likelihood Estimates							
			Standard	ł	Wald			
Parameter	DF E	stimate	Error	r Chi	Square	Pr >	ChiSq	
Intercept	1	-5.6418	1.8061	l	9.7572	(	0.0018	
sex	1	1.3564	0.5464	ŧ	6.1616	(	0.0131	
ecg	1	0.8732	0.3843	3	5.1619	(	0.0231	
age	1	0.0929	0.0351	I	7.0003	(	0.0081	
		Odds R	atio Esti	imates				
		Point		95% Wa	ald			
	Effect	fect Estimate		Confidence Limi		5		
	sex	sex 3.882		1.330		80		
	ecg	2.39	5 1	1.127	5.08	86		
	age	1.09	7 1	.024	1.17	′5		
Associati	ion of Pro	edicted Pr	obabiliti	ies and (	)bserved	l Respons	ses	
Per	rcent Con	cordant	78.2	Somers	D (	.568		
Per	rcent Dis	cordant	21.5	Gamma	C	.569		
Per	rcent Tie	d	0.3	Tau-a	(	.287		
Pai	irs		1517	С	C	.784		
Hosmer and Lemeshow Goodness-of-Fit Test								
	Chi	-Square	DF	Pr > (	ChiSq			
		4.7766	8	0.	7812			