

## PubH 7407 Exam I, Spring 2008

There are four problems, each with multiple parts. Please start a new page for each problem. Put your name at the top of each page.

This exam is closed book and closed notes. You can (and should) use a calculator to numerically evaluate odds ratios and the like.

1. **Brief answer.** Answer the following questions with a few sentences.

- (a) Briefly describe in words what the polychoric correlation  $\hat{\rho}$  and the gamma statistic  $\hat{\gamma}$  measure. For what type of data are these measures valid?
- (b) Show that in a  $2 \times 2 \times 2$  table with outcome  $Y$ , treatment  $X$ , and strata  $Z$ , the additive logistic model:

$$\text{logit } P(Y = 1) = \alpha + \beta I\{X = 1\} + \tau I\{Z = 1\},$$

implies homogeneous association  $\theta_{XY(1)} = \theta_{XY(2)}$ . What is the common odds ratio  $\theta_{XY(k)}$  in terms of  $\{\alpha, \beta, \tau\}$ ?

- (c) Consider a generic  $2 \times 2$  table:

	$Y = 1$	$Y = 2$	
$X = 1$	$n_{11}$	$n_{12}$	$n_{1+}$
$X = 2$	$n_{21}$	$n_{22}$	$n_{2+}$
	$n_{+1}$	$n_{+2}$	$n_{++}$

What is the sample odds ratio  $\hat{\theta}$  in terms of  $\{n_{11}, n_{12}, n_{21}, n_{22}\}$ ? That is, what is the MLE of  $\theta$ ? Assume all cell counts are  $> 0$ .

- (d) What are the three ingredients in a generalized linear model? Explicitly define each of these three parts for the Poisson regression model for data  $\{(x_i, Y_i)\}_{i=1}^n$ :

$$Y_i \sim \text{Pois}(\mu_i), \quad \mu_i = e^{\alpha + \beta x_i}.$$

- (e) For multinomial sampling and product multinomial sampling, the likelihood ratio  $G^2$  and Pearson  $X^2$  tests of  $H_0 : X \perp Y$  in an  $I \times J$  table are the same (True or False)?

2. Consider data relating political affiliation (Democrat, Republican, or Independent) to the college of enrollment of U.S. university students (Letters – essentially literature, Engineering, Agriculture, or Education). SAS’s PROC FREQ and PROC GENMOD produce the following table of observed and expected counts, likelihood ratio and Pearson tests for independence, as well as the standardized Pearson residuals (the *r* below).

Table of College by Affiliation

College	Affiliation			Total
Frequency	Republican	Democrat	Independent	
Expected	Count	Count	Count	
Letters	34	61	16	111
	38.313	50.845	21.842	
Engineering	31	19	17	67
	23.126	30.69	13.184	
Agriculture	19	23	16	58
	20.019	26.568	11.413	
Education	23	39	12	74
	25.542	33.897	14.561	
Total	107	142	61	310

Statistics for Table of College by Affiliation

Statistic	DF	Value	Prob
Chi-Square	6	16.1613	0.0129
Likelihood Ratio Chi-Square	6	16.3901	0.0118

Obs	College	Affiliation	count	r
1	Letters	Republican	34	-1.07469
2	Letters	Democrat	61	2.41451
3	Letters	Independent	16	-1.74079
4	Engineering	Republican	31	2.28541
5	Engineering	Democrat	19	-3.23767
6	Engineering	Independent	17	1.32451
7	Agriculture	Republican	19	-0.31226
8	Agriculture	Democrat	23	-1.04285
9	Agriculture	Independent	16	1.68036
10	Education	Republican	23	-0.71235
11	Education	Democrat	39	1.36463
12	Education	Independent	12	-0.85835

- (a) Do you accept or reject that the college of enrollment is independent of political affiliation? Why or why not? Comment on the validity of the test’s *p*-value in terms of the expected cell counts.
- (b) Are any cells particularly ill-fit by the model of independence? If so, for which college(s) does this occur? Are any pairs of colleges particularly “unlike” each other in terms of political affiliation?

Combining Letters, Agriculture, and Education into one category called Other:

Table of College by Affiliation

College	Affiliation			Total
Frequency	Republic	Democrat	Independ	
Expected	an		ent	
Engineering	31	19	17	67
	23.126	30.69	13.184	
Other	76	123	44	243
	83.874	111.31	47.816	
Total	107	142	61	310

Statistics for Table of College by Affiliation

Statistic	DF	Value	Prob
Chi-Square	2	10.5103	0.0052
Likelihood Ratio Chi-Square	2	10.8539	0.0044

Omitting Engineering from the table:

Table of College by Affiliation

College	Affiliation			Total
Frequency	Republic	Democrat	Independ	
Expected	an		ent	
Letters	34	61	16	111
	34.716	56.185	20.099	
Agriculture	19	23	16	58
	18.14	29.358	10.502	
Education	23	39	12	74
	23.144	37.457	13.399	
Total	76	123	44	243

Statistics for Table of College by Affiliation

Statistic	DF	Value	Prob
Chi-Square	4	5.7698	0.2170
Likelihood Ratio Chi-Square	4	5.5361	0.2366

- (c) Verify that  $G_1^2 + G_2^2$  for the collapsed and reduced tables above add up to  $G^2$  for the full table on the previous page. Verify that  $df_1 + df_2 = df$  as well.
- (d) Partitioning the chi-squared  $G^2$  attempts to locate *why* the original test of  $H_0 : X \perp Y$  is rejected. Carefully interpret the followup tests for independence in the collapsed and partial tables. What do you conclude about political affiliation and college of enrollment among U.S. university students?

3. Consider the horseshoe crab data from your text. Recall that color is 1, 2, 3, 4 for light-medium, medium, dark-medium, and dark. The following SAS program was used to assess the marginal relationship of color to the probability of having one or more satellites, note the `link=identity` here:

```
proc genmod data=crabs1 descending; class color;
  model y=color / dist=bin link=identity;
```

The output is:

Analysis Of Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.3182	0.0993	0.1236	0.5128		10.27	0.0014
color	1	0.4318	0.1596	0.1189	0.7447		7.32	0.0068
color	2	0.4081	0.1093	0.1938	0.6224		13.94	0.0002
color	3	0.2727	0.1239	0.0299	0.5156		4.84	0.0277
color	4	0.0000	0.0000	0.0000	0.0000		.	.

- (a) What is the estimated relative risk of having one or more satellites for light-medium versus dark crabs?
- (b) What is the estimated difference in the probabilities, and 95% CI for the difference, of having one or more satellites for light-medium versus dark crabs? Test that this difference is significant at the 5% level.
4. Data relating occupational aspirations (high or low) to gender, residence (rural, small urban, large urban; e.g. country, small town, or big city), I.Q. (high or low), and socioeconomic status (ses, high or low) are analyzed via logistic regression below; all four predictors are categorical. **Note:** (1) the probability of high occupation aspiration is being modeled, (2) zero/one dummies are being used, but be careful determining which level is baseline. The SAS code follows, with the output on the following page.

```
data d1;
  input gender$ residence$ iq$ ses$ high low @@;
  total=high+low;
  datalines;
1 1 1 1 117 47 1 1 1 2 54 87 1 1 2 1 29 78 1 1 2 2 31 262
1 2 1 1 350 80 1 2 1 2 70 85 1 2 2 1 71 120 1 2 2 2 33 265
1 3 1 1 151 31 1 3 1 2 27 23 1 3 2 1 30 27 1 3 2 2 12 52
2 1 1 1 102 69 2 1 1 2 52 119 2 1 2 1 32 73 2 1 2 2 28 349
2 2 1 1 338 96 2 2 1 2 44 99 2 2 2 1 76 107 2 2 2 2 22 344
2 3 1 1 148 35 2 3 1 2 17 39 2 3 2 1 21 47 2 3 2 2 6 116
;
proc format;
  value $g '1'='Male' '2'='Female'; value $r '1'='Rural' '2'='Small urban' '3'='Large urban';
  value $i '1'='High' '2'='Low'; value $s '1'='High' '2'='Low';
proc logistic;
  class gender residence iq ses / param=ref; format gender$g. residence$r. iq$i. ses$s.;
  model high/total=gender residence iq ses gender*ses residence*ses / aggregate scale=none lackfit;
run;
```

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	18.5519	15	1.2368	0.2348
Pearson	18.5124	15	1.2342	0.2367

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
gender	1	25.3362	<.0001
residence	2	1.9876	0.3702
iq	1	533.0862	<.0001
ses	1	165.0355	<.0001
gender*ses	1	7.1901	0.0073
residence*ses	2	12.1137	0.0023

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.0327	0.1127	325.3931	<.0001
gender Female	1	-0.6002	0.1192	25.3362	<.0001
residence Large urban	1	0.2004	0.1797	1.2429	0.2649
residence Rural	1	-0.0522	0.1286	0.1647	0.6849
iq High	1	1.7720	0.0767	533.0862	<.0001
ses High	1	1.7092	0.1330	165.0355	<.0001
gender*ses Female High	1	0.4112	0.1534	7.1901	0.0073
residence*ses Large urban High	1	-0.0785	0.2198	0.1277	0.7208
residence*ses Rural High	1	-0.5823	0.1721	11.4435	0.0007

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
iq High vs Low	5.883	5.061 6.837

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.2636	7	0.6278

- Characterize how the odds of having high occupational aspirations is related to high versus low I.Q.
- Characterize how the odds of having high occupational aspirations is related to gender and socioeconomic status. In particular, how does the odds *ratio* for females versus males *change* when socioeconomic status changes from high to low?
- Characterize the effect of residence on occupational aspirations.
- Is there replication in this data set? Can you trust the deviance and Pearson GOF  $p$ -values? What do you conclude based on these tests?
- Comment on the residual plot (standardized Pearson residuals  $r_i$  versus the linear predictor  $\eta_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}$  for  $i = 1, \dots, 24$ ). This plot is on the next page.

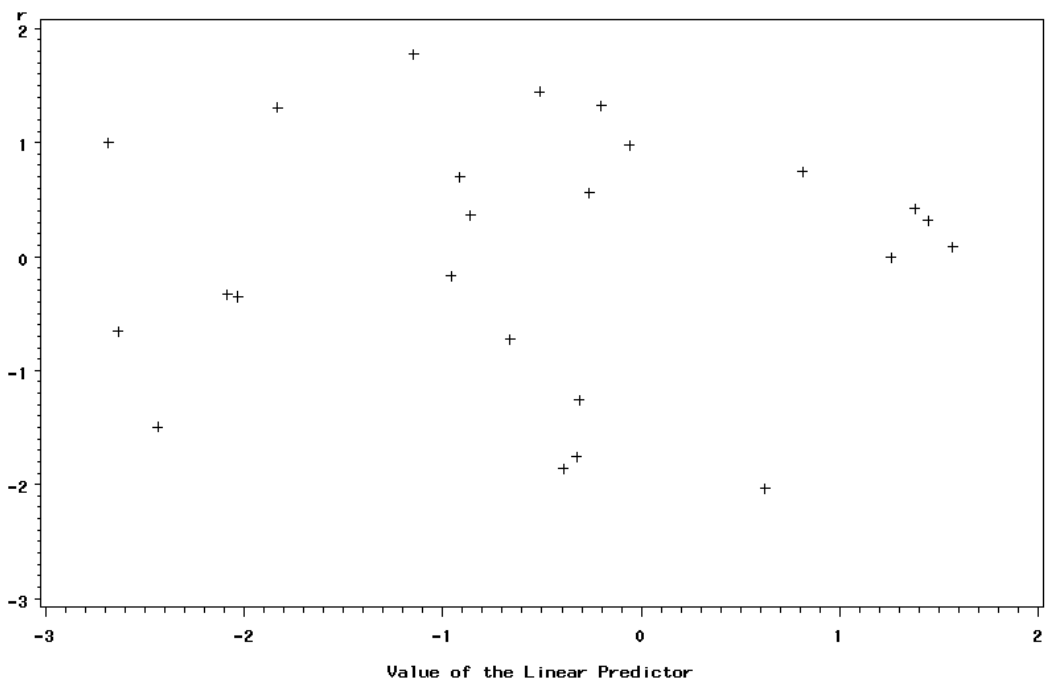


Figure 1: The 24 standardized Pearson residuals  $r_i$  vs. the linear predictors  $\eta_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}$ .