

PubH 7407 Exam I, Spring 2009

There are three problems, each with multiple parts. Each lettered part of a problem is worth 5 points, 100 points total – allocate time accordingly! 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) Show that simple odds ratios “flip” in terms of interpretation, i.e. show

$$\frac{P(Y = 1|X = 1)/P(Y = 2|X = 1)}{P(Y = 1|X = 2)/P(Y = 2|X = 2)} = \frac{P(X = 1|Y = 1)/P(X = 2|Y = 1)}{P(X = 1|Y = 2)/P(X = 2|Y = 2)}.$$

How does this relate to case-control studies?

This is an application of Bayes’ rule:

$$\begin{aligned} \frac{\frac{P(Y=1|X=1)}{P(Y=2|X=1)}}{\frac{P(Y=1|X=2)}{P(Y=2|X=2)}} &= \frac{\frac{P(X=1|Y=1)P(Y=1)/P(X=1)}{P(X=1|Y=2)P(Y=2)/P(X=1)}}{\frac{P(X=2|Y=1)P(Y=1)/P(X=2)}{P(X=2|Y=2)P(Y=2)/P(X=2)}} \\ &= \frac{P(X = 1|Y = 1)/P(X = 2|Y = 1)}{P(X = 1|Y = 2)/P(X = 2|Y = 2)} \end{aligned}$$

The role of X and Y can be switched. For example, in a case control study linking smoking to lung cancer, the odds of smoking for cases may be 3 times the odds for controls. This immediately implies that the odds of lung cancer for smokers is 3 times the odds for non-smokers.

(b) Consider a 2×2 table with fixed row totals n_{1+} and n_{2+} , i.e. product binomial sampling:

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

What are the MLEs $\hat{\pi}_1$ and $\hat{\pi}_2$ of $\pi_1 = P(Y = 1|X = 1)$ and $\pi_2 = P(Y = 1|X = 2)$? Derive the large sample distribution of $\hat{\pi}_1 - \hat{\pi}_2$.

$$\hat{\pi}_1 = \frac{n_{11}}{n_{1+}} \overset{\bullet}{\sim} N\left(\pi_1, \frac{\pi_1(1 - \pi_1)}{n_{1+}}\right) \perp \hat{\pi}_2 = \frac{n_{21}}{n_{2+}} \overset{\bullet}{\sim} N\left(\pi_2, \frac{\pi_2(1 - \pi_2)}{n_{2+}}\right).$$

So

$$\hat{\pi}_1 - \hat{\pi}_2 \overset{\bullet}{\sim} N \left(\pi_1 - \pi_2, \frac{\pi_1(1 - \pi_1)}{n_{1+}} + \frac{\pi_2(1 - \pi_2)}{n_{2+}} \right).$$

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

$$Y_i \sim \text{Bern}(\pi_i), \quad \pi_i = \alpha + \beta x_i.$$

What is the name of link used here?

(1) the random component, here $Y_i \sim \text{Bern}(\pi_i)$, (2) the systematic component, here the linear predictor $\eta_i = \alpha + \beta x_i$, and (3) the link relating $\mu_i = E(Y_i)$ to η_i , here the *identity* link : $E(Y_i) = \pi_i = \eta_i = \alpha + \beta x_i$.

- (d) A diagnostic test has specificity 0.98 and sensitivity 0.92. Find the odds ratio between true disease status and the diagnostic test result.

Sensitivity is $P(T+ | D+)$ and specificity is $P(T- | D-)$, so

$$\frac{\frac{P(D+|T+)}{P(D-|T+)}}{\frac{P(D+|T-)}{P(D-|T-)}} = \frac{\frac{P(T+|D+)}{P(T-|D+)}}{\frac{P(T+|D-)}{P(T-|D-)}} = \frac{0.92}{\frac{0.08}{0.02}} = 563.5.$$

- (e) Let (X, Y, Z) be a triple of dichotomous outcomes collected into a $2 \times 2 \times 2$ table. Find an example of cell counts such that (sample) homogeneous association occurs $\hat{\theta}_{XY(1)} = \hat{\theta}_{XY(2)}$ but independence marginally occurs: $\hat{\theta}_{XY} = 1$.

There are an infinite number of ways this can happen; here's one:

$Z = 1$	$Y = 1$	$Y = 2$
$X = 1$	5	8
$X = 2$	2	5
$Z = 2$	$Y = 1$	$Y = 2$
$X = 1$	5	2
$X = 2$	8	5

We have $\hat{\theta}_{XY(1)} = \hat{\theta}_{XY(2)} = 25/16 = 1.5625$. The collapsed table is

	Y = 1	Y = 2
X = 1	10	10
X = 2	10	10

with $\hat{\theta}_{XY} = 1$.

- (f) A coin is flipped sixteen times and the number of heads observed is $Y = 16$. Assume $Y \sim \text{bin}(16, \pi)$. Construct a score test that the coin is fair, i.e. of $H_0 : \pi = 0.5$ versus $H_a : \pi \neq 0.5$. Note that $P(Z \leq 4) = 0.99997$ to five decimal places for $Z \sim N(0, 1)$.

The score test uses $\pi_0 = 0.5$ to find the variance under the null hypothesis. Under the null H_0

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} \overset{\bullet}{\sim} N(0, 1).$$

So then

$$z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} = \frac{1 - 0.5}{\sqrt{(0.5)(0.5)/16}} = \frac{4(0.5)}{0.5} = 4$$

comes from a $N(0, 1)$ distribution. The p -value is $P(|Z| > |z_0|) = P(Z < -4 \text{ or } Z > 4) = 2(0.00003) = 0.00006$; we'd reject H_0 at any reasonable significance level.

2. The following SAS code reads in data that relates education to attitude toward legalized abortion from a General Social Survey. You analyzed this data in your homework.

```
data abort; input school$ attitude$ count @@;
datalines;
1 1 209 1 2 101 1 3 237
2 1 151 2 2 126 2 3 426
3 1 16 3 2 21 3 3 138
;
proc format;
  value $sc '1'='< high school' '2'='high school' '3'='> high school';
  value $ac '1'='generally disapprove' '2'='middle position' '3'='generally approve';
proc freq; weight count;
format school $sc. attitude $ac.;
table school*attitude/ expected chisq plcorr nopercnt nocol norow;
```

The output is:

school	attitude			Total
Frequency	generally disapprove	middle position	generally approve	
Expected				
< high school	209	101	237	547
	144.33	95.197	307.47	

high school	151	126	426	703
	185.49	122.35	395.16	
> high school	16	21	138	175
	46.175	30.456	98.368	
Total	376	248	801	1425

Statistics for Table of school by attitude

Statistic	DF	Value	Prob
Chi-Square	4	93.0338	<.0001
Likelihood Ratio Chi-Square	4	96.5267	<.0001
Gamma		0.3873	0.0366
Pearson Correlation		0.2530	0.0240
Polychoric Correlation		0.3432	0.0325

Sample Size = 1425

- (a) Is ‘education level’ ordinal or nominal? Is ‘attitude toward legalized abortion’ ordinal, nominal, interval, methodological, or continuous?

Both variables are ordinal.

- (b) Formally test that ‘education level’ and ‘attitude toward legalized abortion’ are independent using X^2 and/or G^2 . Are these tests valid here? Why or why not?

The Pearson $X^2 = 93.0$ with p -value < 0.0001 ; the likelihood ratio test $G^2 = 96.5$ with p -value < 0.0001 . Either test leads us to strongly reject $H_0 : X \perp Y$. Yes these tests are valid. All expected cell counts are at least one; in fact they’re all over ten.

- (c) Create a 3×3 table of ‘+’ and ‘-’ for each cell based on the sign of the Pearson (or raw) residual. Describe *and interpret* any pattern that you see.

+	+	-
-	+	+
-	-	+

We see, *qualitatively*, larger counts along the diagonal (and just above the diagonal) than what we’d expect under independence. This is in line with “concordant” outcomes of more approval with higher education. Information on whether the residuals are ‘significantly’ larger than what we’d expect is not provided.

-
- (d) The gamma, Pearson correlation, and polychoric correlation statistics are also reported. Obtain a 95% CI for each statistic.
-

For gamma, the CI is (0.316, 0.459), for Pearson based on default scores (0.206, 0.300), and for polychoric (0.280, 0.407). These are all obtained as the MLEs ± 1.96 times their standard error. My apologies if you understood the SE above to be p -values for the gamma, Pearson correlation, and polychoric correlations – I removed the part of the SAS output that identified this number as a SE.

- (e) Briefly (i.e. in one sentence each, not formulas) describe what each statistic in part (d) measures and formally test that these are zero.
-

All statistics are between -1 and 1 measure the strength of a particular kind of trend, pattern, or association in the data. The gamma statistic estimates the probability of concordance versus the probability of discordance. The Pearson correlation is literally the Pearson correlation based on replacing variable levels with scores, here $\{1, 2, 3\}$ for both X and Y . The polychoric correlation is the maximum likelihood estimate of the correlation of underlying continuous *latent* variables (Z_1, Z_2) that have a bivariate normal distribution.

The Pearson statistic measures ‘linear trend’ in the ordinal variables, the gamma statistic is said to measure ‘monotone association.’

All measures show a positive, weak to moderate but significant association between education and attitude. We would reject that any of the three measures are zero at the $\alpha = 5\%$ significance level based on the CIs reported in part (d).

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 code regresses the mean number of satellites on color, width, and weight; some output follows the code.

```
data crabs;
input color spine width satell weight; weight=weight/1000; color=color-1;
datalines;
3 3 28.3 8 3050
4 3 22.5 0 1550
...et cetera...
5 3 27.0 0 2625
3 2 24.5 0 2000
;
proc genmod; class color spine;
  model satell = color width weight color*width width*weight / dist=poi link=log type3;
run;
```

The GENMOD Procedure

Model Information

Data Set WORK.CRABS
 Distribution Poisson
 Link Function Log
 Dependent Variable satell

Class Level Information

Class	Levels	Values
color	4	1 2 3 4
spine	3	1 2 3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	163	530.8508	3.2568
Pearson Chi-Square	163	534.1342	3.2769
Log Likelihood		86.9602	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-9.1541	2.6751	-14.3971	-3.9111	11.71	0.0006
color	1	9.1831	3.3887	2.5414	15.8247	7.34	0.0067
color	2	3.5665	2.3722	-1.0830	8.2160	2.26	0.1327
color	3	2.5047	2.6348	-2.6593	7.6688	0.90	0.3418
color	4	0.0000	0.0000	0.0000	0.0000	.	.
width	1	0.3169	0.1034	0.1142	0.5195	9.39	0.0022
weight	1	2.6382	0.7801	1.1092	4.1671	11.44	0.0007
width*color	1	-0.3311	0.1273	-0.5807	-0.0815	6.76	0.0093
width*color	2	-0.1285	0.0898	-0.3044	0.0475	2.05	0.1524
width*color	3	-0.0977	0.0998	-0.2932	0.0979	0.96	0.3277
width*color	4	0.0000	0.0000	0.0000	0.0000	.	.
width*weight	1	-0.0728	0.0259	-0.1236	-0.0219	7.87	0.0050

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
color	3	7.98	0.0464
width	1	5.86	0.0155
weight	1	12.60	0.0004
width*color	3	7.31	0.0627
width*weight	1	8.92	0.0028

- (a) Write down the model fit to data $\{(C_i, W_i, Wt_i, Y_i)\}_{i=1}^{173}$, where C_i is color, Wt_i is weight in *kg*, W_i is width in *cm*, and Y_i is the number of satellites for the i^{th} crab.

Dropping the subscript i ,

$$\log E(Y) = -9.15 + 9.18I\{C = 1\} + 3.57I\{C = 2\} + 2.50I\{C = 3\} + 0.32W + 2.64Wt - 0.33I\{C = 1\}W - 0.13I\{C = 2\}W - 0.10I\{C = 3\}W - 0.07W(Wt).$$

- (b) Assuming the model is correct, characterize how the mean number of satellites changes with color for a fixed weight and width.

$$\begin{aligned} \frac{E(Y|C = 1)}{E(Y|C = 2)} &= \frac{e^{-9.15+9.18+0.32W+2.64Wt-0.33W-0.07W(Wt)}}{e^{-9.15+3.57+0.32W+2.64Wt-0.13W-0.07W(Wt)}} \\ &= \frac{e^{9.18-0.33W}}{e^{3.57-0.13W}} = e^{5.61-0.2W} = 275e^{-0.2W}. \end{aligned}$$

This is $275e^{-0.20(20)} = 5.0$ at $W = 20$ cm and $275e^{-0.20(30)} = 0.7$ at $W = 30$ cm. The expected number of satellites is 5 times higher for light-medium versus medium crabs of width 20 cm; the relationship flips for wider ($W = 30$ cm) crabs. You can continue with this sort of comparison for all 6 pairings of color. Full credit goes to just doing this for one pair of colors.

- (c) Characterize how the mean number of satellites changes with width for a fixed color and weight.

Unfortunately, this will change depending on color. For $C = 1$,

$$\begin{aligned} \frac{E(Y|W + 1, C = 1)}{E(Y|W, C = 1)} &= \frac{e^{-9.15+9.18+0.32(W+1)+2.64Wt-0.33(W+1)-0.07(W+1)(Wt)}}{e^{-9.15+9.18+0.32W+2.64Wt-0.33W-0.07W(Wt)}} \\ &= e^{0.32-0.33-0.07Wt} = e^{-0.014-0.07Wt}. \end{aligned}$$

As in (b), full credit goes to doing this for one color.

- (d) Characterize how the mean number of satellites changes with weight for a fixed color and width.

$$\frac{E(Y|Wt + 1, W)}{E(Y|Wt, W)} = e^{2.64-0.073W}.$$

- (e) What do the Type III tests test? What do you conclude?

The type III tests test whether each continuous or categorical covariate can be dropped from the full model with all covariates. Everything is significant at the

5% level except for `width*color`, which is *almost* significant. This indicates (to me at least) that all predictors are important here.

(f) Why does the `width*color` interaction have 3 degrees of freedom?

The categorical covariate `color` has 4 levels. The main effect for `color` is defined by the inclusion of $4 - 1 = 3$ dummy variables that provide offsets to baseline (`color=4`, dark) log-mean number of satellites. The interaction term allows for this main effect to change with the value of `width` and so has an additional 3 parameters.

(g) Use the output to check for overdispersion. What would you estimate the scale ϕ to be?

$\hat{\phi} = X^2/df = 3.28 > 1$, so there is evidence of overdispersion. The standard errors are scaled by $\sqrt{\hat{\phi}} \approx 1.811$.

(h) Do parameter estimates change using quasi-likelihood? Do the standard errors change? If so, by what factor? Hint: see (g).

The parameter estimates are exactly the same as the score equations do not change. The estimated covariance matrix simply multiplied by ϕ and so standard errors are multiplied by an estimate of $\sqrt{\phi}$ (the estimated variances of $\hat{\beta}_j$ are along the diagonal of the covariance matrix). Here, $\hat{\phi} = X^2/df = 3.2769$ from the Pearson GOF statistic and df .

The following are regression effects under quasi-likelihood.

Analysis Of Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq	
Intercept	1	-9.1541	4.8424	-18.6451	0.3369	3.57	0.0587	
color	1	9.1831	6.1342	-2.8398	21.2060	2.24	0.1344	
color	2	3.5665	4.2943	-4.8501	11.9832	0.69	0.4062	
color	3	2.5047	4.7695	-6.8434	11.8528	0.28	0.5995	
color	4	0.0000	0.0000	0.0000	0.0000	.	.	
width	1	0.3169	0.1871	-0.0499	0.6837	2.87	0.0904	

weight		1	2.6382	1.4121	-0.1295	5.4058	3.49	0.0617
width*color	1	1	-0.3311	0.2305	-0.7829	0.1207	2.06	0.1509
width*color	2	1	-0.1285	0.1625	-0.4470	0.1900	0.63	0.4292
width*color	3	1	-0.0977	0.1806	-0.4517	0.2564	0.29	0.5887
width*color	4	0	0.0000	0.0000	0.0000	0.0000	.	.
width*weight		1	-0.0728	0.0470	-0.1648	0.0193	2.40	0.1213

- (i) How does this differ from the original (likelihood-based) table? Describe what is happening in words.

Describe what the quasi-likelihood approach is adding to the model; i.e. how is the original model generalized? Be specific.

All of the standard errors are multiplied by $\hat{\phi} = \sqrt{3.2769} = 1.81$. This has the effect of increasing the size of the CIs by 1.81 and decreasing the Wald χ^2 test statistics and increasing p -values.

The quasi-likelihood approach here adds one parameter to the model, for Poisson $\text{var}(Y_i) = \phi E(Y_i)$.