CATEGORICAL DATA ANALYSIS, 3rd edition

Extra Exercises

Alan Agresti

Version February 20, 2012, ©Alan Agresti 2012

This file contains extra exercises. Most of these were in the first or second edition of the text, did not fit in the 3rd edition. They are organized by chapter. Instructors are welcome to use them for homeworks or exams.

Chapter 1

1. In the following examples, identify the response variable and the explanatory variables.

   (a) Attitude toward gun control (favor, oppose), Gender (female, male), Mother's education (high school, college).

   (b) Heart disease (yes, no), Blood pressure, Cholesterol level.

   (c) Hospital (A, B, C), Chemotherapy treatment (standard, new), Response of tumor to chemotherapy (complete elimination, partial reduction, stable, growth progression).

   (d) Race (white, nonwhite), Religion (Catholic, Jewish, Protestant, Other), Vote for President (Democrat, Republican, Other), Annual income.

2. Which measurement scale is most appropriate for attitude toward legalization of abortion (disapprove always, approve in certain cases, approve always).

3. Describe a potential study with a categorical response variable. List explanatory variables that would be important. For each variable, identify the measurement scale, and indicate whether to treat it as continuous or discrete, quantitative or qualitative.

4. In a large city, 50% of the population is black. Prospective jurors for court trials are selected from this population. For each selection of a juror, π denotes the probability that a black person is selected. A supposedly random sampling of 12 prospective jurors contains 1 black person. Using the exact binomial test, find the P-value for testing $H_0 : \pi = 0.5$ against $H_a : \pi \neq 0.5$.

5. A sample of 100 adults suffer from migraine headaches. A new analgesic is claimed to provide greater relief than a standard one. After using each analgesic in a crossover experiment, 40 reported greater relief with the standard analgesic and 60 reported greater relief with the new one. Analyze these data.

6. A criminologist studies the proportion of U. S. citizens who live in a home in which firearms are available. The 1991 General Social Survey asked respondents, "Do you have in your home any guns or revolvers?" Of the respondents, 393 answered 'yes' and 583 answered 'no.' Analyze these data.

7. Suppose that $P(Y_i = 1|\pi_i) = \pi_i, i = 1, \ldots, n$, where $\{\pi_i\}$ are independent from $g(\cdot)$. Explain why Y has a bin(n, ρ) distribution unconditionally but not conditionally on $\{\pi_i\}$. (*Hint:* In each case, is Y a sum of independent, identical Bernoulli trials?)

8. For a binomial parameter π, show how the inversion process for constructing a confidence interval works with (**a**) the Wald test, and (**b**) the score test.

9. Show that the Jeffreys prior for π equals the beta density with $\alpha_1 = \alpha_2 = .50$.

Chapter 2

1. A diagnostic test has sensitivity = specificity = 0.80. Find the odds ratio between true disease status and the diagnostic test result.

2. An estimated odds ratio for adult females between the presence of squamous cell carcinoma (yes, no) and smoking behavior (smoker, nonsmoker) equals 11.7 when the smoker category has subjects whose smoking level s is $0 < s < 20$ cigarettes per day; it is 26.1 for smokers with $s \geq 20$ cigarettes per day (R. C. Brownson et al., *Epidemiology* **3**: 61–64, 1992). Show that the estimated odds ratio between carcinoma (yes, no) and the smoking levels ($s \geq 20, 0 < s < 20$) equals 2.2.

3. Table 1.1 refers to a retrospective study of lung cancer and tobacco smoking among patients in several English hospitals. The table compares male lung cancer patients with control patients having other diseases, according to the average number of cigarettes smoked daily over a 10-year period preceding the onset of the disease.

   a. Find the sample odds of lung cancer at each smoking level and the five odds ratios that pair each level of smoking with no smoking. As smoking increases, is there a trend? Interpret.

   b. If the log odds of lung cancer is linearly related to smoking level, the log odds in row i satisfies $\log(\text{odds}_i) = \alpha + \beta i$. Show that this implies that the local odds ratios are identical.

   c. Using these data, can you estimate the probability of lung cancer at each level of smoking? Are the estimated odds ratios in part (a) meaningful? Explain.

   d. Show that the disease groups are *stochastically ordered* with respect to their distributions on smoking of cigarettes. Interpret.

Table 1.1:

Daily Average Number of Cigarettes	Disease Group: Lung Cancer Patients	Disease Group: Control Patients
None	7	61
<5	55	129
5–14	489	570
15–24	475	431
25–49	293	154
50+	38	12

Source: Reprinted with permission from R. Doll and A. B. Hill, *British Med. J.*, **2**: 1271–1286 (1952).

Table 1.2:

Oral Contraceptive Practice	Myocardial Infarction: Yes	Myocardial Infarction: No
Used	23	34
Never used	35	132
Total	58	166

Reprinted with permission from Mann et al., *British J. Med.* **2**: 241-245 (1975).

4. Binomial parameters for two groups are graphed, with π_1 on the horizontal axis and π_2 on the vertical axis. Plot the locus of points for a 2×2 table having (**a**) relative risk = 0.5, (**b**) odds ratio = 0.5, and (**c**) difference of proportions = -0.5.

5. Table 1.2 comes from a study that investigated the effect of oral contraceptive use on the likelihood of heart attacks. The 58 subjects in the first column represent married women under 45 years of age treated for myocardial infarction in two hospital regions in England and Wales during 1968–1972. Each case was matched with three control patients in the same hospitals who were not being treated for myocardial infarction. All subjects were then asked whether they had ever used oral contraceptives. Analyze these data.

6. When X and Y are ordinal with counts $\{n_{ij}\}$:

   a. Explain why the $n(n-1)/2$ pairs of observations partition into $C + D + T_X + T_Y - T_{XY}$, where $T_X = \sum n_{i+}(n_{i+} - 1)/2$ pairs are tied on X, T_Y pairs are tied on Y, and T_{XY} pairs are tied on X and Y.

   b. For each ordered pair of observations (X_a, Y_a) and (X_b, Y_b), let $X_{ab} = \text{sign}(X_a - X_b)$ and $Y_{ab} = \text{sign}(Y_a - Y_b)$. Show that the sample correlation for the $n(n-1)$ distinct (X_{ab}, Y_{ab}) pairs is

$$\tau_b = \frac{C - D}{\{[n(n-1)/2 - T_X]\,[n(n-1)/2 - T_Y]\}^{1/2}}$$

This ordinal measure is called *Kendall's tau-b* (Kendall 1945).

c. Let $d = (C - D)/\,[n(n-1)/2 - T_X]$. Explain why d is the difference between the proportions of concordant and discordant pairs out of those pairs untied on X (Somers 1962). (For $2 \times c$ tables d is the sample version of Δ in (2.15). For 2×2 tables, d equals the difference of proportions, and tau-b equals the correlation between X and Y.)

7. Refer to Δ in (2.15) for comparing two ordinal distributions.

   a. Show that Δ relates to $\alpha = P(Y_1 > Y_2) + (\frac{1}{2})P(Y_1 = Y_2)$ by

   $$\alpha = (\Delta + 1)/2, \quad \Delta = 2\alpha - 1,$$

   with α having range [0, 1].

   b. Consider $\theta = \frac{P(Y_1 > Y_2)}{P(Y_2 > Y_1)}$. If we artificially identify row 1 as the higher level of the group variable, show $\hat{\theta} = C/D$. When $c = 2$, show $\hat{\theta}$ is an odds ratio. (For more about such measures, see Agresti 1980, 2010).

8. When all rows and columns have positive probability, show that independence is equivalent to all odds ratios $\{\alpha_{ij} = 1\}$.

9. A $2 \times J$ table compares the distributions on an ordinal response for two groups. The *cumulative odds ratios* are

   $$\theta_j = \frac{P(Y_1 \le j)/P(Y_1 > j)}{P(Y_2 \le j)/P(Y_2 > j)}, \qquad j = 1, \ldots, J-1.$$

   a. Show that $\log \theta_j \ge 0$ for all j is equivalent to row 2 being stochastically higher than row 1.

   b. If all local log odds ratios are nonnegative, $\log \theta_j \ge 0$ for $1 \le j \le J-1$ (Lehmann 1966). Show by counterexample that the converse is not true.

10. Show that Yule's Q falls between -1 and 1. State conditions under which $Q = -1$ or $Q = 1$.

11. The odds ratio between whether a boy scout (yes, no) and juvenile delinquent behavior (yes, no) is 1.0 at each fixed level of socioeconomic status (SES), but 0.5 marginally. Why is it misleading to claim that scouting leads to lower delinquency rates?

12. Explain why for three events E_1, E_2, and E_3 and their complements, it is possible that $P(E_1|E_2) > P(E_1|\overline{E}_2)$ even if both $P(E_1|E_2E_3) < P(E_1|\overline{E}_2E_3)$ and $P(E_1|E_2\overline{E}_3) < P(E_1|\overline{E}_2\overline{E}_3)$. (*Hint*: Use Simpson's paradox for a three-way table.)

Chapter 3

1. Refer to Table 1.2. Is there evidence of an association between myocardial infarction and use of oral contraceptives? Use an inferential procedure, and interpret.

2. In a study of the relationship between stage of breast cancer at diagnosis (local or advanced) and a woman's living arrangement, of 144 women living alone, 41.0% had an advanced case; of 209 living with spouse, 52.2% were advanced; of 89 living with others, 59.6% were advanced. The authors reported the P-value for the relationship as 0.02 (D. J. Moritz and W. A. Satariano, *J. Clin. Epidemiol.* **46**: 443–454, 1993). Reconstruct the analysis performed to obtain this P-value.

3. In the 2008 General Social Survey, of the 36 subjects who said they were gay or bisexual, 32 agreed with the statement that homosexuals should have the right to marry, whereas of the 985 heterosexuals, 459 agreed. Analyze these data, and summarize your conclusions.

4. According to a survey by the European Commission in late 2000 of about 16,000 Europeans (Eurobarometer 54), 1000 in each country, the percent support for a common currency (the euro) was 64% in the Netherlands, 21% in the U.K., and 79% in Italy. Analyze these data.

5. Refer to Table 2.1. Partition G^2 for testing whether the incidence of heart attacks is independent of aspirin intake into two components. Interpret.

6. For multinomial sampling in an $I \times J$ table, assuming statistical independence show that the ML estimator $\hat{\pi}_{ij} = n_{i+}n_{+j}/n^2$.

7. Refer to the example in Section 3.3.8. Explain how the range of n for which $\text{MSE}(\{\hat{\pi}_{ij}\}) < \text{MSE}(\{p_{ij}\})$ changes as δ increases.

8. Refer to text Exercise 2.27 on the attributable risk. For multinomial sampling, show how to obtain a confidence interval for AR by first finding one for $\log(1 - AR)$.

9. For ordinal variables, consider gamma (2.14). Let

$$\pi_{ij}^{(c)} = \sum_{a<i}\sum_{b<j}\pi_{ab} + \sum_{a>i}\sum_{b<j}\pi_{ab}, \quad \pi_{ij}^{(d)} = \sum_{a<i}\sum_{b>j}\pi_{ab} + \sum_{a>i}\sum_{b<j}\pi_{ab},$$

   where i and j are fixed in the summations. Show that the probabilities of concordance and discordance are $\Pi_c = \sum_i \sum_j \pi_{ij}\pi_{ij}^{(c)}$ and $\Pi_d = \sum_i \sum_j \pi_{ij}\pi_{ij}^{(d)}$. Use the delta method to show that the large-sample normality applies for $\hat{\gamma}$ (Goodman and Kruskal 1963) with (3.9) using

$$\phi_{ij} = 4[\Pi_d\pi_{ij}^{(c)} - \Pi_c\pi_{ij}^{(d)}]/(\Pi_c + \Pi_d)^2, \quad \sum_i\sum_j \pi_{ij}\phi_{ij} = 0,$$

$$\sigma^2 = \frac{16}{(\Pi_c + \Pi_d)^4} \sum_i \sum_j \pi_{ij} \left[\Pi_d \pi_{ij}^{(c)} - \Pi_c \pi_{ij}^{(d)}\right]^2.$$

Chapter 4

1. For games in baseball's National League between 1900 and 1990, consider the probability π that the starting pitcher pitched a complete game. Let x = decade since 1900 ($x = 1, 2, \ldots$).

   a. Between 1900 and 1990, data (from George Will, *Newsweek*, Apr. 10, 1989) were fit well by the linear probability model $\hat{\pi} = 0.7578 - 0.0694x$. Interpret 0.7578 and -0.0694, and substitute $x = 12$ to predict the percentage of complete games between 2010 and 2019. Is this prediction plausible? Why?

   b. Between 1900 and 1990, data were fit well by the logistic regression model $\hat{\pi} = \exp(1.148 - 0.315x)/[1 + \exp(1.148 - 0.315x)]$. Obtain $\hat{\pi}_i$ for $x = 12$. Is this more plausible than the prediction in (**a**)?

2. For the horseshoe crab counts of satellites, using the identity link with x = weight, $\hat{\mu} = -2.60 + 2.264x$, where $\hat{\beta} = 2.264$ has SE = 0.228. Interpret, and conduct inference.

3. One hundred leukemia patients were randomly assigned to two treatments. During the study, 10 subjects on treatment A died and 18 subjects on treatment B died. The total time at risk was 170.4 years for treatment A and 147.3 years for treatment B. Test whether the two treatments have the same death rates. Compare the rates with a confidence interval.

4. For Table 4.5, fit a model in which death rate depends only on age. Interpret the age effect.

5. An article by W. A. Ray et al. (*Amer. J. Epidemiol.* **132**: 873–884, 1992) dealt with motor vehicle accident rates for 16,262 subjects aged 65–84 years, with data on each for up to 4 years. In 17.3 thousand years of observation, the women had 175 accidents in which an injury occurred. In 21.4 thousand years, men had 320 injurious accidents.

   a. Find a 95% confidence interval for the true overall rate of injurious accidents.

   b. Using a model, compare the rates for men and women by interpreting a model parameter estimate and giving a confidence interval.

6. A table at the Web site for the second edition (*www.stat.ufl.edu/∼aa/cda2/cda.html*) shows the number of train miles (in millions) and the number of collisions involving British Rail passenger trains between 1970 and 1984. A Poisson model assuming a constant log rate α over the 14-year period has $\hat{\alpha} = -4.177$ (SE = 0.1325) and $X^2 = 14.8$ (df = 13). Interpret.

Table 1.3:

Team	Attendance (thousands)	Arrests	Team	Attendance (thousands)	Arrests
Aston Villa	404	308	Shrewsbury	108	68
Bradford City	286	197	Swindon Town	210	67
Leeds United	443	184	Sheffield Utd.	224	60
Bournemouth	169	149	Stoke City	211	57
West Brom	222	132	Barnsley	168	55
Hudderfield	150	126	Millwall	185	44
Middlesbro	321	110	Hull City	158	38
Birmingham	189	101	Manchester City	429	35
Ipswich Town	258	99	Plymouth	226	29
Leicester City	223	81	Reading	150	20
Blackburn	211	79	Oldham	148	19
Crystal Palace	215	78			

Source: The *Independent* (London), Dec. 21, 1988. Thanks to P. M. E. Altham for showing me these data.

7. Table 1.3 lists total attendance (in thousands) and the total number of arrests in the 1987–1988 season for soccer teams in the Second Division of the British football league. Let $Y =$ number of arrests for a team, and let $t =$ total attendance. Explain why the model $E(Y) = \mu t$ might be plausible. Assuming Poisson sampling, fit it and interpret. Plot arrests against attendance, and overlay the prediction equation. Use residuals to identify teams that had arrest counts much different than expected.

8. Refer to Exercise 4.7. The wafers are also classified by thickness of silicon coating ($z = 0, \text{low}; z = 1, \text{high}$). The first five imperfection counts reported for each treatment refer to $z = 0$ and the last five refer to $z = 1$. Analyze these data.

9. Refer to Exercise 14.10 on frequency of sexual intercourse. Analyze these data.

10. Refer to model (4.15). Given the times at risk $\{t_{ij}\}$, show that sufficient statistics are $\{n_{i+}\}$ and $\{n_{+j}\}$.

11. Table 1.4 reports the frequency of all reported game-related concussions for players on 49 college football teams, between 1975 and 1982. The total time at risk for these data was 216,690 athlete-games. Suppose the total was identical for offense and defense, the total for blocking was six times that for tackling, and the total for rushing plays was 2.2 times that for passing plays. Find the total time at risk per cell and the sample rates of concussion. Which activity has greatest sample rate? Use loglinear models to summarize these rates.

12. Conditional on λ, Y has a Poisson distribution with mean λ. Values of λ vary according to gamma density (14.12), which has $E(\lambda) = \mu$, $\text{var}(\lambda) = \mu^2/k$. Greenwood

Table 1.4:

		Activity	
Team	Situation	Tackle	Block
Offense	Rushing	125	129
	Passing	85	31
Defense	Rushing	216	61
	Passing	62	16

Source: Reprinted with permission from Buckley, W. E. (1988), *Amer. J. Sports Med.*, **16**: 51–56.

and Yule (1920) noted that marginally Y has the negative binomial distribution (4.13). Show this.

13. For binary data with sample proportion y_i based on n_i trials, we use quasi-likelihood to fit a model using variance function (4.53). Show that parameter estimates are the same as for the binomial GLM but that the covariance matrix multiplies by ϕ.

14. In a GLM, suppose that $\text{var}(Y) = v(\mu)$ for $\mu = E(Y)$. Show that the link function g satisfying $g'(\mu) = [v(\mu)]^{-1/2}$ has the same weight matrix $\boldsymbol{W}^{(t)}$ at each cycle. Show this link for a Poisson random component is $g(\mu) = 2\sqrt{\mu}$.

15. When k is unknown show the negative binomial does not have exponential dispersion form. Jørgensen (1986) argued that a more appropriate form for two-parameter discrete distributions is

$$f(y; \theta, \phi) = \exp\{y\theta - b(\theta)/a(\phi) + c(y, \phi)\}.$$

Show the negative binomial distribution has this form.

Chapter 5

1. Refer to Table 1.1. Using scores (0, 3, 9.5, 19.5, 37, 55) for cigarette smoking, analyze these data using a logistic model. Is the intercept estimate meaningful? Explain.

2. For the model for the horseshoe crab data with weight and color predictors, the estimated color effects are monotone across the four categories. Fit a simpler model that treats color as quantitative and assumes a linear effect. Interpret its color effect. Compare the fit to the model that treats color as nominal scale. Now add width to the model. What effect does the strong positive correlation between width and weight have? Are both needed in the model?

3. Table 1.5 appeared in a national study of 15- and 16-year-old adolescent. The event of interest is ever having sexual intercourse. Analyze, including description

Table 1.5:

		Intercourse	
Race	Gender	Yes	No
White	Male	43	134
	Female	26	149
Black	Male	29	23
	Female	22	36

Source: S. P. Morgan and J. D. Teachman, *J. Marriage Fam.* **50:** 929–936 (1988). Reprinted with permission from the National Council on Family Relations.

and inference about the effects of gender and race, goodness of fit, and summary interpretations.

4. The National Collegiate Athletic Association studied graduation rates for freshman student athletes during the 1984–1985 academic year. The (sample size, number graduated) totals were (796, 498) for white females, (1625, 878) for white males, (143, 54) for black females, and (60, 197) for black males (J. J. McArdle and F. Hamagami, *J. Amer. Statist. Assoc.* **89:** 1107–1123, 1994). Analyze and interpret.

5. Let Y denote a subject's opinion about current laws legalizing abortion (1 = support), for gender h ($h = 1, \text{female}; h = 2, \text{male}$), religious affiliation i ($i = 1$, Protestant; $i = 2$, Catholic; $i = 3$, Jewish), and political party affiliation j ($j = 1$, Democrat; $j = 2$, Republican; $j = 3$, Independent). For survey data, software for fitting the model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_h^G + \beta_i^R + \beta_j^P$$

reports $\hat{\alpha} = 0.62, \hat{\beta}_1^G = 0.08, \hat{\beta}_2^G = -0.08, \hat{\beta}_1^R = -0.16, \hat{\beta}_2^R = -0.25, \hat{\beta}_3^R = 0.41, \hat{\beta}_1^P = 0.87, \hat{\beta}_2^P = -1.27, \hat{\beta}_3^P = 0.40$.

   a. Interpret how the odds of support depends on religion.

   b. Estimate the probability of support for the group most (least) likely to support current laws.

   c. If, instead, parameters used constraints $\beta_1^G = \beta_1^R = \beta_1^P = 0$, report the estimates.

6. Refer to Table 9.1, treating marijuana use as the response variable. Analyze these data.

7. Refer to model (5.13) for the horseshoe crabs.

   a. Fit the model using $x =$ weight. Interpret effects of weight and color.

   b. Does the model permitting interaction provide an improved fit? Interpret.

Table 1.6:

	Student Smokes	Student Does Not Smoke
Both parents smoke	400	1380
One parent smokes	416	1823
Neither parent smokes	188	1168

By permission, S.V. Zagona, *Studies and Issues in Smoking Behavior*, Tucson: The University of Arizona Press, Copyright 1967.

Table 1.7:

		1980 Presidential Vote	
Race	Political Views[a]	Reagan	Carter or other
White	1	1	12
	2	13	57
	3	44	71
	4	155	146
	5	92	61
	6	100	41
	7	18	8
Nonwhite	1	0	6
	2	0	16
	3	2	23
	4	1	31
	5	0	8
	6	2	7
	7	0	4

Source: 1982 General Social Survey.
[a]Political views range from 1 = extremely liberal to 7 = extremely conservative.

c. For part (b), construct a confidence interval for a difference between the slope parameters for medium-light and dark crabs. Interpret.

d. Using models that treat color as quantitative, repeat the analyses in parts (a) to (c).

8. Use models to analyze Table 1.6 on smoking habits of students in Arizona high schools.

9. Table 1.7, reported by Clogg and Shockey (1988), comes from the 1982 General Social Survey.

   (a) Treating vote as the response, fit logistic model (5.12) with nominal main effects. Does there seem to be a trend in the effects at the seven levels of political views?

Table 1.8:

Blood Pressure	Serum Cholesterol (mg/100 ml)						
	<200	200–209	210–219	220–244	245–259	260–284	>284
<177	2/53	0/21	0/15	0/20	0/14	1/22	0/11
117–126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
127–136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
137–146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
147–156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
157–166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
167–186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
>186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Source: Reprinted with permission from Cornfield (1962).

(b) Fit a logit model that uses the ordinal nature of political views. Carefully interpret parameter estimates for this model.

(c) Test the fit of the models in (a) and (b), and analyze whether the model in (a) gives a significantly better fit.

10. A research study used multiple logistic regression to predict the stage of breast cancer (1 = advanced, 0 = local) at diagnosis for a sample of women. A table referring to demographic factors reported the estimated odds ratio for the effect of living arrangement (three categories) as 2.02 for spouse versus alone and 1.71 for others versus alone. Estimate the odds ratios for spouse versus others.

11. For data from the 1998 General Social Survey on Y = whether one favors the death penalty for persons convicted of murder (1 = yes), x_1 = race (1 = white, 0 = other), and x_2 = opinion about how courts treat criminals (1 = not harsh enough, 2 = about right, 3 = too harshly), $\text{logit}[\hat{P}(Y = 1)] = 1.30 + 1.24x_1 - 0.82x_2$. Interpret the predictor effects. Find the estimate of $P(Y = 1)$ when $x_1 = 0$ and $x_2 = 3$.

12. Table 1.8 is based on data reported by Cornfield (1962). Subjects were classified on blood pressure, cholesterol level, and whether they developed coronary heart disease during a follow-up period. For instance, at the lowest level of both predictors, 2 of 53 cases had heart disease. Plot sample logits or smooth the data to show the trend using cholesterol level alone to predict heart disease. Fit and interpret a logit model that describes the trend.

13. For Table 1.8, fit a logit model that simultaneously describes effects of cholesterol and blood pressure on heart disease. Interpret effects.

14. Refer to the previous exercise. Describe each predictor's effect by estimating (a) the model slope for a standard deviation change in the predictor, (b) the change in the probability of heart disease between the scores for the first and last categories, at the mean score for the other predictor.

Table 1.9:

Gender	Cumulative GPA	Black: High Self-Esteem	Black: Low Self-Esteem	White: High Self-Esteem	White: Low Self-Esteem
Males	High	15	9	17	10
	Low	26	17	22	26
Females	High	13	22	22	32
	Low	24	23	3	17

Source: Reprinted with permission of the Helen Dwight Reid Educational Foundation from D. H. Demo and K. D. Parker, *J. Social Psych.*, **127**: 345–355 (1987). Published by Heldref Publications, copyright ©1987.

Table 1.10:

Follow Politics Regularly	USSR Yes	USSR No	USA Yes	USA No	UK Yes	UK No	Italy Yes	Italy No	Mexico Yes	Mexico No
Primary	94	84	227	112	356	144	166	526	447	430
Secondary	318	120	371	71	256	76	142	103	78	25
College	473	72	180	8	22	2	47	7	22	2

Source: Reprinted with permission from DiFrancesco and Critelman (1984).

15. Table 1.9 refers to the effect of academic achievement on self-esteem among black and white college students. Treating self-esteem as a response variable, analyze these data.

16. Table 1.10, from DiFrancisco and Critelman (1984), refers to effects of nationality and education level on whether one follows politics regularly. Analyze these data.

17. Refer to Section 5.3.7. When Y is $N(\mu_i, \sigma^2)$, consider the comparison of $(\mu_1, \ldots, \mu_I)$ based on independent samples at the I categories of X. When approximately $\mu_i = \alpha + \beta x_i$, explain why the t or F test of H_0: $\beta = 0$ is more powerful than the one-way ANOVA F test. Describe a pattern for $\{\mu_i\}$ for which the ANOVA test would be more powerful.

Chapter 6

1. Table 1.11 classifies 174 poliomyelitis cases in Des Moines, Iowa by age of subject, paralytic status, and by whether the subject had been injected with the Salk vaccine.

   (a) Test the hypothesis that severity is independent of whether vaccinated, controlling for age.

Table 1.11:

		Paralysis	
Age	Salk Vaccine	No	Yes
0–4	Yes	20	14
	No	10	24
5–9	Yes	15	12
	No	3	15
10–14	Yes	3	2
	No	3	2
15–19	Yes	7	4
	No	1	6
20–39	Yes	12	3
	No	7	5
40+	Yes	1	0
	No	3	2

Source: Reprinted with permission, based on data from Chin et al. (1961).

(b) Use another procedure for testing this hypothesis, and compare results to those obtained in (a).

(c) Estimate the common odds ratio between severity and whether vaccinated, using (i) the Mantel–Haenszel estimator, (ii) the unconditional ML estimator. Interpret.

(d) If you have appropriate software, obtain the conditional ML estimator. Compare results to those in (c).

2. Refer to Table 6.13. An alternative scenario has P(Nonroutine Care) values equal to (0.50, 0.45, 0.40, 0.45, 0.25, 0.15). Calculate the noncentrality for the likelihood-ratio model-based test of NO partial association. Find the approximate powers for sample sizes 500 and 1000, for a 0.05-level test. How large a sample is needed to achieve power 0.90?

3. Refer to the example in Sec. 6.6.5. Suppose instead we used a linear logit model with equally-spaced scores for categories of N. Calculate the power for the test of conditional independence, with $df = 1$. Find the approximate power for sample size 1000. How do these compare to powers for the additive logit model? Explain the discrepancy.

Chapter 8

Table 1.12:

Logit	Intercept	Schooling	Experience	Race	Gender
$\log(\pi_B/\pi_M)$	1.056	−0.124	−0.015	0.700	1.252
$\log(\pi_C/\pi_M)$	−3.769	−0.001	−0.008	1.458	3.112
$\log(\pi_W/\pi_M)$	−3.305	0.225	0.003	1.762	−0.523
$\log(\pi_P/\pi_M)$	−5.959	0.429	0.008	0.976	0.656

Source: Reprinted with permission from P. Schmidt and R. P. Strauss, *Intern. Econ. Rev.*, 16, pp. 471-486 (1975).

Table 1.13:

		Breathing Test Results		
Age	Smoking Status	Normal	Borderline	Abnormal
<40	Never smoked	577	27	7
	Former smoker	192	20	3
	Current smoker	682	46	11
40–59	Never smoked	164	4	0
	Former smoker	145	15	7
	Current smoker	245	47	27

Source: From p. 21 of *Public Program Analysis* by R. N. Forthofer and R. G. Lehnen. Copyright©1981 by Lifetime Learning Publications, Belmont, CA 94002, a division of Wadsworth, Inc. Reprinted by permission of Van Nostrand Rienhold. All rights reserved.

1. Refer to the model discussed for Table 8.1 in Sec. 8.1.2. Show that for small alligators in Lake Hancock, the estimated probabilities of primary food choice (fish, invertebrates, reptile, bird, other) are (0.54, 0.09, 0.05, 0.07, 0.25).

2. Table 1.12 shows results of logit modeling of occupational attainment in the U.S. using S = years of schooling, E = labor market experience (calculated as age − years of schooling − 5), R = race (1 = white, 0 = black), and G = gender (1 = male, 0 = female). The categories of occupational attainment are professional (P), white collar (W), blue collar (B), craft (C), and menial (M).

   (a) Obtain parameter estimates for modeling $\log(\pi_W/\pi_B)$, and interpret.

   (b) Explain why the estimates in the Race column indicate that occupational groups are ordered (W, C, P, B, M) in terms of relative number of white workers, controlling for the other factors.

3. Table 1.13 displays associations among smoking status (S), breathing test results (B), and age (A) for workers in certain industrial plants. Treat B as a response.

   a. Specify a baseline-category logit model with additive factor effects of S and

Table 1.14:

Party Identification	Non-Protestors	Protestors
Strong Democrat	10	18
Weak Democrat	59	38
Leaning Democrat	41	22
Independent	26	7
Leaning Republican	44	10
Weak Republican	47	7
Strong Republican	29	2

Source: Reprinted with permission, based on data from M. K. Jennings, *Amer. Political Sci. Rev.*, **81**, 367–382 (1987).

A. This model has deviance $G^2 = 25.9$. Show that df $= 4$, and explain why this model treats all variables as nominal.

b. Treat B as ordinal and S as ordinal in terms of how recently one was a smoker, with scores $\{s_i\}$. Consider the model

$$\log \frac{P(B = k+1|S = i, A = j)}{P(B = k|S = i, A = j)} = \alpha_k + \beta_1 s_i + \beta_2 a_j + \beta_3 s_i a_j$$

with $a_1 = 0$ and $a_2 = 1$. Show that this assumes a linear effect of S with slope β_1 for age < 40 and $\beta_1 + \beta_3$ for age 40–59. Using $\{s_i = i\}, \hat{\beta}_1 = 0.115, \hat{\beta}_2 = 0.311$, and $\hat{\beta}_3 = 0.663$ (SE $= 0.164$). Interpret the interaction.

c. From part (b), for age 40–59 show that the estimated odds of abnormal rather than borderline breathing for current smokers are 2.18 times those for former smokers and $\exp(2 \times 0.778) = 4.74$ times those for never smokers. Explain why the squares of these values are estimated odds of abnormal rather than normal breathing.

4. Explain how the probability-based prior specification approach presented in Section 7.2.4 could be extended to multinomial response models.

5. Table 1.14 refers to subjects who graduated from high school in 1965. They were classified as protestors if they took part in at least one demonstration, protest march, or sit-in, and classified according to their party identification in 1982. Analyze the data, using response (**a**) party identification, (**b**) whether a protestor. Compare interpretations.

6. Table 1.15 cross-classifies assessment of cognitive impairment, Alzheimer's disease, and age. Analyze these data, treating (**a**) Alzheimer's disease, and (**b**) cognitive impairment, as the response variable.

Table 1.15:

Age	Alzheimer's Disease	Cognitive Impairment				
		Severe	Moderate	Mild	Borderline	Unaffected
65–69	Highly probable	1	1	0	0	0
	Probable	0	4	5	0	0
	Possible	0	4	11	9	0
	Unaffected	0	0	2	1	45
70–74	Highly probable	1	0	0	0	0
	Probable	1	8	3	0	0
	Possible	1	6	16	11	0
	Unaffected	0	1	3	3	40
75–79	Highly probable	1	4	0	0	0
	Probable	5	17	8	0	0
	Possible	1	5	17	14	0
	Unaffected	0	0	2	2	30
80–84	Highly probable	4	7	0	0	0
	Probable	2	15	9	0	0
	Possible	1	7	24	12	0
	Unaffected	0	0	0	3	28
85+	Highly probable	9	8	1	0	0
	Probable	17	16	8	0	0
	Possible	0	13	22	9	0
	Unaffected	0	0	2	2	11

Source: Dr. Laurel Smith, Department of Biostatistics, Harvard University.

Table 1.16:

Age	Survival Status[a]	Dose (rad)					
		Not in City	0–9	10–49	50–99	100–199	200+
0–9	LD	0	7	3	1	4	11
	NLD	5015	10752	2989	694	418	387
10–19	LD	5	4	6	1	3	6
	NLD	5973	11811	2620	771	792	820
20–34	LD	2	8	3	1	3	7
	NLD	5669	10828	2798	797	596	624
35–49	LD	3	19	4	2	1	10
	NLD	6158	12645	3566	972	694	608
50+	LD	3	7	3	2	2	6
	NLD	3695	9053	2415	655	393	289

Source: Reprinted from Sugiura and Otake (1974), by courtesy of Marcel Dekker, Inc.
[a]LD = death from leukemia, NLD = nondeath from leukemia.

7. Table 1.16, analyzed by Sugiura and Otake (1974) and Landis et al. (1978), shows the relationship between the deaths from leukemia during 1950–1970 and estimated radiation dosage from atomic bombing at the end of World War II. Subjects are stratified according to their age at time of bombing. Using the midpoint scores (0, 5, 30, 75, 150, 300) for the levels of dosage, Table 1.17 shows results of CMH tests between dosage and survival status.

   (a) Interpret the CMH statistics. Why are two of the statistics identical?

   (b) Interpret the effect by fitting a logit model with a linear effect of dose on the probability of death from leukemia.

8. Table 1.18 classifies 1398 children on tonsil size and on whether they are carriers of the virus *Streptococcus pyogenes*. Analyze these data.

9. Table 1.19 is from Bock and Jones (1968), one of the first books to present sophisticated models for categorical data. Using an ordinal scale, subjects indicated their preference for black olives. The sample consisted of independent samples of Armed Forces personnel selected from six combinations of urbanization (urban, rural) and location (NE, MW, SW). Analyze these data.

10. For an ordinal response, what are advantages and disadvantages of using baseline-category logits compared to cumulative logit modeling?

11. For an $I \times J$ contingency table, show that statistical independence is equivalent to

$$\text{logit}[P(Y \leq j \mid X = i)] = \alpha_j, \quad i = 1, \ldots, I, \ j = 1, \ldots, J-1.$$

Table 1.17:

Summary Statistics for survival by dose
Controlling for age

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	397.0091	<.0001
2	Row Mean Scores Differ	1	397.0091	<.0001
3	General Association	5	427.0519	<.0001

Table 1.18:

	Tonsil Size		
	Not Enlarged	Enlarged	Greatly Enlarged
Noncarriers	497	560	269
Carriers	19	29	24

Source: From M. C. Holmes and R. E. O. Williams, *J. Hyg.*, **52**, 165-179 (1954). Reprinted with permission from Cambridge University Press.

Table 1.19:

		Preference					
Urbanization	Location	A	B	C	D	E	F
Urban	MW	20	15	12	17	16	28
	NE	18	17	18	18	6	25
	SW	12	9	23	21	19	30
Rural	MW	30	22	21	17	8	12
	NE	23	18	20	18	10	15
	SW	11	9	26	19	17	24

Source: Reprinted with permission from Holden-Day (Bock and Jones 1968, p. 244).
Key: A, Dislike extremely; B, dislike very much or moderately; C, dislike slightly or neither like or dislike; D, like slightly; E, like moderate ly; F, like very much or like extremely.

Table 1.20:

President	Busing	Home		
		1	2	3
	1	41	65	0
1	2	71	157	1
	3	1	17	0
	1	2	5	0
2	2	3	44	0
	3	1	0	0
	1	0	3	1
3	2	0	10	0
	3	0	0	1

Source: 1991 General Social Survey, with categories 1 = yes, 2 = no, 3 = don't know

12. For the cumulative probit model $\Phi^{-1}[P(Y \leq j)] = \alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}$, explain why a 1-unit increase in x_i corresponds to a β_i standard deviation increase in the expected underlying latent response, controlling for other predictors.

13. Use a literature search to find an article that utilized discrete choice models for some application. Summarize results of the research. What justification was given for the particular model used?

14. Consider the model $\text{Link}[\omega_j(\boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}$, where $\omega_j(\boldsymbol{x})$ is (8.13). Explain why this model can be fitted separately for $j = 1, \ldots, J-1$. For the complementary log-log link and the common effect structure $\boldsymbol{\beta}_1^T = \cdots = \boldsymbol{\beta}_{J-1}^T$, show that this model is equivalent to one using the same link for cumulative probabilities (Läärä and Matthews 1985).

15. Suppose that model (8.5) holds for a $2 \times J$ table with $J > 2$, and let $x_2 - x_1 = 1$. Explain why you would expect local log odds ratios to typically be smaller in absolute value than the cumulative log odds ratio β.

Chapter 9

1. Table 1.20 is a $3 \times 3 \times 3$ table from a General Social Survey in which white subjects were asked: (B) "Do you favor busing of (Negro/Black) and white school children from one school district to another?", (P) "If your party nominated a (Negro/Black) for President, would you vote for him if he were qualified for the job?", (D) "During the last few years, has anyone in your family brought a friend who was a (Negro/Black) home for dinner?" The response scale for each item was (yes, no, don't know). Fit the model denoted by (*BD, BP, DP*).

a. Using the yes and no categories, estimate the conditional odds ratio for each pair of variables. Interpret.

b. Analyze the model's goodness of fit. Interpret.

c. Conduct inference for the BP conditional association using a Wald or likelihood-ratio confidence interval and test. Interpret.

2. Opposition to the legal availability of abortion is stronger among the religious than the nonreligious, and stronger among those with conservative sexual attitudes than those with more permissive attitudes. Does this imply that the religious are more likely than the nonreligious to have conservative sexual attitudes? Use sample tables in your answer.

3. Construct a loglinear model for a $2 \times 2 \times K$ table that has homogeneous XY association except for a different association in the first stratum. Derive the likelihood equations and show the first stratum has a perfect fit. Show residual $df = K - 2$.

4. Refer to Table 2.6. Consider the nested set $\{(DVP), (DP, VP, DV), (VP, DV), (P, DV), (D, V, P)\}$. Partition chi-squared to compare the four pairs, ensuring that the overall type I error probability for the four comparisons does not exceed $\alpha = 0.10$. Which model would you select, using a backward comparison starting with (DVP)? Show that the final model selected depends on the choice of nested set, by repeating the analysis with (DP, VP, DV), (DP, DV), (P, DV), (D, V, P).

5. Refer to Table 2.6. Let D = defendant's race, V = victims' race, and P = death penalty verdict. Fit the loglinear model (DV, DP, PV) and the corresponding logistic model, treating P as the response. Show the correspondence between parameter estimates and fit statistics.

a. Using the fitted values, estimate and interpret the odds ratio between D and P at each level of V. Note the common odds ratio property.

b. Calculate the marginal odds ratio between D and P, (**i**) using the fitted values, and (**ii**) using the sample data. Why are they equal? Contrast the odds ratio with part (a). Explain why Simpson's paradox occurs.

c. Fit the corresponding logit model, treating P as the response. Show the correspondence between parameter estimates and fit statistics.

d. Is there a simpler model that fits well? Interpret, and show the logit–loglinear connection.

6. For a three-way table with binary response Y, give the equivalent loglinear and logit models for which (a) Y is jointly independent of X and Z, (b) no interaction exists between X and Z in their effects on Y.

7. For a 3-way table, the general model between X and Y at level k of Z is

$$\log \mu_{ijk} = \lambda(k) + \lambda_i^X(k) + \lambda_j^Y(k) + \lambda_{ij}^{XY}(k).$$

Show that parameters in model (XYZ) satisfy $\lambda = [\sum \lambda(k)]/K$, $\lambda_i^X = [\sum_k \lambda_i^X(k)/K$, $\lambda_{ij}^{XY} = [\sum_k \lambda_{ij}^{XY}(k)]/K$, $\lambda_k^Z = \lambda(k) - \lambda$, $\lambda_{ik}^{XZ} = \lambda_i^X(k) - \lambda_i^X$, $\lambda_{ijk}^{XYZ} = \lambda_{ij}^{XY}(k) - \lambda_{ij}^{XY}$.

8. For loglinear models for a three-way table, define parameters such that

$$\lambda_1^X = \lambda_1^Y = \lambda_1^Z = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = \cdots = \lambda_{ij1}^{XYZ} = 0.$$

Show the two-factor terms are log odds ratios using the cell at the first level of each variable, and a three-factor term is a log of ratios of odds ratios. Illustrate for a $2 \times 2 \times 2$ table, showing $\lambda_{22}^{XY} = \log[\theta_{11(1)}]$ and $\lambda_{222}^{XYZ} = \log[\theta_{11(1)}/\theta_{11(2)}]$. Explain how to set up dummy variables so model fitting yields estimates having these constraints.

9. In a $2 \times 2 \times 2$ table, show $\theta_{11(1)} = \theta_{11(2)}$ implies $\theta_{1(1)1} = \theta_{1(2)1}$ and $\theta_{(1)11} = \theta_{(I)11}$.

10. When $\{n_i\}$ has a multinomial distribution with probabilities $\{\pi_i = \mu_i/(\sum_a \mu_a)\}$, show that the part of the log likelihood involving both the data and parameters is $\sum_i n_i \log(\mu_i)$, the same as for Poisson sampling.

11. Consider loglinear model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta I(ab)$$

where $I(ab) = 1$ in cell (a, b) and equals 0 otherwise.

(a) Find the likelihood equations, and note $\hat{\mu}_{ab} = n_{ab}$.

(b) Show that residual $df = IJ - I - J$.

(c) State an IPF algorithm for finding fitted values that satisfy the model. (Hint: Replace the entry in cell (a, b) by 0. Apply IPF for the independence model, with a starting value of 0 in cell (a, b), to obtain other fitted values.)

12. Suppose that X and Y are conditionally independent, given Z, and X and Z are marginally independent.

a. Show that X is jointly independent of Y and Z.

b. Show that if X and Z are conditionally (rather than marginally) independent, then X and Y are still marginally independent.

13. Refer to the baseline constraints $\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = 0$ for an $I \times J$ table. Show that when a $n_{ij} = 0$ for non-baseline categories and consequently $\hat{\lambda}_{ij} = -\infty$, some association parameters can have finite estimates. This is an advantage of these constraints compared to the sum equal to 0 constraints.

14. For a four-way table response Y, give the equivalent loglinear and logistic models that have interaction between A and B in their effects on Y, which is binary, and C has main effects.

15. For model (*XY, Z*), derive (**a**) minimal sufficient statistics, (**b**) likelihood equations, (**c**) fitted values, and (**d**) residual df for tests of fit.

16. Verify the df values shown in Table 9.14 for models (*XY, Z*), (*XY, YZ*), and (*XY, XZ, YZ*).

17. Verify that loglinear model (*GLS, GI, LI, IS*) implies logistic model (9.16). Show that the conditional log odds ratio for the effect of S on I equals $\beta_1^S - \beta_2^S$ in the logistic model and $\lambda_{11}^{IS} + \lambda_{22}^{IS} - \lambda_{12}^{IS} - \lambda_{21}^{IS}$ in the loglinear model.

18. Show that ML estimates for Poisson loglinear models are identical to those obtained after splitting the sample into several independent multinomial samples. Specifically, suppose a set of Poisson means $\{\mu_{ij}\}$ satisfy

$$\log \mu_{ij} = \alpha_i + \mathbf{x}_{ij}\boldsymbol{\beta}.$$

    Decompose the Poisson log likelihood so part refers to the row totals and part refers to the effect of conditioning on those totals.

19. Explain how IPF can standardize a three-way table so that each marginal two-way table has uniform cell frequencies.

20. Given target row totals $\{r_i > 0\}$ and column totals $\{c_j > 0\}$:

    a. Explain how to use IPF to adjust sample proportions $\{p_{ij}\}$ to have these totals but maintain the sample odds ratios.

    b. Show how to find cell proportions that have these totals and for which all local odds ratios equal $\theta > 0$. (*Hint:* Take initial values of 1.0 in all cells in the first row and in the first column. This determines all other initial cell entries such that all local odds ratios equal θ.)

    c. Explain how cell proportions are determined by the marginal proportions and the local odds ratios.

21. For logit model $\text{logit}[P(Y = 1|X = i, Z = k)] = \alpha + \beta i + \beta_k^Z$, $i = 0, 1$, is $\hat{\beta}$ the same as with model $\text{logit}[P(Y = 1|X = i)] = \alpha + \beta i$ for the table collapsed over Z? Explain.

22. Show that $G^2[(Y, XZ) \mid (XY, XZ)]$ is identical to $G^2[(X, Y)]$ for the XY marginal table.

23. Suppose $\{y_i\}$ are independent Poisson random variables with means $\{\mu_i\}, i = 1, \cdots, N$.

Table 1.21:

Gender	Income	Job Satisfaction: Very Dissatisfied	A Little Satisfied	Moderately Satisfied	Very Satisfied
Female	< 5000	1	3	11	2
	5000-15,000	2	3	17	3
	15,000-25,000	0	1	8	5
	$> 25{,}000$	0	2	4	2
Male	< 5000	1	1	2	1
	5000-15,000	0	3	5	1
	15,000-25,000	0	0	7	3
	$> 25{,}000$	0	1	9	6

Source: General Social Survey, 1991

(a) Let $z_i = (y_i - \mu_i)/(\mu_i)^{1/2}$. Show $sum_i \text{Var}(z_i) = N$.

(b) Let $e_i = (y_i - \hat{\mu}_i)/(\hat{\mu}_i)^{1/2}$, where $\{\hat{\mu}_i\}$ are fitted values for a model $\{\mu_i\}$ satisfy. Give a heuristic argument that $\sum_i \text{Var}(e_i)$ asymptotically equals *df* for testing the model fit.

24. Suppose we fit a multiplicative model M to a table, except for certain structural-zero cells where $\mu_a = 0$. The model is $\mu_i = E_i M$, where $E_a = 0$ for those cells and all other $E_i = 1$. Explain how to fit this using the model-with-offset representation. (In practice, E_a must be a very small constant, such as 10^{-8}, so that its logarithm exists. Some software allows the user to fit this model by assigning zero *weights* to certain cells.)

25. Suppose $n_{11+} = 0$. Do finite ML estimates exist of all parameters for loglinear model (XY, XZ, YZ)? Explain.

Chapter 10

1. Refer to Table 1.21. Fit the homogeneous linear-by-linear association model, and interpret. Test conditional independence between income (I) and job satisfaction (S), controlling for gender (G), using (**a**) that model, and (**b**) model (IS, IG, SG). Explain why the results are so different.

2. In a $2 \times 2 \times K$ table, the true XY conditional odds ratios are identical, but different from the XY marginal odds ratio. Is there three-factor interaction? Is Z conditionally independent of X or Y? Explain.

3. For a 3×3 table with ordered rows having scores $\{x_i\}$, identify all terms in the generalized loglinear model (10.10) for models (**a**) $\text{logit}[P(Y \leq j)] = \alpha_j + \beta x_i$, and (**b**) $\log[P(Y = j)/P(Y = 3)] = \alpha_j + \beta_j x_i$.

4. Refer to Section 10.2.3. Show that $G^2(M_j|M_{j-1})$ equals G^2 for independence in the 2×2 table comparing columns 1 through $j - 1$ with column j.

5. Consider the $L \times L$ model with $\{v_j = j\}$ replaced by $\{v_j = 2j\}$. Explain why $\hat{\beta}$ is halved but $\{\hat{\mu}_{ij}\}, \{\hat{\theta}_{ij}\}$, and G^2 are unchanged.

6. In a three-way table, refer to the homogeneous linear-by-linear XY association model.

   a. Show that the likelihood equations are, for all i, j, and k,

   $$\hat{\mu}_{i+k} = n_{i+k}, \quad \hat{\mu}_{+jk} = n_{+jk}, \quad \sum_i \sum_j u_i v_j \hat{\mu}_{ij+} = \sum_i \sum_j u_i v_j n_{ij+}.$$

   b. Show that residual df $= K(I - 1)(J - 1) - 1$.

   c. Show how the last likelihood equation above changes for heterogeneous linear-by-linear XY association. Explain why, in each stratum, the fitted XY correlation equals the sample correlation.

7. Construct a model having general XZ and YZ associations, but row effects for the XY association that are (**a**) homogeneous, and (**b**) heterogeneous across levels of Z. Interpret.

8. When $I = 2$, explain why the row effects model is equivalent to the linear-by-linear association model.

9. Express the RC model as a probability function for cell probabilities $\{\pi_{ij}\}$. Demonstrate the similarity of this function to the bivariate normal density having unit standard deviations. Show that β in the RC model corresponds to $\rho/(1 - \rho^2)$ for the bivariate normal density, where ρ is the correlation. See Goodman (1981a,b, 1985) and Becker (1989b).

10. For three dimensions, state a generalization of the RC model for the XY association that is a special case of (XY, XZ, YZ) and contains the homogeneous $L \times L$ model as a special case.

Chapter 11

1. For a poll of a random sample of 1600 voting-age British citizens, 944 indicated approval of the Prime Minister's performance in office. Six months later, of these same 1600 people, 880 indicated approval. Table 1.22 summarizes results.

Table 1.22:

First	Second Survey		
Survey	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Table 1.23:

Adult	Juvenile Court	
Court	Rearrest	No Rearrest
Rearrest	158	515
No Rearrest	290	1134

Source: Based on a study at the Univ. of Florida by D. Bishop, C. Frazier, L. Lanza-Kaduce, and L. Winner. Thanks to Dr. Larry Winner for showing me these data.

a. Compare the marginal proportions using a confidence interval.

b. Perform McNemar's test, and interpret.

c. Explain why inferences about the difference in approval ratings are more precise than if we had the same sample proportions but with independent samples of size 1600 each.

2. Table 1.23 refers to a sample of juveniles convicted of a felony in Florida in 1987. Matched pairs were formed using criteria such as age and the number of prior offenses. For each pair, one subject was handled in the juvenile court and the other was transferred to the adult court. The response of interest was whether the juvenile was rearrested by the end of 1988. Compare the true proportions rearrested for the adult and juvenile court assignments. Interpret.

3. Table 1.24 shows results when subjects of age between 18 and 29 were asked "Do you think a person has the right to end his or her own life if this person (1) has an incurable disease? (2) is tired of living an ready to die?"

a. Compare the marginal proportions using a confidence interval.

b. Perform McNemar's test, and interpret.

Table 1.24:

	Let Patient Die		
Suicide	Yes	No	Total
Yes	1097	90	1187
No	203	435	638

Source: 1994 General Social Survey

Table 1.25: Occupational Status for British Father–Son Pairs

Father's Status	Son's Status					Total
	1	2	3	4	5	
1	50	45	8	18	8	129
2	28	174	84	154	55	495
3	11	78	110	223	96	518
4	14	150	185	714	447	1510
5	3	42	72	320	411	848
Total	106	489	459	1429	1017	3500

Source: Reprinted with permission from Glass (1954).

c. Find the conditional ML estimate of β for model (11.7). Interpret.

4. For Table 1.25, use kappa to describe agreement. Interpret.

5. Refer to Table 11.8. Based on the reported standardized residuals, explain why the linear-by-linear association model might fit well. Fit it and describe the association.

6. A nonmodel-based ordinal measure of marginal heterogeneity is

$$\hat{\Delta} = \sum \sum_{a<b} p_{a+}p_{+b} - \sum \sum_{a>b} p_{a+}p_{+b}.$$

Show that $\hat{\Delta}$ estimates $\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1)$, where Y_1 has distribution $\{\pi_{a+}\}$ and Y_2 is independent from $\{\pi_{+b}\}$. Show that marginal homogeneity implies that $\Delta = 0$. Show that the estimated asymptotic variance of $\hat{\Delta}$ is

$$\left[\sum_a \sum_b \hat{\phi}_{ab}^2 p_{ab} - \left(\sum_a \sum_b \hat{\phi}_{ab} p_{ab}\right)^2\right] \Big/ n,$$

where $\hat{\phi}_{ab} = \hat{F}_{b1} + \hat{F}_{b-1,1} - \hat{F}_{a2} - \hat{F}_{a-1,2}$ with $\hat{F}_{a1} = (p_{1+} + \cdots + p_{a+})$ and $\hat{F}_{a2} = (p_{+1} + \cdots + p_{+a})$ (Agresti 2010, pp. 228–229).

7. For ordered scores $\{u_a\}$, let $\bar{y}_1 = \sum_a u_a p_{a+}$ and $\bar{y}_2 = \sum_a u_a p_{+a}$. Show that marginal homogeneity implies that $E(\overline{Y}_1) = E(\overline{Y}_2)$ and

$$\left[\sum_a \sum_b (u_a - u_b)^2 p_{ab} - (\bar{y}_1 - \bar{y}_2)^2\right] \Big/ n.$$

estimates $\text{var}(\bar{Y}_1 - \bar{Y}_2)$. Construct a test of marginal homogeneity (Bhapkar 1968).

8. A model for agreement on an ordinal response partitions beyond-chance agreement into that due to a baseline association and a main-diagonal increment (A. Agresti, *Biometrics* **44**: 539–548, 1988). For ordered scores $\{u_a\}$, the model is

$$\log \mu_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \beta u_a u_b + \delta I(a = b).$$

a. Show that this is a special case of quasi-symmetry and of quasi-uniform association.

b. For agreement odds (11.29), show that $\log \tau_{ab} = (u_b - u_a)^2\beta + 2\delta$. For unit-spaced scores, show the local odds ratios have $\log \theta_{ab} = \beta$ when none of the four cells falls on the main diagonal.

c. Find the likelihood equations and show that $\{\hat{\mu}_{ab}\}$ and $\{n_{ab}\}$ share the same marginal distributions, correlation, and prevalence of exact agreement.

d. For Table 11.8 using $\{u_a = a\}$, show that the model of linear-by-linear association has $G^2 = 4.8$ (df = 7), with $\hat{\delta} = 0.842$ (SE = 0.427) and $\hat{\beta} = 1.316$ (SE = 0.420). Interpret using $\hat{\tau}_{a,a+1}$ and $\hat{\theta}_{ab}$ for $|a - b| > 1$.

9. For ordered classifications, when symmetry does not hold, often either $\pi_{ab} > \pi_{ba}$ for all $a < b$, or $\pi_{ab} < \pi_{ba}$ for all $a < b$. A generalization of symmetry with this property is the conditional symmetry model mentioned in Note 11.2.

a. For Table 11.6 on opinions about premarital and extramarital sex, show that the conditional symmetry model has $\hat{\tau} = -4.130$ ($SE = 0.451$). Interpret.

b. Show that it has the loglinear representation

$$\log \mu_{ab} = \lambda_{\min(a,b),\max(a,b)} + \tau I(a < b),$$

where $I(\cdot)$ is an indicator.

c. Show that the likelihood equations are

$$\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \quad \text{for all } a \le b, \quad \sum\sum\nolimits_{a<b} \hat{\mu}_{ab} = \sum\sum\nolimits_{a<b} n_{ab}.$$

d. Show that $\hat{\tau} = \log\left[\left(\sum\sum_{a<b} n_{ab}\right) / \left(\sum\sum_{a>b} n_{ab}\right)\right]$, $\hat{\mu}_{aa} = n_{aa}, a = 1, \ldots, I, \hat{\mu}_{ab} = \exp[\hat{\tau} I(a < b)](n_{ab} + n_{ba})/[\exp(\hat{\tau}) + 1]$ for $a \neq b$.

e. Show that the estimated asymptotic variance of $\hat{\tau}$ is

$$\left(\sum\sum_{a<b} n_{ab}\right)^{-1} + \left(\sum\sum_{a>b} n_{ab}\right)^{-1}.$$

f. Show that residual df $= (I + 1)(I - 2)/2$.

g. Show that conditional symmetry + marginal homogeneity = symmetry. Explain why G^2(S|CS) tests marginal homogeneity (df = 1). When the model holds G^2(S|CS) is more powerful asymptotically than G^2(S|QS). Why?

h. Explain how the conditional symmetry model is a special case of *diagonals-parameter symmetry*,

$$\log(\pi_{ab}/\pi_{ba}) = \tau_{b-a}, \quad a < b.$$

See Goodman (1979b, 1985) and Hout et al. (1987).

Table 1.26:

Affiliation	Religious Affiliation Now			
at Age 16	1	2	3	4
1	863	30	1	52
2	50	320	0	33
3	1	1	28	1
4	27	8	0	33

Source: 1991 General Social Survey

10. Another ordinal model generalizes quasi-independence. Let $\{u_a\}$ be ordered scores. The model

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \beta u_a u_b + \delta_a I(a = b)$$

permits linear-by-linear association off the main diagonal. It is a special case of quasi-symmetry, and quasi-independence is the special case $\beta = 0$. For equal-interval scores, it implies uniform local association, given that responses differ. Goodman (1979a) called it *quasi-uniform association*. For Table 11.6 on opinions about premarital and extramarital sex, show that the quasi-uniform association model has $\hat{\beta} = 0.632$ $(SE = 0.106)$. Explain why, off the main diagonal, the estimated local odds ratio equals 1.88.

11. Table 1.26 reports subjects' religious affiliation in 1991 and when their age was 16, for categories (1) Protestant, (2) Catholic, (3) Jewish, (4) None or Other.

    (a) The symmetry model has $G^2 = 32.2$ $(df = 6)$. Interpret, and use residuals to analyze transition patterns.

    (b) The quasi-symmetry model has $G^2 = 2.0$ $(df = 3)$. Interpret.

    (c) Test marginal homogeneity. Show the small P-value mainly reflects the large sample size, a small decrease in the proportion classified Catholic, and an increase in the proportion classified None or Other.

    (d) Fit the quasi-independence model, and interpret.

12. Table 1.27, from Breslow (1982), compares 80 esophageal cancer patients with 80 matched control subjects. The response is the number of beverages reported drunk at "burning hot" temperatures. In analyzing whether cases tended to drink more beverages burning hot than did controls, use X^2 to check model fit, since most cell counts are small.

    (a) Fit the symmetry model, and explain how $n_{ab} \leq n_{ba}$ whenever $a < b$ contributes to the lack of fit $(X^2 = 15.1, df = 6)$.

    (b) Show the quasi-symmetry model has $X^2 = 2.5$ $(df = 3)$.

    (c) Fit an ordinal model, interpret the effect, and use it to test marginal homogeneity.

Table 1.27:

	Control			
Case	0	1	2	3
0	31	5	5	0
1	12	1	0	0
2	14	1	2	1
3	6	1	1	0

Source: Reprinted with permission from the Biometric Society; data from Breslow (1982).

Table 1.28:

	Left Eye Grade			
Right Eye Grade	Best	Second	Third	Worst
Best	1520	266	124	66
Second	234	1512	432	78
Third	117	362	1772	205
Worst	36	82	179	492

Source: Reprinted with permission from the Biometrika Trustees (Stuart 1955).

13. In the previous exercise, treating the response as continuous, use a normal paired-difference procedure to compare cases and controls on the mean number of beverages drunk burning hot. Compare the results to an ordinal test of marginal homogeneity. List assumptions on which each procedure is based.

14. Table 1.28 describes unaided distance vision for a sample of women. Analyze these data.

15. A sample of married couples indicate their candidate preference in a presidential election. Table 1.29 reports the results. Analyze these data.

16. A wildlife biologist wants to estimate the number of alligators in Lake Lochloosa, Florida. She catches n_{1+} alligators, tags them, and releases them back into the lake. Two weeks later, she catches a second sample of n_{+1} alligators, of which n_{11} were also in the first sample. She cannot observe n_{22}, the number not caught either time, and hence the population size N. If whether an alligator is captured in the

Table 1.29:

Husband's	Wife's Preference	
Preference	Democrat	Republican
Democrat	200	25
Republican	75	200

Table 1.30:

		Control		
Cataract Case	Always or Almost Always	Frequently	Occasionally	Never
Always or almost always	29	3	3	4
Frequently	5	0	1	1
Occasionally	9	0	2	0
Never	7	3	1	0

Source: J. M. Dolezal *et al., Amer. J. Epidemiol., 129*: 559-568 (1989).

Table 1.31:

	Loser				
Winner	Edberg	Lendl	Agassi	Sampras	Becker
Edberg	–	5	3	2	4
Lendl	4	–	3	1	2
Agassi	2	0	–	1	3
Sampras	0	1	2	–	0
Becker	6	4	2	1	– –

Source: Reprinted with permission from World Tennis magazine.

second sample is independent of whether it was captured in the first sample, argue that a reasonable estimator is $\hat{N} = n_{1+}n_{+1}/n_{11}$ (Sekar and Deming 1949).

17. Show the ML fit of loglinear model (W, XYZ) for the $6 \times I^3$ table with entries $\{y^*_{1ijk} = y_{ijk},\ y^*_{2ijk} = y_{ikj},\ y^*_{3ijk} = y_{jik},\ y^*_{4ijk} = y_{jki},\ y^*_{5ijk} = y_{kij},\ y^*_{6ijk} = y_{kji}\}$ has entries $\{\hat{\mu}^*_{hijk}\}$ related to the ML fit $\{\hat{\mu}_{ijk}\}$ for the complete symmetry model by $\{\hat{\mu}_{ijk} = \hat{\mu}^*_{1ijk}\}$.

18. Construct a loglinear model for an I^3 table having the following quasi-independence interpretation: Conditional on the event that the three responses are completely different, the responses are mutually independent. Find the residual df.

19. Table 1.30 refers to a case-control study investigating a possible relationship between cataracts and the use of head coverings during the summer. Each case reporting to a clinic for care for a cataract was matched with a control of the same sex and similar age not having a cataract. The row and column categories refer to the frequency with which the subject used head coverings. Analyze these data.

20. Table 1.31 refers to matches among five men tennis players during 1989-1990. Analyze these data.

21. Table 1.32 reports respondents' current region of residence and region of residence at age 16. Fit the quasi-independence model. Describe lack of fit. What can you

Table 1.32:

Residence at Age 16	Residence in 1991			
	Northeast	Midwest	South	West
Northeast	245	16	40	20
Midwest	12	333	31	51
South	14	31	321	16
West	3	51	12	309

Source: 1991 General Social Survey

Table 1.33:

Mother's Education	Father's Education			
	8th Grade or less	Part High School	High School	College
8th Grade or less	81	3	9	11
Part High School	14	8	9	6
High School	43	7	43	18
College	21	6	24	87

Source: Reprinted with permission from E. J. Mullins and P. Sites, *Amer. Sociol. Rev.*, **49**: 672-685 (1984).

say about the numbers of people who moved from the Northeast to the South and from the Midwest to the West, relative to what quasi independence predicts?

22. Table 1.33 relates mother's education to father's education for a sample of eminent black Americans (defined as persons having biographical sketch in the publication, *Who's Who Among Black Americans*). Analyze these data.

23. For a longitudinal binary matched-pairs study, data are available for some subjects at both times, for others only at the first time or the second time. Of n subjects observed both times, let p_{ab} denote the proportion having outcome a at time 1 and b at time 2. Of n_t subjects observed only at time t, let q_t denote the proportion making the first outcome. Treat n, n_1, and n_2 as fixed, and let $a = n/(n+n_1)$, $b = n/(n+n_2)$, and $\mathbf{p}^T = (p_{11}, p_{12}, p_{21}, q_1, q_2)$.

    (a) Of all subjects observed at time t, let P_t denote the proportion having the first outcome. Show that $P_t = \mathbf{d}_t^T\mathbf{p}$, with $\mathbf{d}_1^T = (a, a, 0, 1-a, 0)$ and $\mathbf{d}_2^T = (b, 0, b, 0, 1-b)$. Thus, $\widehat{Var}(P_i) = \mathbf{d}_i^T\mathbf{S}\mathbf{d}_i$, and $\widehat{Var}(P_1 - P_2) = (\mathbf{d}_1 - \mathbf{d}_2)^T\mathbf{S}(\mathbf{d}_1 - \mathbf{d}_2)$.

    (b) Table 1.34 is from a study at the Univ. of Florida about drug use in an elderly population. Subjects were asked whether they took tranquilizers. Some were interviewed in 1979, some in 1985, and others both times. Find P_1 and P_2. Assuming $E(p_{1+}) = E(q_1) = \pi_1$ and $E(p_{1+}) = E(q_2) = \pi_2$, construct a 95% confidence interval for $\pi_1 - \pi_2$. (This approach is reasonable when data are *missing completely at random*.)

Table 1.34:

	1985		
Take Drug	Yes	No	Not Sampled
1979			
Yes	175	190	230
No	139	1518	982
Not Sampled	64	595	

Source: Mary Moore.

24. For an $I \times I$ table $\{n_{ab}\}$, construct the $I \times I \times 2$ tables $\{n_{ab1} = n_{ab},\ n_{ab2} = n_{ba}\}$ and $\{\mu_{ab1} = \mu_{ab},\ \mu_{ab2} = \mu_{ba}\}$.

    (a) If quasi symmetry holds for $\{\mu_{ab}\}$, show $\theta_{ab(1)}/\theta_{ab(2)} = 1$ for $\{\mu_{abc}\}$, for all a and b.

    (b) Show that likelihood equations for the quasi-symmetry model for $\{\mu_{ab}\}$ correspond to likelihood equations for loglinear model $(XY,\ XZ,\ YZ)$ for $\{\mu_{abc}\}$.

    (c) Show that $\{\hat{\mu}_{ab}\}$ for the quasi-symmetry model are identical to $\{\hat{\mu}_{ab1}\}$ for model $(XY,\ XZ,\ YZ)$ fitted to $\{n_{abc}\}$ (Bishop et al. 1975, pp. 289–290).

    (d) Show that model $\log \mu_{abc} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY}$ for $\{\mu_{abc}\}$ corresponds to symmetry for $\{\mu_{ab}\}$.

25. Consider complete symmetry for $T = 3$ matched observations. Show that

$$\hat{\mu}_{abc} = (n_{abc} + n_{acb} + n_{bac} + n_{bca} + n_{cab} + n_{cba})/6.$$

    How does this simplify for $\hat{\mu}_{aaa}$, $a = 1, \ldots, I$?

Chapter 12

1. Analyze the cereal diet and cholesterol study of text Exercise 8.17 with marginal models.

2. Use GEE with the cumulative logit model for marginal distributions to model the esophageal cancer data in Table 1.27. Interpret the marginal effect, and show how to use the model to test marginal homogeneity.

3. Suppose that loglinear model $(Y_0, Y_1, \ldots, Y_T)$ holds. Is this a Markov chain?

4. For a univariate response, how is quasi-likelihood (QL) inference different from ML inference? When are they equivalent?

Chapter 13

1. Refer to the voting data in Table 13.1 and the estimated effects given for the random effects model in Section 13.1.3 and the marginal model in Section 13.2.3. Which model and which estimate seems more appropriate to you for describing these data? Why? (*Note*: There is no "correct" answer here.)

2. For the insomnia example in Section 13.4.2, according to SAS the maximized log likelihood equals -593.0, compared to -621.0 for the simpler model forcing $\sigma = 0$. Compare models, using either a likelihood-ratio test or AIC. What do you conclude?

3. The analyses in Section 13.3.2 comparing opinions on some topic extend to ordinal responses. Using an ordinal random effects model, analyze the 4^3 table in Agresti (1993).

4. Refer to the crossover study in Exercise 12.6. Kenward and Jones (1991) reported results using the ordinal response scale (none, moderate, complete) for relief. Explain how to formulate an ordinal logit random effects model for these data.

5. For ordinal square $I \times I$ tables of counts $\{n_{ab}\}$, model (13.2) for binary matched-pairs responses (Y_{i1}, Y_{i2}) for subject i extends to

$$\text{logit}[P(Y_{it} \leq j | u_i)] = \alpha_j + \beta x_t + u_i$$

with $\{u_i\}$ independent $N(0, \sigma^2)$ variates and $x_1 = 0$ and $x_2 = 1$.

   a. Explain how to interpret β, and compare to the interpretation of β in the corresponding marginal model.

   b. This model implies model (13.2) for each 2×2 collapsing that combines categories 1 through j for one outcome and categories $j+1$ through I for the other. Use the form of the conditional ML (or random effects ML) estimator for binary matched pairs to explain why

$$\log\left[\left(\sum_{a>j}\sum_{b<j} n_{ab}\right) \Big/ \left(\sum_{a<j}\sum_{b>j} n_{ab}\right)\right]$$

   is a consistent estimator of β.

   c. Treat these $(I-1)$ collapsed 2×2 tables naively as if they are independent samples. Show that adding the numerators and adding the denominators of the separate estimates of e^β motivates the summary estimator of β,

$$\tilde{\beta} = \log\left\{\left[\sum_{a>b}(a-b)n_{ab}\right] \Big/ \left[\sum_{b>a}(b-a)n_{ab}\right]\right\}.$$

   Explain why $\tilde{\beta}$ is consistent for β even recognizing the actual dependence.

d. A standard error for $\tilde{\beta}$ that treats the collapsed tables in part (c) as independent is inappropriate. Treating $\{n_{ab}\}$ as a multinomial sample, show that an estimated asymptotic variance of $\tilde{\beta}$ is (Agresti and Lang 1993a)

$$\left\{\sum_{b>a}(b-a)^2 n_{ab} \Big/ \left[\sum_{b>a}(b-a)n_{ab}\right]^2\right\} + \left\{\sum_{a>b}(a-b)^2 n_{ab} \Big/ \left[\sum_{a>b}(a-b)n_{ab}\right]^2\right\}.$$

6. You are a statistical consultant asked to conduct a meta-analysis of Table 4 in B. Efron, *Statistical Science* **13:** 95–122 (1998), which shows 2×2 tables from a clinical trial in 41 cities. Analyze, and write a report summarizing your analysis.

7. Refer to the example in Section 13.4.6. Hedeker also considered an alternative coding scheme whereby one logit contrasts independent vs. community housing, giving estimated part of the linear predictor involving the certificate effect $0.13c + 1.07(c \times t_1) + 1.23(c \times t_2) + 0.63(c \times t_3)$, and the other contrasts (independent and community) categories combined vs. street/shelter, giving linear predictor $0.43c + 0.62(c \times t_1) - 0.46(c \times t_2) - 0.22(c \times t_3)$. Interpret.

8. A data set from a General Social Survey on subjects' opinions on four items (the environment, health, law enforcement, education) related to whether they believed government spending on each item should increase, stay the same, or decrease. Subjects were also classified by their gender and race. For subject i, let $G_i = 1$ for females and 0 for males, let $R_{1i} = 1$ for whites and 0 otherwise, $R_{2i} = 1$ for blacks and 0 otherwise, and $R_{1i} = R_{2i} = 0$ for the other category of race. Let y_{it} denote the response for subject i on spending item t, where outcomes (1, 2, 3) represent (increase, stay the same, decrease).

   a. With constraint $\beta_4 = 0$, the random-intercept model

$$\text{logit}[P(Y_{it} \le j | u_i)] = \alpha_j + \beta_t + \beta_g G_i + \beta_{r1} R_{i1} + \beta_{r2} R_{2i} + u_i, \quad j = 1, 2,$$

   has $\hat{\beta}_1 = -0.55, \hat{\beta}_2 = -0.60, \hat{\beta}_3 = -0.49$, with $\hat{\sigma} = 1.03$. These estimates are greater than five standard errors in absolute value. Interpret.

   b. Table 1.35 shows results with a race-by-item interaction. Interpret.

9. Use a GLMM to analyze the esophageal cancer data in Table 1.27.

10. Use a logistic-normal model to analyze the data in Larsen et al. (2000).

Table 1.35:

Variable	Estimate	SE
Intercept-1	1.065	0.391
Intercept-2	1.919	0.051
Gender	0.409	0.088
Race1-w	−0.055	0.397
Race2-b	0.434	0.452
Item1-envir	−0.357	0.539
Item2-health	−0.319	0.493
Item3-crime	−0.585	0.480
Race1 * Item1	−0.170	0.549
Race1 * Item2	−0.387	0.503
Race1 * Item3	0.197	0.491
Race2 * Item1	−0.452	0.606
Race2 * Item2	0.454	0.598
Race2 * Item3	−0.518	0.560

Note: Coding 0 for item 4 (education) and race 3 (other).

Chapter 14

1. Refer to the data in Crowder (1978). Analyze these data using at least two different approaches for overdispersed binary data. Compare results and interpret.

2. Analyze Table 9.1 using a latent class model with $q = 2$.

   a. For a subject in the first latent class, estimate the probability of having used (**i**) marijuana, (**ii**) alcohol, (**iii**) cigarettes, (**iv**) all three, and (**v**) none of them.

   b. Estimate the probability a subject is in the first latent class, given they have used (**i**) marijuana, (**ii**) alcohol, (**iii**) cigarettes, (**iv**) all three, and (**v**) none of them.

3. Use or write software to replicate the analyses of the opinions about abortion data in Section 14.2.2 using (**a**) nonparametric random effects fitting of logit model (14.3), and (**b**) the quasi-symmetry model.

4. For the toxicity study of Table 1.36, collapsing to a binary response, consider linear logit models for the probability a fetus is normal.

   a. Does the ordinary binomial model show evidence of overdispersion?

   b. Fit the linear logit model using the quasi-likelihood approach with inflated binomial variance. How do the standard errors change?

Table 1.36: Response Counts for 94 Litters of Mice on (Number Dead, Number Malformed, Number Normal)

Dose = 0.00 g/kg	Dose = 0.75 g/kg	Dose = 1.50 g/kg	Dose = 3.00 g/kg
(1, 0, 7), (0, 0, 14)	(0, 3, 7), (1, 3, 11)	(0, 8, 2), (0, 6, 5)	(0, 4, 3), (1, 9, 1)
(0, 0, 13), (0, 0, 10)	(0, 2, 9), (0, 0, 12)	(0, 5, 7), (0, 11, 2)	(0, 4, 8), (1, 11, 0)
(0, 1, 15), (1, 0, 14)	(0, 1, 11), (0, 3, 10)	(1, 6, 3), (0, 7, 6)	(0, 7, 3), (0, 9, 1)
(1, 0, 10), (0, 0, 12)	(0, 0, 15), (0, 0, 11)	(0, 0, 1), (0, 3, 8)	(0, 3, 1), (0, 7, 0)
(0, 0, 11), (0, 0, 8)	(2, 0, 8), (0, 1, 10)	(0, 8, 3), (0, 2, 12)	(0, 1, 3), (0, 12, 0)
(1, 0, 6), (0, 0, 15)	(0, 0, 10), (0, 1, 13)	(0, 1, 12), (0, 10, 5)	(2, 12, 0), (0, 11, 3)
(0, 0, 12), (0, 0, 12)	(0, 1, 9), (0, 0, 14)	(0, 5, 6), (0, 1, 11)	(0, 5, 6), (0, 4, 8)
(0, 0, 13), (0, 0, 10)	(1, 1, 11), (0, 1, 9)	(0, 3, 10), (0, 0, 13)	(0, 5, 7), (2, 3, 9)
(0, 0, 10), (1, 0, 11)	(0, 1, 10), (0, 0, 15)	(0, 6, 1), (0, 2, 6)	(0, 9, 1), (0, 0, 9)
(0, 0, 12), (0, 0, 13)	(0, 0, 15), (0, 3, 10)	(0, 1, 2), (0, 0, 7)	(0, 5, 4), (0, 2, 5)
(1, 0, 14), (0, 0, 13)	(0, 2, 5), (0, 1, 11)	(0, 4, 6), (0, 0, 12)	(1, 3, 9), (0, 2, 5)
(0, 0, 13), (1, 0, 14)	(0, 1, 6), (1, 1, 8)		(0, 1, 11)
(0, 0, 14)			

Source: Study described in article by C. J. Price, C. A Kimmel, R. W. Tyl, and M. C. Marr, *Toxicol. and Appl. Pharmacol.* **81**, 113-127 (1985).

c. Fit the linear logit model using quasi-likelihood with beta-binomial variance. Interpret and compare with previous results.

d. Fit the linear logit model using a GEE approach with exchangeable working correlation among fetuses in the same litter. Interpret and compare with previous results, including comparing the estimated GEE correlation with the estimate $\hat{\rho}$ from part (c).

e. Fit the linear logit GLMM after adding a litter-specific normal random effect. Interpret and compare with previous results.

5. Refer to Problems 13.14 and 13.15. Using width and qualitative color as predictors, fit a (**a**) negative binomial GLM, and (**b**) Poisson GLMM, checking for interaction and interpreting the final model.

6. Refer to Table 14.6. For those with race classified as "other," the sample counts for (0, 1, 2, 3, 4, 5, 6) homicides were (55, 5, 1, 0, 1, 0, 0). Fit an appropriate model simultaneously to these data and those for white and black race categories. Interpret by making pairwise comparisons of the three pairs of means.

7. Conduct the analyses of Exercise 4.7 on defects in the fabrication of computer chips, but use a negative binomial GLM. Compare results to those for the Poisson GLM. Indicate why results are similar.

8. Conduct a latent class analysis of the data in Espeland and Handelman (1989).

9. Refer to the teratology study in Liang and Hanfelt (1994). Analyze these data using at least two different approaches for overdispersed binary data. Compare results and interpret.

10. Let $\mathbf{\Pi}$ denote an $I \times J$ matrix of cell probabilities for the joint distribution of X and Y. Suppose that there exist $I \times 1$ column vectors $\boldsymbol{\pi}_{1k}$ and $J \times 1$ column vectors $\boldsymbol{\pi}_{2k}$ of probabilities, $k = 1, \ldots, q$, and a set of probabilities $\{\rho_k\}$ such that

$$\mathbf{\Pi} = \sum_{k=1}^{q} \rho_k \boldsymbol{\pi}_{1k} \boldsymbol{\pi}_{2k}^T.$$

Explain why this implies that there is a latent variable Z such that X and Y are conditionally independent, given Z.

11. Refer to Exercises 12.6 and 13.6. Let $\mu_k(a, b, c)$ denote the expected frequency of outcomes (a, b, c) for treatments (A, B, C) under treatment sequence k, where outcome 1 = relief and 0 = nonrelief. With a non-parametric random effects approach, show that one can estimate treatment effects in model (13.22) by fitting the quasi-symmetry model

$$\log \mu_k(a, b, c) = a\beta_A + b\beta_B + c\beta_C + \lambda_k(a, b, c),$$

where $\lambda_k(a, b, c) = \lambda_k(a, c, b) = \lambda_k(b, a, c) = \lambda_k(b, c, a) = \lambda_k(c, a, b) = \lambda_k(c, b, a)$. Fit the model, and show that $\hat{\beta}_B - \hat{\beta}_A = 1.64$ (SE = 0.34), $\hat{\beta}_C - \hat{\beta}_A = 2.23$ (SE = 0.39), $\hat{\beta}_C - \hat{\beta}_B = 0.59$ (SE = 0.39). Interpret.

12. When y_i is the sum of n_i binary responses each having mean μ_i, refer to the quasi-likelihood approach with $v(\mu_i) = \phi n_i \mu_i (1 - \mu_i)$. Explain why this variance function has a structural problem, with only $\phi = 1$ making sense when $n_i = 1$.

13. The negative binomial distribution is unimodal with a mode at the integer part of $\mu(k-1)/k$ (Johnson et al. 2005, p. 217). Show that the mode is 0 when $\mu \leq 1$, and that when $\mu > 1$ the mode is still 0 if $k < \mu/(\mu - 1)$. (This gives greater scope than the Poisson, for which the mode equals the integer part of the mean.)

Chapter 16

1. If $c > 0$, show that $n^{-c} = o(1)$ as $n \to \infty$.

2. Refer to Section 16.1.4, with $\mathbf{\Sigma} = \mathbf{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$ the covariance matrix of $\sqrt{n}\,(p_1, \ldots, p_{N-1})^T$. Let

$$Z = \begin{cases} c_i & \text{with probability } \pi_i, \quad i = 1, \ldots, N-1 \\ 0 & \text{with probability } \pi_N \end{cases}$$

and let $\boldsymbol{c} = (c_1, \ldots, c_{N-1})^T$.

a. Show that $E(Z) = \boldsymbol{c}^T\boldsymbol{\pi}$, $E(Z^2) = \boldsymbol{c}^T\mathbf{Diag}(\boldsymbol{\pi})\boldsymbol{c}$, and $\text{var}(Z) = \boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}$.

b. Suppose that at least one $c_i \neq 0$, and all $\pi_i > 0$. Show $\text{var}(Z) > 0$, and deduce that $\boldsymbol{\Sigma}$ is positive definite.

c. If $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^T$, so $\boldsymbol{\Sigma}$ is $N \times N$, prove that $\boldsymbol{\Sigma}$ is not positive definite.

3. Suppose a sample measure has the same form when expressed in terms of $\{p_{ij}\}$ as it does when expressed in terms of $\{n_{ij}\}$ (i.e., substituting n_{ij} for p_{ij} in its formula gives the same value, as is the case for measures such as gamma and odds ratio measures). Show $\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$ in the two-way version of (3.9). (See Fleiss 1982.)

4. Explain why $\{n_{+j}\}$ are sufficient for $\{\pi_{+j}\}$ in (16.26).

5. For loglinear model $(XY,\ XZ,\ YZ)$, ML estimates of $\{\mu_{ijk}\}$ and hence the X^2 and G^2 statistics are not direct. Alternative approaches may yield direct analyses. For $2 \times 2 \times 2$ tables, find a statistic for testing the hypothesis of no three-factor interaction, using the delta method with the asymptotic normality of $\log \hat{\theta}_{111}$, where

$$\hat{\theta}_{111} = \frac{p_{111}\, p_{221}/p_{121}\, p_{211}}{p_{112}\, p_{222}/p_{122}\, p_{212}}.$$

6. Justify the use of *estimated* asymptotic covariance matrices. For instance, for large samples, why is $\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}$ close to $\boldsymbol{A}^T\boldsymbol{A}$?

7. Motivate partitioning (3.15) by showing that the multivariate hypergeometric distribution (16.27) for $\{n_{ij}\}$ factors as the product of hypergeometric distributions for the separate component tables (Lancaster, 1949a).

8. For a given set of parameter constraints, show that weak identifiability conditions hold for the independence loglinear model for a two-way table; that is, when two values for $\boldsymbol{\theta}$ give the same $\boldsymbol{\pi}$, those parameter vectors must be identical.

9. Consider Table 9, which cross-classifies level of smoking and myocardial infarction for a sample of young women in a case–control study.

a. Given the marginal counts, explain why the only table having greater evidence of positive association between smoking and myocardial infarction has counts (25,26,11) for row 1 and (0,0,4) in row 2.

b. Conditional on both sets of margins, (i) find the null probability of the observed table and this more extreme table [based on formula (16.27)], (ii) show that the exact $P(X^2 \geq X_o^2) = P(X^2 \geq 6.96) = 0.052$.

10. For WLS with $\boldsymbol{F}(\boldsymbol{\pi}) = \boldsymbol{C}[\log(\boldsymbol{A}\boldsymbol{\pi})]$, show that $\boldsymbol{Q} = \boldsymbol{C}[\mathbf{Diag}(\boldsymbol{A}\boldsymbol{\pi})]^{-1}\boldsymbol{A}$.

Table 1.37:

	Smoking Level (cigarettes/day)		
	0	1–24	> 25
Control	25	25	12
Myocardial infarction	0	1	3

Source: Reprinted with permission, based on Table 5 in S. Shapiro et al., *Lancet* 743–746 (1979).

Table 1.38: Coalminers Classified by Breathlessness, Wheeze, and Age

	Breathlessness				
	Yes		No		
Age	Wheeze Yes	Wheeze No	Wheeze Yes	Wheeze No	Std. Pearson Residual[a]
20–24	9	7	95	1841	0.75
25–29	23	9	105	1654	2.20
30–34	54	19	177	1863	2.10
35–39	121	48	257	2357	1.77
40–44	169	54	273	1778	1.13
45–49	269	88	324	1712	−0.42
50–54	404	117	245	1324	0.81
55–59	406	152	225	967	−3.65
60–64	372	106	132	526	−1.44

Source: Reprinted with permission from Ashford and Sowden (1970).
[a]Residual refers to yes-yes and no-no cells; reverse sign for yes–no and no–yes cells.

11. Show how to use WLS to fit a linear probability model to an $I \times 2$ contingency table with row scores $\{u_i\}$. Identify the number of multinomial samples, the number of response categories, the response functions $\boldsymbol{F}$, the model matrix $\boldsymbol{X}$, the parameter vector $\boldsymbol{\beta}$, and the estimated covariance matrix $\hat{\boldsymbol{V}}_F$.

12. Use WLS to conduct the longitudinal analysis of depression in Sec. 12.2.1. Using software (e.g., SAS: PROC CATMOD), obtain WLS estimates and standard errors and compare to the ML results.

13. Refer to the previous problem. Using these data, describe the differences between (**a**) WLS and ML, (**b**) WLS and GEE methods for marginal models with multivariate categorical response data.

14. Refer to Table 1.38. Consider the model that simultaneously assumes a linear trend for the conditional log odds ratio between wheeze and breathlessness (given age) as well as linear logit relationships for the marginal effects of age on breathlessness and on wheeze.

    (a) Specify $\boldsymbol{C}$, $\boldsymbol{A}$, and $\boldsymbol{X}$ for which this model has form $\boldsymbol{C}\log(\boldsymbol{A}\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}$.

(b) Using software, fit the model and interpret estimates.

15. Show that the loglinear model of homogenous association for an $I \times J \times K$ table is specified by $(I-1)(J-1)(K-1)$ constraint equations, such as

$$\log[(\pi_{ijk}\pi_{i+1,j+1,k})/(\pi_{i+1,jk}\pi_{i,j+1,k})] \\ -\log[(\pi_{ij,k+1}\pi_{i+1,j+1,k+1})/\pi_{i+1,j,k+1}\pi_{i,j+1,k+1})] = 0.$$

For it, are WLS estimates the same as minimum modified chi-squared estimates?