AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS, 3rd ed.

EXTRA EXERCISES

Chapter 1

1. Which scale of measurement is most appropriate for the following variables — nominal, or ordinal?

   (a) Political party affiliation (Democrat, Republican, Independent)

   (b) Appraisal of a company's inventory level (too low, about right, too high)

2. When the observation falls at the boundary of the sample space, explain why Wald methods of inference often don't provide sensible answers.

3. Suppose a researcher routinely conducts significance tests by rejecting H_0 if the P-value satisfies $P \leq 0.05$. Suppose a test using a test statistic T and right-tail probability for the P-value has null distribution $P(T = 0) = 0.30$, $P(T = 3) = 0.62$, and $P(T = 9) = 0.08$.

   (a) Show that with the usual P-value, the actual $P(\text{Type I error}) = 0$ rather than 0.05.

   (b) Show that with the mid P-value, the actual $P(\text{Type I error}) = 0.08$.

   (c) Repeat (a) and (b) using $P(T = 0) = 0.30$, $P(T = 3) = 0.66$, and $P(T = 9) = 0.04$. Note that the test with mid P-value can be *conservative* (having actual $P(\text{Type I error})$ below the desired value) or *liberal* (having actual $P(\text{Type I error})$ above the desired value). The test with the ordinary P-value cannot be liberal.

4. For a given sample proportion p, show that a value π_0 for which the test statistic $z = (p - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n}$ takes some fixed value z_0 (such as 1.96) is a solution to the equation $(1 + z_0^2/n)\pi_0^2 + (-2p - z_0^2/n)\pi_0 + p^2 = 0$. Hence, using the formula $x = (-b \pm \sqrt{b^2 - 4ac})/2a$ for solving the quadratic equation $ax^2 + bx + c = 0$, obtain the limits for the 95% confidence interval for the probability of success when a clinical trial has 9 successes in 10 trials.

Chapter 2

5. An estimated odds ratio for adult females between the presence of squamous cell carcinoma (yes, no) and smoking behavior (smoker, non-smoker) equals 11.7 when the smoker category consists of 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. Brownson *et al.*, *Epidemiology* **3**: 61-64, (1992)). Show that the estimated odds ratio between carcinoma and smoking levels ($s \geq 20, 0 < s < 20$) equals $26.1/11.7 = 2.2$.

6. Refer to Table 2.1 about belief in an afterlife.

   (a) Construct a 90% confidence interval for the difference of proportions, and interpret.

   (b) Construct a 90% confidence interval for the odds ratio, and interpret.

7. Refer to Exercise 2.12. Given that a murderer was white, can you estimate the probability that the victim was white? What additional information would you need to do this? (Hint: How could you use Bayes Theorem?)

8. A statistical analysis that combines information from several studies is called a *meta analysis*. A meta analysis compared aspirin to placebo on incidence of heart attack and of stroke, separately for men and for women (*J. Amer. Med. Assoc.*, vol. 295, pp. 306-313, 2006). For the Women's Health Study, heart attacks were reported for 198 of 19,934 taking aspirin and for 193 of 19,942 taking placebo.

   (a) Construct the 2×2 table that cross classifies the treatment (aspirin, placebo) with whether a heart attack was reported (yes, no).

   (b) Estimate the odds ratio. Interpret.

   (c) Find a 95% confidence interval for the population odds ratio for women. Interpret. (As of 2006, results suggested that for women, aspirin was helpful for reducing risk of stroke but not necessarily risk of heart attack.)

9. A European study estimated that the lifetime probability that a woman develops lung cancer during her lifetime were 0.185 for heavy smokers (more than 5 cigarettes for day) and 0.004 for nonsmokers. Find and interpret the difference of proportions and the relative risk.

10. The study described in Exercise 2.15 was a "prospective cohort study." Explain what is meant by this.

11. A large-sample confidence interval for the log of the relative risk is

$$\log(p_1/p_2) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}} \ .$$

    Antilogs of the endpoints yield an interval for the true relative risk. Verify the 95% confidence interval of (1.43, 2.30) for the aspirin and heart attack study.

12. For the aspirin and heart attacks example, find the P-value for testing that the incidence of heart attacks is independent of aspirin intake using (a) X^2, (b) G^2. Interpret results.

13. In an investigation of the relationship between stage of breast cancer at diagnosis (local or advanced) and a woman's living arrangement (D. J. Moritz and W. A. Satariano, *J. Clin. Epidemiol.* **46**: 443–454 (1993)), 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. Reconstruct the analysis they performed to obtain this P-value.

Table 1: Data for Exercise 14.

Diagnosis	Drugs	No Drugs
Schizophrenia	105	8
Affective disorder	12	2
Neurosis	18	19
Personality disorder	47	52
Special symptoms	0	13

Source: E. Helmes and G. C. Fekken, *J. Clin. Psychol.* **42**: 569-576 (1986). Copyright by Clinical Psychology Publishing Co., Inc., Brandon, VT. Reproduced by permission of the publisher.

14. Table 1 classifies a sample of psychiatric patients by their diagnosis and by whether their treatment prescribed drugs.

    (a) Conduct a test of independence, and interpret the P-value.

    (b) Obtain standardized residuals, and interpret.

    (c) Partition chi-squared into three components to describe differences and similarities among the diagnoses, by comparing (i) the first two rows, (ii) the third and fourth rows, (iii) the last row to the first and second rows combined and the third and fourth rows combined.

15. In Exercise 2.16, show how to obtain the estimated expected cell count of 35.8 for the first cell.

16. For tests of H_0: independence, $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$.

    (a) Show that $\{\hat{\mu}_{ij}\}$ have the same row and column totals as $\{n_{ij}\}$.

    (b) For 2×2 tables, show that $\hat{\mu}_{11}\hat{\mu}_{22}/\hat{\mu}_{12}\hat{\mu}_{21} = 1.0$. Hence, $\{\hat{\mu}_{ij}\}$ satisfy H_0.

17. A chi-squared variate with degrees of freedom equal to df has representation $Z_1^2 + ... + Z_{df}^2$, where $Z_1, \ldots, Z_{df}$ are df independent standard normal variates.

    (a) If Z has a standard normal distribution, what distribution does Z^2 have?

    (b) Show that if Y_1 and Y_2 are independent chi-squared variates with degrees of freedom df_1 and df_2, then $Y_1 + Y_2$ has a chi-squared distribution with $df = df_1 + df_2$.

18. By trial and error, find a 3×3 table of counts for which the P-value is greater than 0.05 for the X^2 test but less than 0.05 for the M^2 ordinal test. Explain why this happens.

19. Of the six candidates for three managerial positions, three are female and three are male. Denote the females by F1, F2, F3 and the males by M1, M2, M3. The result of choosing the managers is (F2, M1, M3).

    (a) Identify the 20 possible samples that could have been selected, and construct the contingency table for the sample actually obtained.

(b) Let $\hat{\pi}_1$ denote the sample proportion of males selected and $\hat{\pi}_2$ the sample proportion of females. For the observed table, $\hat{\pi}_1 - \hat{\pi}_2 = 1/3$. Of the 20 possible samples, show that 10 have $\hat{\pi}_1 - \hat{\pi}_2 \geq 1/3$. Thus, if the three managers were randomly selected, the probability would equal $10/20 = 0.50$ of obtaining $\hat{\pi}_1 - \hat{\pi}_2 \geq 1/3$. This reasoning provides the P-value for Fisher's exact test with H_a: $\pi_1 > \pi_2$.

20. Refer to Exercise 2.27. If half the newborns are of each gender, for each race, find the marginal odds ratio between race and whether a murder victim.

21. For three-way contingency tables:

    (a) When any pair of variables is conditionally independent, explain why there is homogenous association.

    (b) When there is not homogeneous association, explain why no pair of variables can be conditionally independent.

22. For the happiness variable with categories (very, pretty, not), the General Social Survey gave counts (486, 855, 265) in 1972 and (786, 1403, 341) in 2016. Analyze these data.

Chapter 3

23. For the snoring and heart disease data, refer to the linear probability model. Would the least squares fit differ from the ML fit for the 2484 binary observations? (Hint: The least squares fit is the same as the ML fit of the GLM assuming normal rather than binomial random component.)

24. From equation (3.1) for logistic regression, explain why the odds ratio naturally arises as a measure for comparing two groups with that model.

25. Show that the logistic regression equation follows from formula (3.1) for $P(Y = 1)$.

26. One question in a recent General Social Survey asked subjects how many times they had sexual intercourse in the previous month.

    (a) The sample means were 5.9 for males and 4.3 for females; the sample variances were 54.8 and 34.4. Does an ordinary Poisson GLM seem appropriate? Explain.

    (b) The GLM with log link and a dummy variable for gender (1 = males, 0 = females) has gender estimate 0.308. The SE is 0.038 assuming a Poisson distribution and 0.127 for a model (assuming a negative binomial distribution) that allows overdispersion. Why are the SE values so different?

    (c) The Wald 95% confidence interval for the ratio of means is $(1.26, 1.47)$ for the Poisson model and $(1.06, 1.75)$ for the negative binomial model. Which interval do you think is more appropriate? Why?

Table 2: Table for Exercise on Oral Contraceptive Use

Variable	Coding=1 if:	Estimate	*SE*
Age	35 or younger	−1.320	0.087
Race	white	0.622	0.098
Education	$\geq$ 1 year college	0.501	0.077
Marital status	married	−0.460	0.073

Source: Debbie Wilson, College of Pharmacy, Univ. of Florida.

27. Fit the Poisson GLM with identity link to the horseshoe crab data for predicting the number of satellites, and verify the prediction equation shown in Section 3.3.3.

28. Refer to Exercise 3.11. The wafers are also classified by thickness of silicon coating ($z = 0$, low; $z = 1$, 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, making inferences about the effects of treatment type and of thickness of coating.

Chapter 4

29. A study[1] used logistic regression to predict whether the stage of breast cancer at diagnosis was advanced or was local for a sample of 444 middle-aged and elderly women. A table referring to a particular set of 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; it reported the effect of income (three categories) as 0.72 for \$10,000-24,999 versus < \$10,000 and 0.41 for \$25,000+ versus < \$10,000. Estimate the odds ratios for the third pair of categories for each factor.

30. A study used the Behavioral Risk Factors Social Survey to consider factors associated with American women's use of oral contraceptives. Table 2 summarizes effects for a logistic regression model for the probability of using oral contraceptives. Each predictor uses an indicator variable, and the table lists the category having value 1.

    (a) Interpret effects.

    (b) Construct and interpret a confidence interval for the conditional odds ratio between contraceptive use and education.

31. A sample of subjects were asked their opinion about current laws legalizing abortion (support, oppose). For the explanatory variables gender (female, male), religious affiliation (Protestant, Catholic, Jewish), and political party affiliation (Democrat, Republican, Independent), the model for the probability π of supporting legalized abortion,

$$\text{logit}(\pi) = \alpha + \beta_h^G + \beta_i^R + \beta_j^P,$$

[1] Moritz and Satariano, *J. Clin. Epidemiol.*, 46: 443-454 (1993)

has reported parameter estimates (setting the parameter for the last category of a variable equal to 0.0) $\hat{\alpha} = -0.11$, $\hat{\beta}_1^G = 0.16$, $\hat{\beta}_2^G = 0.0$, $\hat{\beta}_1^R = -0.57$, $\hat{\beta}_2^R = -0.66$, $\hat{\beta}_3^R = 0.0$, $\hat{\beta}_1^P = 0.84$, $\hat{\beta}_2^P = -1.67$, $\hat{\beta}_3^P = 0.0$.

(a) Interpret how the odds of supporting legalized abortion depend on gender.

(b) Find the estimated probability of supporting legalized abortion for (i) Male Catholic Republicans, (ii) Female Jewish Democrats.

(c) If we defined parameters such that the *first* category of a variable has value 0, then what would $\hat{\beta}_2^G$ equal? Show then how to obtain the odds ratio that describes the conditional effect of gender.

32. For the horseshoe crab data file `Crabs` at the text website, fit the logistic regression model for the probability of a satellite, using weight as the predictor.

(a) Construct a 95% confidence interval to describe the effect of weight on the odds of a satellite. Interpret.

(b) Conduct the Wald or likelihood-ratio test of the hypothesis that weight has no effect. Report the P-value, and interpret.

33. Refer to model (4.3) with width and color effects for the horseshoe crab data. Using the data file `Crabs` at the text website:

(a) Fit the model, treating color as nominal-scale but with weight instead of width as x. Interpret the parameter estimates.

(b) Controlling for weight, conduct a likelihood-ratio test of the hypothesis that having a satellite is independent of color. Interpret.

(c) Using models that treat color in a quantitative manner with scores (1, 2, 3, 4), repeat the analyses in (a) and (b).

34. Using indicators for the first three color categories, Model (4.3) for the probability π of a satellite for horseshoe crabs with color and width predictors has fit

$$\text{logit}(\hat{\pi}) = -12.715 + 1.330c_1 + 1.402c_2 + 1.106c_3 + 0.468x.$$

Consider this fit for crabs of width $x = 20$ cm. This yields $\hat{\pi}_i = 0.0954$ for medium dark crabs ($c_3 = 1$) and $\hat{\pi}_i = 0.0337$ for dark crabs, for a ratio of 2.8. Estimate the odds of a satellite for medium-dark crabs and the odds for dark crabs. Show two ways that the odds ratio equals 3.0. (When each probability is close to zero, the odds ratio is similar to the ratio of probabilities, providing another interpretation for logistic regression parameters. For widths at which $\hat{\pi}$ is small, $\hat{\pi}$ for medium-dark crabs is about 3 times that for dark crabs.)

Table 3: Data for Exercise on Teenagers and Sex

		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 by The National Council on Family Relations.

35. For recent General Social Survey data, the logistic regression model relating Y = whether attended college (1 = yes) to family income (thousands of dollars), whether mother attended college (1 = yes, 0 = no), and whether father attended college (1 = yes, 0 = no), has output shown. In a report of about 100 words, explain how to interpret the model fit, indicating limitations due to information not reported.

```
----------------------------------------
                  Estimate
(Intercept)       -1.90
income             0.02
mother             0.82
father             1.33
----------------------------------------
```

36. For the model, $\text{logit}[\pi(x)] = \alpha + \beta x$, show that e^{α} equals the odds of success when $x = 0$. Construct the odds of success when $x = 1$, $x = 2$, and $x = 3$. Use this to provide an interpretation of β. Generalize these results to the multiple logistic regression model.

37. Table 3 appeared in a national study of 15 and 16 year-old adolescents. The event of interest is ever having sexual intercourse. Create a data file and analyze these data. Summarize in a one-page report, including description and inference about the effects of both gender and race.

38. See `http://bmj.com/cgi/content/full/317/7153/235` for a meta analysis of studies about whether administering albumin to critically ill patients increases or decreases mortality. Analyze the data for the three studies with burn patients using logistic regression methods. Summarize your analyses in a short report.

39. Refer to Exercise 4.12 about MBTI and alcohol drinking. The area under the ROC curve equals 0.658 for the model with the four main effects and the six interaction terms, 0.640 for the model with only the four main effect terms, and 0.568 for the model with only T/F as a predictor. According to this criterion, which model would you choose (i) if you want to maximize sample predictive power (ii) if you think model parsimony is important?

40. For data from Florida on y = whether someone convicted of multiple murders receives the death penalty (1 = yes, 0 = no), the prediction equation is $\text{logit}[\hat{P}(Y = 1)] = -2.06 + 0.87d - 2.40v$,

where d and v are defendant's race and victims' race (1 = black, 0 = white). The following are true-false questions:

(a) The estimated probability of the death penalty is lowest when the defendant is white and victims are black.

(b) Controlling for victims' race, the estimated odds of the death penalty for white defendants equal 0.87 times the estimated odds for black defendants. If we instead let $d = 1$ for white defendants and 0 for black defendants, the estimated coefficient of d would be $1/0.87 = 1.15$ instead of 0.87.

(c) The lack of an interaction term means that the estimated odds ratio between the death penalty outcome and defendant's race is the same for each category of victims' race.

(d) The intercept term -2.06 is the estimated probability of the death penalty when the defendant and victims were white (i.e., $d = v = 0$).

(e) If there were 500 cases with white victims and defendants, then the model fitted count for the number who receive the death penalty equals $500e^{-2.06}/(1 + e^{-2.06})$.

Chapter 5

41. Exercise 4.1 used a labelling index (LI) to predict π = the probability of remission in cancer patients.

(a) When the data for the 27 subjects are 14 binomial observations (for the 14 distinct levels of LI), the deviance for this model is 15.7 with df = 12. Is it appropriate to use this to check the fit of the model? Why or why not?

(b) The model that also has a quadratic term for LI has deviance = 11.8. Conduct a test comparing the two models.

(c) The model in (b) has fit, $\text{logit}(\hat{\pi}) = -13.096 + 0.9625(LI) - 0.0160(LI)^2$, with $SE = 0.0095$ for $\hat{\beta}_2 = -0.0160$. If you know basic calculus, explain why $\hat{\pi}$ is increasing for LI between 0 and 30. Since LI varies between 8 and 38 in this sample, the estimated effect of LI is positive over most of its observed values.

42. Refer to Exercise 5.4. Use a process (such as backward elimination) or a criterion (such as AIC) to select a model, with *affirm* as the response variable. Interpret the parameter estimates for that model.

43. The Metropolitan Police in London reported[2] 30,475 people as missing one year. For those of age 13 or less, 33 of 3271 missing males and 38 of 2486 missing females were still missing a year later. For ages 14-18, the values were 63 of 7256 males and 108 of 8877 females; for ages 19 and above,

[2]From *Independent* newspaper (March 8, 1994), shown to me by Dr. P. M. E. Altham

Table 4: Data for Exercise on Penicillin in Rabbits

Penicillin Level	Delay	Response: Cured	Response: Died
1/8	None	0	6
	$1\frac{1}{2}$h	0	5
1/4	None	3	3
	$1\frac{1}{2}$h	0	6
1/2	None	6	0
	$1\frac{1}{2}$h	2	4
1	None	5	1
	$1\frac{1}{2}$h	6	0
4	None	2	0
	$1\frac{1}{2}$h	5	0

Source: Reprinted with permission from article by N. Mantel, *J. Amer. Statist. Assoc.*, (1963).

the values were 157 of 5065 males and 159 of 3520 females. Create a data file and analyze these data, including checking model fit and interpreting parameter estimates.

44. Table 4 refers to the effectiveness of immediately injected or $1\frac{1}{2}$-hour-delayed penicillin in protecting rabbits against lethal injection with β-hemolytic streptococci. Let X = delay, Y = whether cured, and Z = penicillin level. Fit the model, $\text{logit}[P(Y = 1)] = \beta x + \beta_k^Z$, deleting an intercept term so each level of Z has its own parameter. Argue that the pattern of 0 cell counts suggests that $\hat{\beta}_1^Z = -\infty$ and $\hat{\beta}_5^Z = \infty$. What does your software report?

45. The data file `Incontinence` at the text website[3] describes results from a study in which subjects received a drug and the outcome measures whether the subject became incontinent ($y = 1$, yes; $y = 0$, no). The three explanatory variables are lower urinary tract variables that represent drug-induced physiological changes.

    (a) Report the prediction equations when each predictor is used separately in logistic regressions.

    (b) Try to fit a main-effects logistic regression model containing all three predictors. What does your software report for the effects and their standard errors? (The ML estimates are actually $-\infty$ for x_1 and x_2 and ∞ for x_3.)

    (c) Use the penalized likelihood or Bayesian approach to fit the model. Summarize the information you get from this that you would not get using ordinary ML fitting.

46. In Chapter 4, various exercises asked for a data analysis and report. Select one of those analyses, and conduct a goodness-of-fit test for the model you used. Interpret.

47. Consider the probit model with explanatory variable x.

[3] *Source:* D. M. Potter, *Statist. Med.* **24**: 693–708 (2005)

(a) The fit of the probit model to the horseshoe crab data using x = width is probit$[\hat{\pi}(x)] = -7.502 + 0.302x$. At which x-value does the estimated probability of a satellite equal 0.50?

(b) Interpret the parameter estimates in (**a**).

48. Refer to Table 2.6 on mother's drinking and infant malformations.

(a) Fit the logistic regression model using scores $(0, 0.5, 1.5, 4, 7)$ for alcohol consumption. Check goodness of fit.

(b) Test independence using the likelihood-ratio test for the model in (**a**). (The trend test of Section 2.5.5 is the score test for this model.)

(c) The sample proportion of malformations is much higher in the highest alcohol category because, although it has only one malformation, its sample size is only 38. Are the results sensitive to this single observation? Re-fit the model without it, entering 0 malformations for 37 observations, and compare results of the likelihood-ratio test. (Because results are sensitive to a single observation, it is hazardous to make conclusions, even though n was extremely large.)

49. The slope of the line drawn tangent to the probit regression curve at a particular x value equals $(0.40)\beta \exp[-(\alpha + \beta x)^2/2]$. Show this is highest when $x = -\alpha/\beta$, where it equals 0.40β. At this point, $\pi(x) = 0.50$.

50. Section 4.1.3 showed that the horseshoe crab data with x = width has fit, logit$[\hat{\pi}(x)] = -12.351 + 0.497x$.

(a) Show that the curve for $\hat{\pi}(x)$ has the shape of a logistic *cdf* with mean 24.8 and standard deviation 3.6.

(b) Since about 95% of a bell-shaped distribution occurs within two standard deviations of the mean, argue that the probability of a satellite increases from near 0 to near 1 as width increases from about 17 to 32 cm.

51. True or false? For the model, logit$[P(Y = 1)] = \alpha + \beta x$, suppose $y = 1$ for all $x \leq 50$ and $y = 0$ for all $x > 50$. Then, the ML estimate $\hat{\beta} = -\infty$.

52. When $\beta > 0$, the logistic regression curve has the shape of the *cdf* of a logistic distribution with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1.814/\beta$. Compare to corresponding results for probit models, and explain what this suggests about relative sizes of effects in logistic and probit models.

Chapter 6

53. Refer to the alligator food choice example in Section 6.1.2. Estimate the length at which the outcomes *invertebrate* and *other* are equally likely.

Table 5: Job Satisfaction and Income, Controlling for Gender

		Job Satisfaction			
Gender	Income	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

54. Table 5, from a General Social Survey, refers to the relationship between Y = job satisfaction and income, stratified by gender, for black Americans. Treating job satisfaction as the response variable, analyze the data using a cumulative logit model with income scores (3, 10, 20, 35).

    (a) Describe the effects of income and gender.

    (b) Compare the estimated income effect to the estimate obtained after combining categories *Very dissatisfied* and *A little satisfied.* What property of the model does this reflect?

55. Fit an adjacent-categories logit model with main effects to the job satisfaction data in Table 5, using scores (1, 2, 3, 4) for income.

    (a) Use proportional odds structure. Interpret the estimated effect of income.

    (b) Fit the model allowing different effects for each logit, which is equivalent to a baseline-category logit model. Interpret the income effect. How does this model differ in terms of how it treats job satisfaction?

56. Analyze the job satisfaction data of Table 5 using sequential logits. Prepare a short summary of your analyses and interpretations.

57. The sample in Table 5 consists of 104 black Americans. The table relating income and job satisfaction for the combined sample of black and white subjects and females and males in the same General Social Survey had counts (by row) of (6, 18, 52, 39 / 10, 16, 104, 72 / 8, 18, 96, 98 / 11, 19, 110, 164). Test the hypothesis of independence between income and job satisfaction, (a) using a model that treats income and job satisfaction as nominal, (b) using a model that incorporates the category orderings. Interpret, and compare results, indicating the extent to which conclusions suffer when you do not use the ordinality.

58. For a recent GSS, counts in the happiness categories (not too, pretty, very) were (67, 650, 555) for those who were married and (65, 276, 93) for those who were divorced. Analyze these data

Table 6: GSS Data for exercise on Happiness and Religious Attendance

	Happiness		
Religion	Not too happy	Pretty happy	Very happy
1	189	908	382
2	53	311	180
3	46	335	294

Table 7: Data for exercise on Busing

President	Busing	Home	
		1	2
1	1	41	65
	2	72	175
2	1	2	9
	2	4	55

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

using a Bayesian or frequentist approach. Prepare a short report summarizing your analyses and interpretations.

59. Table 6 shows results from the 2000 General Social Survey relating happiness and religious attendance (1 = at most several times a year, 2 = once a month to several times a year, 3 = every week to several times a week).

    (a) Fit a multinomial model. Conduct descriptive and inferential analyses about the association.

    (b) Analyze the model goodness of fit.

Chapter 7

60. Table 7 comes from a General Social Survey. White subjects in the sample 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 (1 = Yes, 2 = No or Don't know). Table 8 shows output from fitting model (BD, BP, DP).

    (a) Analyze the model goodness of fit. Interpret.

    (b) Estimate the conditional odds ratios for each pair of variables. Interpret.

    (c) Show all steps of the likelihood-ratio test for the BP association, including explaining which loglinear model holds under the null hypothesis. Interpret.

    (d) Construct a 95% confidence interval for the BP conditional odds ratio. Interpret.

Table 8: Software Output for Fitting Model to Table 7

```
Criteria For Assessing Goodness Of Fit
Criterion                 DF           Value
Deviance                   1          0.4794
Pearson Chi-Square         1          0.5196

        Analysis Of Parameter Estimates
Parameter           DF  Estimate  Std Error
president*busing     1    0.7211    0.3539
president*home       1    1.5520    0.4436
busing*home          1    0.4672    0.2371

                    LR Statistics
        Source                 DF    Chi-Square    Pr > ChiSq
        president*busing        1       4.64        0.0313
        president*home          1      17.18        <.0001
        busing*home             1       3.83        0.0503
```

Table 9: Data for exercise on Measures to Deal with AIDS

	Information	Health Opinion	
Gender	Opinion	Support	Oppose
Male	Support	76	160
	Oppose	6	25
Female	Support	114	181
	Oppose	11	48

Source: General Social Survey

61. In a General Social Survey respondents were asked "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms." Table 9 shows responses on these two items, classified also by the respondent's gender. Denote the variables by G for Gender, H for opinion on Health care costs, and I for opinion on an Information program.

    (a) Fit the model (GH, GI, HI) and test its goodness of fit.

    (b) For this model, estimate the GI conditional odds ratio, construct a 95% confidence interval, and interpret.

    (c) Given the model, test whether G and I are conditionally independent. Do you think the GI term needs to be in the model?

62. For the MBTI data of Exercise 7.4, refer to Table 10. The estimates shown use N for the first category of the S/N scale and J for the first category of the J/P scale. Suppose you instead use S

Table 10: Partial Output for Fitting Loglinear Model to MBTI data

```
-------------------------------------------------------------------------
               Criteria For Assessing Goodness Of Fit
    Criterion                 DF           Value
    Deviance                   7         12.3687
    Pearson Chi-Square         7         12.1996

                  Analysis Of Parameter Estimates
                                Standard   LR 95% Confidence  Wald Chi-
Parameter         DF  Estimate     Error        Limits         Square
EI*SN      e  n    1    0.3219    0.1360    0.0553    0.5886     5.60
SN*TF      n  f    1    0.4237    0.1520    0.1278    0.7242     7.77
SN*JP      n  j    1   -1.2202    0.1451   -1.5075   -0.9382    70.69
TF*JP      f  j    1   -0.5585    0.1350   -0.8242   -0.2948    17.12
-------------------------------------------------------------------------
```

for the first category of the S/N scale. Then, report the estimated conditional odds ratio and the 95% likelihood-ratio confidence interval, and interpret.

63. Refer to text Exercise 7.4. PROC GENMOD in SAS reports the maximized log likelihood as 3475.19 for the model of mutual independence ($df = 11$), 3538.05 for the model of homogeneous association ($df = 5$), and 3539.58 for the model containing all the three-factor interaction terms.

    (a) Write the loglinear model for each case, and show that the numbers of parameters are 5, 11, and 15.

    (b) According to AIC, which of these models seems best? Why?

64. Refer to text Exercise 7.4. Table 10 shows the fit of the model that assumes conditional independence between E/I and T/F and between E/I and J/P but has the other pairwise associations.

    (a) Compare this to the fit of the model containing all the pairwise associations, which has deviance 10.16 with df = 5. What do you conclude?

    (b) Show how to use the limits reported to construct a 95% likelihood-ratio confidence interval for the conditional odds ratio between the S/N and J/P scales. Interpret.

65. Refer to Table 7.8 on the substance use survey which also classified students by gender (G) and race (R).

    (a) Analyze these data using logistic models, treating marijuana use as the response variable. Select a model.

    (b) Which loglinear model is equivalent to your choice of logistic model?

66. Table 11 refers to applicants to graduate school at the University of California, Berkeley for the fall 1973 session. Admissions decisions are presented by gender of applicant, for the six largest graduate departments. Denote the three variables by A = whether admitted, G = gender, and D = department. Fit loglinear model (AD, AG, DG).

Table 11: Data for Exercise on Admissions to Berkeley

	Whether Admitted			
	Male		Female	
Department	Yes	No	Yes	No
1	512	313	89	19
2	353	207	17	8
3	120	205	202	391
4	138	279	131	244
5	53	138	94	299
6	22	351	24	317
Total	1198	1493	557	1278

Note: For further details, see Bickel *et al.*, *Science, 187*, 398-403 (1975).

(a) Report the estimated AG conditional odds ratio, and compare it to the AG marginal odds ratio. Why are they so different?

(b) Report deviance and *df* values, and comment on the quality of fit. Conduct a residual analysis. Describe the lack of fit.

(c) Deleting the data for Department 1, re-fit the model. Interpret.

(d) Deleting the data for Department 1 and treating A as the response variable, fit an equivalent logistic model for model (AD, AG, DG) in (**c**). Show how to use each model to obtain an odds ratio estimate of the effect of G on A, controlling for D.

67. For the Maine accident data modeled in Section 7.2.7:

(a) Verify the logistic model that follows from loglinear model (GLS, GI, LI, IS).

(b) 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.

68. For a four-way table, are X and Y independent, given Z alone, for model (a) (WX, XZ, YZ, WZ), (b) (WX, XZ, YZ, WY)?

69. Consider logit models for a four-way table in which X_1, X_2, and X_3 are predictors of Y. When the table is collapsed over X_3, indicate whether the association between X_1 and Y remains unchanged, for the model (a) that has main effects of all predictors, (b) that has main effects of X_1 and X_2 but assumes no effect for X_3.

70. Generalizations of the linear-by-linear association model analyze association between ordinal variables X and Y while controlling for a categorical variable that may be nominal or ordinal. The model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ},$$

with ordered scores $\{u_i\}$ and $\{v_j\}$, is a special case of model (XY, XZ, YZ) that replaces λ_{ij}^{XY} by a linear-by-linear term.

(a) Show that the XY conditional independence model (XZ, YZ) is the special case of this model. Thus, assuming the ordinal model, explain how one could construct a test with $df = 1$ for testing XY conditional independence.

(b) For this model, the ordinary linear-by-linear equation applies for the cells at each fixed level of Z. With unit-spaced scores, explain why the model implies that every local odds ratio in each partial table equals $\exp(\beta)$.

(c) If we replace β in this model by β_k, is there homogeneous association? Why or why not? (The fit of this model is equivalent to fitting the $L \times L$ association model separately for each partial table.)

71. For the linear-by-linear association model applied with column scores $\{v_j = j\}$, show the form for the adjacent-category logits within row i, identifying α_j with $(\lambda_{j+1}^Y - \lambda_j^Y)$ and the row scores $\{u_i\}$ with the levels of x. In fact, the two models are equivalent. The logit representation provides an interpretion for the loglinear model.

72. A recent General Social Survey asked subjects, "Within the past 12 months, how many people have you known personally that were victims of homicide?" The sample mean for the 159 blacks was 0.522, with a variance of 1.150. The sample mean for the 1149 whites was 0.092, with a variance of 0.155.

(a) Let Y_{ij} denote the response for subject j of race i, and let $\mu_{ij} = E(Y_{ij})$. The Poisson model $\log(\mu_{ij}) = \alpha + \beta x_{ij}$ with $x_{1j} = 1$ (blacks) and $x_{2j} = 0$ (whites) has fit $\log(\hat{\mu}_{ij}) = -2.38 + 1.733 x_{ij}$. Show that the estimated population means are 0.522 for blacks and 0.092 for whites, which are the sample means.

(b) For the Poisson GLM, the standard error of $\hat{\beta}$ is 0.147. Show that the Wald 95% confidence interval for the ratio of means for blacks and whites is $(4.2, 7.5)$. (Hint: Note that β is the log of the ratio of the means.)

(c) The negative binomial loglinear model has the same estimates as in (a), but the standard error of $\hat{\beta}$ increases to 0.238 and the Wald 95% confidence interval for the ratio of means is $(3.5, 9.0)$. Based on the sample means and variances, which confidence interval is more believeable? Why?

(d) The negative binomial model has $\hat{D} = 4.94$ ($SE = 1.00$). Based on this, which model do you think is more appropriate – the Poisson, or the negative binomial GLM? Why?

73. Table 12 lists total attendance (in thousands) and the total number of arrests, in a season for soccer teams in the Second Division of the British football league. (Thanks to Dr. P. M. E. Altham for showing me these data.)

(a) Let Y denote the number of arrests for a team with total attendance t. Explain why the model $E(Y) = \mu t$ might be plausible. Show that it has alternative form $\log[E(Y)/t] = \alpha$, where $\alpha = \log(\mu)$, and express this model with an offset term.

Table 12: Data for exercise on Soccer Game Arrests

Team	Attendance	Arrests	Team	Attendance	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 Dr. P.M.E. Altham for showing me these data.

(b) Assuming Poisson sampling, fit the model. Report and interpret $\hat{\mu}$.

(c) Plot arrests against attendance, and overlay the prediction equation. Use residuals to identify teams that had a much larger or smaller than expected number of arrests.

(d) Now fit the model $\log[E(Y)/t] = \alpha$ by assuming a negative binomial distribution. Compare $\hat{\alpha}$ and its SE to what you got in (a). Based on this information and the estimate of the dispersion parameter and its SE, does the Poisson assumption seem appropriate?

74. Table 13 shows data on accidents involving trains.

(a) Is it plausible that the collision counts are independent Poisson variates with constant rate over the 29-years? Respond by comparing a Poisson GLM for collision rates that contains only an intercept term to a Poisson GLM that contains also a time trend. The deviances of the two models are 35.1 and 23.5.

(b) For a negative binomial model, the estimated collision rate x years after 1975 was $e^{-4.20}(e^{-0.0337})^x = (0.015)(0.967)^x$. The ML estimate $\hat{\beta} = -0.0337$ has $SE = 0.0130$. Conduct the Wald test of H_0: $\beta = 0$ against H_a: $\beta \neq 0$.

(c) The likelihood-ratio 95% confidence interval for β is $(-0.060, -0.008)$. Find the interval for the multiplicative annual effect on the accident rate, and interpret.

75. Table 14, based on a study with British doctors conducted by R. Doll and A. Bradford Hill.[4]

[4]From article by N. R. Breslow in *A Celebration of Statistics*, Springer-Verlag, (1985)

Table 13: Collisions Involving Trains in Great Britain

Year	Train-km	Train Collisions	Train-road Collisions	Year	Train-km	Train Collisions	Train-road C ollisions
2003	518	0	3	1988	443	2	4
2002	516	1	3	1987	397	1	6
2001	508	0	4	1986	414	2	13
2000	503	1	3	1985	418	0	5
1999	505	1	2	1984	389	5	3
1998	487	0	4	1983	401	2	7
1997	463	1	1	1982	372	2	3
1996	437	2	2	1981	417	2	2
1995	423	1	2	1980	430	2	2
1994	415	2	4	1979	426	3	3
1993	425	0	4	1978	430	2	4
1992	430	1	4	1977	425	1	8
1991	439	2	6	1976	426	2	12
1990	431	1	2	1975	436	5	2
1989	436	4	4				

Source: British Department of Transport

Table 14: Data for Exercise on the Doll/Hill Study

	Person-years		Coronary Deaths	
Age	Nonsmokers	Smokers	Nonsmokers	Smokers
35-44	18,793	52,407	2	32
45-54	10,673	43,248	12	104
55-64	5710	28,612	28	206
65-74	2585	12,663	28	186
75-84	1462	5317	31	102

Source: R. Doll and A. B. Hill, *Natl. Cancer Inst. Monogr., 19*: 205-268 (1966).

(a) For each age, compute the sample coronary death rates per 1000 person-years, for nonsmokers and smokers. To compare them, take their ratio and describe its dependence on age.

(b) Specify a main-effects Poisson model for the log rates having four parameters for age and one for smoking. Explain why this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over levels of age. Based on (a), would you expect this model to be appropriate?

(c) Based on (a), explain why it is sensible to add a quantitative interaction of age and smoking. Specify this model, and show that the log of the ratio of coronary death rates changes linearly with age.

(d) Fit the model in (b). Assign scores to the levels of age for a product interaction term between age and smoking, and fit the model in (c). Compare the fits by comparing the deviances. Interpret.

76. For rate data, a GLM with identity link is

$$\mu/t = \alpha + \beta x.$$

Explain why you could fit this model using t and tx as explanatory variables and with no intercept or offset terms.

77. A study dealing with motor vehicle accident rates for elderly drivers indicated that the entire cohort of elderly drivers had 495 injurious accidents in 38.7 thousand years of driving. Using a Poisson GLM, find a 95% confidence interval for the true rate. (Hint: Find a confidence interval first for the log rate by obtaining the estimate and standard error for the intercept term in a loglinear model that has no other predictor and uses log(38.7) as an offset.)

Chapter 8

Table 15: Data for Exercise on crossover experiment

Treatment Order	Treatment That Is Better	
	First	Second
A then B	25	10
B then A	12	20

78. For Table 9 on opinions about measures to deal with AIDS, treat the data as matched pairs on opinion, stratified by gender.

    (a) For females, test equality of the true proportions supporting government action for the two items.

    (b) Refer to (a). Construct a 90% confidence interval for the difference between the true proportions of support. Interpret.

    (c) For females, estimate the odds ratio $\exp(\beta)$ for (i) a marginal model, (ii) a subject-specific model. Interpret.

    (d) Explain how you could construct a 90% confidence interval for the difference between males and females in their differences of proportions of support for a particular item. (Hint: The gender samples are independent.)

79. A crossover experiment with 100 subjects compares two treatments for migraine headaches. The response scale is success (+) or failure (−). Half the study subjects, randomly selected, used drug A the first time they get a migraine headache and drug B the next time. For them, 6 had responses $(A+, B+)$, 25 had responses $(A+, B-)$, 10 had responses $(A-, B+)$, and 9 had responses $(A-, B-)$. The other 50 subjects took the drugs in the reverse order. For them, 10 were $(A+, B+)$, 20 were $(A+, B-)$, 12 were $(A-, B+)$, and 8 were $(A-, B-)$.

    (a) Ignoring treatment order, use the McNemar test to compare the success probabilities for the two treatments. Interpret.

    (b) The McNemar test uses only the pairs of responses that differ. For this study, Table 15 shows such data from both treatment orders. Explain why a test of independence for this table tests the hypothesis that success rates are identical for the two treatments. Analyze these data, and interpret.

80. Refer to Table 8.1 on ways to help the environment. Suppose sample proportions of approval of 0.314 and 0.292 were based on *independent* samples of size 1144 each. Construct a 95% confidence interval for the true difference of proportions. Compare to the result in Section 8.1.2, and comment on how the use of dependent samples can improve precision.

81. A case-control study has 8 pairs of subjects. The cases have colon cancer, and the controls are matched with the cases on gender and age. A possible explanatory variable is the extent of red meat in a subject's diet, measured as *low* or *high.* For 3 pairs, both the case and the control were

high; for 1 pair, both the case and the control were low; for 3 pairs, the case was high and the control was low; for 1 pair, the case was low and the control was high.

(a) Display the data in a 2×2 cross-classification of diet for the case against diet for the control. Display the $2 \times 2 \times 8$ table with partial tables relating diet to response (case, control) for the matched pairs. Successive parts refer to these as Table A and Table B.

(b) Find the McNemar z^2 statistic for Table A.

(c) This sample size is too small for a large-sample test. Find the exact P-value for testing marginal homogeneity against the alternative hypothesis of a higher incidence of colon cancer for the *high* red meat diet.

82. For the subject-specific logistic model for matched pairs,

$$\text{logit}[P(Y_{i1} = 1)] = \alpha_i + \beta, \quad \text{logit}[P(Y_{i2} = 1)] = \alpha_i,$$

the estimated variance for the conditional ML estimate $\hat{\beta} = \log(n_{12}/n_{21})$ of β is $(1/n_{12} + 1/n_{21})$. Find a 95% confidence interval for the odds ratio $\exp(\beta)$ for Table 8.1 on helping the environment. Interpret.

83. For text Table 7.1 on the student survey, viewing the table as matched triplets, you can compare the proportion of *yes* responses among alcohol, cigarettes, and marijuana. Construct the marginal distribution for each substance, and find the three sample proportions of *yes* responses.

84. Refer to text Table 6.10 on a study about whether cereal containing psyllium had a desirable effect in lowering LDL cholesterol. For both the control and treatment groups, use methods of this chapter to compare the beginning and ending cholesterol levels. Compare the changes in cholesterol levels for the two groups. Interpret.

85. Refer to Table 8.5 on coffee choice. Fit the quasi-independence model. Calculate the fitted odds ratio for the four cells in the first two rows and the last two columns. Interpret. Analyze the data from the perspective of describing agreement between choice of coffee at the two times.

86. Table 16 refers to journal citations among four statistical theory and methods journals (*Biometrika, Communications in Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society Series B*) during 1987-1989. The more often that articles in a particular journal are cited, the more prestige that journal accrues. For citations involving a pair of journals X and Y, view it as a victory for X if it is cited by Y and a defeat for X if it cites Y.

(a) Fit the Bradley–Terry model. Interpret the fit, and give a prestige ranking of the journals.

(b) For citations involving *Comm. Stat.* and *JRSS-B*, estimate the probability that the *Comm. Stat.* article cites the *JRSS-B* article.

Table 16: Data for Exercise on journal citations

	Cited Journal			
Citing Journal	Biometrika	Comm. Stat.	JASA	JRSS-B
Biometrika	714	33	320	284
Comm. Stat.	730	425	813	276
JASA	498	68	1072	325
JRSS-B	221	17	142	188

Source: Stigler (1994). Reprinted with permission from the Institute of Mathematical Statistics.

87. In loglinear model form, the quasi-symmetry (QS) model is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij},$$

where $\lambda_{ij} = \lambda_{ji}$ for all i and j.

(a) For this model, by finding $\log(\mu_{ij}/\mu_{ji})$ show that the model implies logit model

$$\log(\pi_{ij}/\pi_{ji}) = \beta_i - \beta_j \text{ for all } i \text{ and } j.$$

(b) Show that the special case of QS with $\lambda_i^X = \lambda_i^Y$ for all i is the symmetry model in loglinear form.

(c) Show that the quasi-independence model is the special case in which $\lambda_{ij} = 0$ for $i \neq j$.

88. For matched pairs, to obtain conditional ML $\{\hat{\beta}_j\}$ for the multiple logistic regression model using software for ordinary logistic regression, let

$$y_i^* = 1 \text{ when } (y_{i1} = 1, y_{i2} = 0),\ y_i^* = 0 \text{ when } (y_{i1} = 0, y_{i2} = 1).$$

Let $x_{1i}^* = x_{1i1}^* - x_{1i2}^*, \ldots, x_{ki}^* = x_{ki1}^* - x_{ki2}^*$. Fit the ordinary logistic model to y^* with predictors $\{x_1^*, \ldots, x_k^*\}$, forcing the intercept parameter α to equal zero. This works because the likelihood is the same as the conditional likelihood for the ordinary model after eliminating $\{\alpha_i\}$.

(a) Apply this approach with Table 8.1 and report $\hat{\beta}$ and its SE.

(b) The pairs $(y_{i1} = 1, y_{i2} = 1)$ and $(y_{i1} = 0, y_{i2} = 0)$ do not contribute to the likelihood or to estimating $\{\beta_j\}$. Identify the counts for such pairs in Table 8.1. Do these counts contribute to McNemar's test?

89. Table 17, from the 2004 General Social Survey, reports subjects' current religious affiliation and at age 16, for categories (1) Protestant, (2) Catholic, (3) Jewish, (4) None or Other.

(a) The symmetry model has deviance 150.6 with $df = 6$. Use residuals for the model to analyze transition patterns between pairs of religions.

Table 17: Data for Exercise on Religious Affiliation

Affiliation at Age 16	Religious Affiliation Now			
	1	2	3	4
1	1228	39	2	158
2	100	649	1	107
3	1	0	54	9
4	73	12	4	137

Table 18: Data from General Social Survey on Recycling and Pesticides

Chemical Free	Recycle		
	1	2	3
1	66	39	3
2	227	359	48
3	150	216	108

(b) The quasi-symmetry model has deviance 2.3 with $df = 3$. Test marginal homogeneity by comparing its fit to symmetry. (The small P-value mainly reflects the large sample size and is due to a small decrease in the proportion classified Catholic and increase in the proportion classified None or Other, with little evidence of change for other categories.)

90. Table 18 is from a General Social Survey. Subjects were asked "How often do you make a special effort to buy fruits and vegetables grown without pesticides or chemicals?" and "How often do you make a special effort to sort glass or cans or plastic or papers and so on for recycling?" The categories are 1 = always, 2 = often or sometimes, 3 = never. Analyze these data using models. Prepare a short report summarizing your analyses, with edited sofware output as an appendix.

Chapter 9

91. Table 9.1 shows General Social Survey responses on attitudes toward legalized abortion. For the response Y_t about legalization (1 = support, 0 = oppose) for question t ($t = 1, 2, 3$) and for gender g (1 = female, 0 = male), consider the model $\text{logit}[P(Y_t = 1)] = \alpha + \gamma g + \beta_t$ with $\beta_3 = 0$.

(a) A GEE analysis using unstructured working correlation gives correlation estimates 0.826 for questions 1 and 2, 0.797 for 1 and 3, and 0.832 for 2 and 3. What does this suggest about a reasonable working correlation structure?

(b) Table 19 shows a GEE analysis with exchangeable working correlation. Interpret effects.

Chapter 10

Table 19: Software Output for Abortion Survey Data

```
-------------------------------------------------------------
                    Working Correlation Matrix
                         Col1         Col2         Col3
            Row1       1.0000       0.8173       0.8173
            Row2       0.8173       1.0000       0.8173
            Row3       0.8173       0.8173       1.0000

        Analysis Of GEE Parameter Estimates
        Empirical Standard Error Estimates
  Parameter   Estimate  Std Error     Z   Pr > |Z|
  Intercept    -0.1253   0.0676    -1.85   0.0637
  question  1   0.1493   0.0297     5.02   <.0001
  question  2   0.0520   0.0270     1.92   0.0544
  question  3   0.0000   0.0000      .      .
  female        0.0034   0.0878     0.04   0.9688
-------------------------------------------------------------
```

92. In Exercise 10.1, compare $\hat{\beta}$ and its SE for this approach to their values for the conditional ML approach.

93. Refer to text Table 7.16 on government spending. Analyze these data using a cumulative logit model with random effects. Interpret. Compare results to those with a marginal model.

94. Refer to text Table 4.8 about an eight-center clinical trial comparing a drug to placebo for curing an infection. Model the data in a way that allows the odds ratio to vary by center. Summarize your analyses and interpretations.

95. See `http://bmj.com/cgi/content/full/317/7153/235` for a meta analysis of studies about whether administering albumin to critically ill patients increases or decreases mortality. Analyze the data for the 13 studies with hypovolaemia patients using logistic models with (i) fixed effects, (ii) random effects. Summarize your analyses in a two-page report.

96. Refer to the insomnia data in text Table 9.2.

    (a) From results for the GLMM, explain how to get the interpretation that response distributions are similar initially for the two treatment groups, but the interaction suggests that at the follow-up response the active treatment group tends to fall asleep more quickly.

    (b) According to SAS, the maximized log likelihood equals -593.0, compared to -621.0 for the simpler model forcing $\sigma = 0$. Compare models, using a likelihood-ratio test. What do you conclude?

97. For General Social Survey data on responses of 1308 subjects to the question, "Within the past 12 months, how many people have you known personally that were victims of homicide?" one can

use Poisson and negative binomial GLMs for count data. Here is a possible GLMM: For response y_i for subject i of race x_i (1 = black, 0 = white),

$$\log[E(Y_i)] = u_i + \alpha + \beta x_i,$$

where conditional on u_i, y_i has a Poisson distribution, and where $\{u_i\}$ are independent $N(0, \sigma^2)$. Like the negative binomial GLM, unconditionally (when $\sigma > 0$) this model can allow more variability than the Poisson GLM. The Poisson GLMM has $\hat{\alpha} = -3.69$ and $\hat{\beta} = 1.90$, with $\hat{\sigma} = 1.6$. Interpret $\hat{\beta}$.

98. True, or false? In a logistic regression model containing a random effect as a way of modeling within-subject correlation in repeated measures studies, the greater the estimate $\hat{\sigma}$ for the random effects distribution, the greater the heterogeneity of the subjects, and the larger in absolute value the estimated effects tend to be compared to the marginal model approach (with effects averaged over subjects, rather than conditional on subjects). Explain your choice of answer.

99. Create a data file for five situations by downloading data for the items labeled (ABRAPE, ABHLTH, ABSINGLE, ABDEFECT, ABPOOR) in the most recent GSS at `http://sda.berkeley.edu/archive.htm`. Fit a latent class model. For each latent class, find the estimated probability of supporting legalized abortion the five situations. Suggest a tentative interpretation for the classes.

100. You plan to apply the subject-specific matched-pairs model to a data set for which y_{i1} is whether the subject agrees that abortion should be legal if the woman cannot afford the child (1 = yes, 0 = no), and y_{i2} is whether the subject opposes abortion if a woman wants it because she is unmarried (1 = yes, 0 = no). Indicate a way in which this model would probably be inappropriate. (Hint: Do you think these variables would have a positive, or negative, log odds ratio?) How could you reword the second question so the model would be more appropriate?

101. A dataset on pregnancy rates among girls in 13 north central Florida counties has information on the total in 2017 for each county i on T_i = number of births and y_i = number of those for which mother's age was under 18. Let π_i be the probability that a pregnancy in county i is to a mother of age under 18. The logistic-normal model, $\text{logit}(\pi_i) = u_i + \alpha$, has $\hat{\alpha} = -3.24$ and $\hat{\sigma} = 0.33$.

    (a) Find $\hat{\pi}_i$ for a county estimated to be (i) at the mean, (ii) two standard deviations below the mean, (iii) two standard deviations above the mean on the random effects distribution.

    (b) For estimating $\{\pi_i\}$, what advantage does this model have over the fixed effects model, $\text{logit}(\pi_i) = \beta_i$?

102. You plan to apply the subject-specific matched-pairs model to a data set for which y_{i1} is whether the subject agrees that abortion should be legal if the woman cannot afford the child (1 = yes, 0 = no), and y_{i2} is whether the subject opposes abortion if a woman wants it because she is unmarried (1 = yes, 0 = no). Indicate a way in which this model would probably be inappropriate. How could you reword the second question so the model would be more appropriate?

103. Raudenbush and Bryk (2002, pp. 296-304) analyzed data from a survey of 7516 sixth graders in 356 schools in Thailand. The response variable y_{it} measured whether student t in school i had to repeat at least one grade during the primary school years (1 = yes, 0 = no). The student-level model was

$$\text{logit}[P(y_{it} = 1)] = \alpha_i + \beta_1 SES_{it} + \beta_2 M_{it} + \beta_3 D_{it} + \beta_4 B_{it} + \beta_5 P_{it}$$

for SES = socioeconomic status, whether the student was male (M = 1 if yes, 0 if female), spoke Central Thai dialect (D = 1 if yes, 0 if no), had breakfast daily (B = 1 if yes, 0 if no), and had some preprimary experience (P = 1, 0 if no). The school-level model was

$$\alpha_i = u_i + \alpha + \gamma_1 MSES_i + \gamma_2 S_i + \gamma_3 T_i$$

for MSES = the school mean SES, S = size of school enrollment, and T = a measure of availability of textbooks in the school. Specify the full multilevel model and explain how to interpret parameters.