```
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -4.05251     1.17143   -3.459  0.000541
width          0.19207     0.04406    4.360   1.3e-05
---
(Dispersion parameter for Negative Binomial(0.9046) family taken to be 1)
              # D in our notation is 1/0.9046 = 1.11
Residual deviance: 195.81  on 171  degrees of freedom
------------------------------------------------------------------------
```

The corresponding negative binomial GLM has

$$\log(\hat{\mu}) = -4.05 + 0.192x$$

with $SE = 0.044$ for $\hat{\beta}$. Moreover, $\hat{D} = 1.1$, so at an estimated mean $\hat{\mu}$, the estimated variance is $\hat{\mu} + 1.1\hat{\mu}^2$, compared to $\hat{\mu}$ for the Poisson GLM. Fitted values are similar, but the greater estimated variance in the negative binomial model and the resulting greater SE for $\hat{\beta}$ reflect the overdispersion uncaptured with the Poisson GLM. Inspection of Figure 3.3 shows that some zero counts occur even when the sample mean response is relatively large, reflecting this overdispersion.

For the Poisson model, the 95% Wald confidence interval for the effect of width (β) is $0.164 \pm 1.96(0.020)$, which is (0.125, 0.203). For the negative binomial model, it is $0.192 \pm 1.96(0.044)$, which is (0.105, 0.278). The profile likelihood confidence intervals are similar. Confidence intervals for β with the Poisson GLM are unrealistically narrow, because of not allowing for the overdispersion.

EXERCISES

7.1 For the recent General Social Survey data on X = gender (males, females) and Y = belief in an afterlife (no, yes), Table 7.12 shows results of fitting the independence loglinear model.

a. The deviance is 0.82 with $df = 1$. What does this suggest?

b. Report $\{\hat{\lambda}_j^Y\}$. Interpret $\hat{\lambda}_1^Y - \hat{\lambda}_2^Y$.

c. For the saturated model, software reports for $\{\hat{\lambda}_{ij}^{XY}\}$:

```
-----------------------------------------------------------------
                                 Estimate  Std Error
genderfemales:beliefyes            0.1368     0.1507
-----------------------------------------------------------------
```

Estimate the odds ratio.

Table 7.12 Software output for Exercise 7.1 on belief in afterlife.

```
-----------------------------------------------------------------------------
                Estimate   Std. Error
Intercept         4.5849       0.0752
genderfemales     0.2192       0.0599
beliefyes         1.4165       0.0752
-----------------------------------------------------------------------------
```

d. The text website has a data file Postlife that cross-classifies belief in life after death with race (black, white, other). Fit the independence model and interpret $\hat{\lambda}_1^Y - \hat{\lambda}_2^Y$.

7.2 For Table 2.9, let D = defendant's race, V = victims' race, and P = death penalty verdict. Table 7.13 shows the output for fitting loglinear model (DV, DP, PV).

a. Report the estimated conditional odds ratio between D and P at each category of V. Interpret.

b. Test the goodness of fit of this model. Interpret.

c. Using the DeathPenalty data file at the text website, obtain the results shown in this output. Specify the corresponding logistic model with P as the response. Set up a grouped-data file to fit that model and show how estimated effects of D and V relate to loglinear model estimates. Which model seems more relevant for these data?

Table 7.13 Software output for Exercise 7.2 on death penalty verdicts.

```
-------------------------------------------------------------------------------
 Coefficients: # not showing intercept and main effect terms
               Estimate  Std. Error  z value  Pr(>|z|)
Dwhite:Vwhite   4.59497     0.31353   14.656   < 2e-16
Dwhite:Pyes    -0.86780     0.36707   -2.364    0.0181
Vwhite:Pyes     2.40444     0.60061    4.003  6.25e-05
---
Residual deviance:    0.37984  on 1  degrees of freedom
-------------------------------------------------------------------------------
```

7.3 Table 7.14 is based on automobile accident records supplied by the state of Florida Department of Highway Safety and Motor Vehicles. Subjects were classified by whether they were wearing a seat belt, whether ejected, and whether killed.

a. Find a loglinear model that describes the data well. Interpret the associations.

b. Since the sample size is large, goodness-of-fit statistics are large unless the model fits very well. Calculate the dissimilarity index for the model you found in (**a**) and interpret.

c. Conduct a Bayesian analysis of the homogeneous association model. Find and interpret posterior intervals for the conditional odds ratios. How do interpretations differ from ones you make with a frequentist analysis?

Table 7.14 Data for Exercise 7.3.

Safety Equipment	Whether	Injury	
in Use	Ejected	Nonfatal	Fatal
Seat belt	Yes	1,105	14
	No	411,111	483
None	Yes	4,624	497
	No	157,342	1008

Source: Florida Department of Highway Safety and Motor Vehicles.

7.4 At the website `www.stat.ufl.edu/~aa/intro-cda/data` for the second edition of this book, the `MBTI` data file cross-classifies the MBTI Step II National Sample on four binary scales of the Myers–Briggs personality test: Extroversion/Introversion (E/I), Sensing/iNtuitive (S/N), Thinking/Feeling (T/F), and Judging/Perceiving (J/P). Fit the loglinear model of homogeneous association and conduct a goodness-of-fit test. Based on the fit, show that (i) the estimated conditional association is strongest between the S/N and J/P scales, (ii) there is not strong evidence of conditional association between the E/I and T/F scale or between the E/I and J/P scales.

7.5 Refer to the auto accident injury data shown in Table 7.5.

a. Explain why the fitted odds ratios in Table 7.7 for model (GL, GS, LS, GI, LI, SI) suggest that the most likely case for injury is accidents for females not wearing seat belts in rural locations.

b. Consider the following two-stage model. The first stage is a logistic model with S as the response, for the three-way $G \times L \times S$ table. The second stage is a logistic model with these three variables as predictors for I in the four-way table. Explain why this composite model is sensible, fit the models, and interpret results.

7.6 Table 7.15, from a General Social Survey, relates responses on $R =$ religious service attendance ($1 =$ at most a few times a year, $2 =$ at least several times a year), $P =$ political views ($1 =$ liberal, $2 =$ moderate, $3 =$ conservative), $B =$ birth control availability to teenagers between the ages of 14 and 16 ($1 =$ agree, $2 =$ disagree), $S =$ sex relations before marriage ($1 =$ wrong only sometimes or not wrong at all, $2 =$ always or almost always wrong).

a. Investigate the complexity needed for loglinear modeling by fitting models having only single-factor terms, all two-factor terms, and all three-factor terms. Select a model and interpret it by estimating conditional odds ratios.

b. Draw the independence graph for model (BP, BR, BS, PS, RS). Remark on conditional independence patterns. Are any fitted marginal and conditional associations identical?

c. Fit the loglinear model that corresponds to the logistic model that predicts S using the other variables as main effects, without any interaction. Does it fit adequately?

Table 7.15 Data (file `BPRS` at text website) for Exercise 7.6.

		Premarital Sex							
		1				2			
	Religious Attendence	1		2		1		2	
	Birth control	1	2	1	2	1	2	1	2
Political Views	1	99	15	73	25	8	4	24	22
	2	73	20	87	37	20	13	50	60
	3	51	19	51	36	6	12	33	88

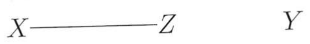
7.7 The data in Table 7.16, from a General Social Survey, are in the `Spending` data file at the text website. Subjects were asked about government spending on the environment (E), health (H), assistance to big cities (C), and law enforcement (L).

Table 7.16 Opinions about government spending.

Cities Law Enforcement		1			2			3		
Environment	Health	1	2	3	1	2	3	1	2	3
1	1	62	17	5	90	42	3	74	31	11
	2	11	7	0	22	18	1	19	14	3
	3	2	3	1	2	0	1	1	3	1
2	1	11	3	0	21	13	2	20	8	3
	2	1	4	0	6	9	0	6	5	2
	3	1	0	1	2	1	1	4	3	1
3	1	3	0	0	2	1	0	9	2	1
	2	1	0	0	2	1	0	4	2	0
	3	1	0	0	0	0	0	1	2	3

The common response scale was (1 = too little, 2 = about right, 3 = too much). Compare the models with all two-factor and with all three-factor terms. For the homogeneous association model. estimate the conditional odds ratios using the *too much* and *too little* categories for each pair of variables. Summarize the associations. Based on these results, which term(s) might you consider dropping from the model? Why?

7.8 For a three-way contingency table, consider the independence graph,



Write the corresponding loglinear model. Which pairs of variables are conditionally independent? Which pairs of variables have the same marginal association as their conditional association?

7.9 For a multiway contingency table, when is a logistic model more appropriate than a loglinear model? When is a loglinear model more appropriate?

7.10 For loglinear model (WXZ, WYZ), draw its independence graph and identify variables that are conditionally independent.

7.11 Refer to Exercise 7.7 with Table 7.16 and the `Spending` data file.

a. Beginning with the homogeneous association model, show that backward elimination yields (CE, CL, EH, HL). Interpret its fit.

b. Based on the independence graph for (CE, CL, EH, HL), show that (i) every path between C and H involves a variable in $\{E, L\}$; (ii) collapsing over H, one obtains the same associations between C and E and between C and L, and, collapsing over C, one obtains the same associations between H and E and between H and L; (iii) the conditional independence patterns between C and H and between E and L are not collapsible.

7.12 For the substance use data in Table 7.8, consider loglinear model $(AC, AM, CM, AG, AR, GM, GR)$.

a. Explain why the *AM* conditional odds ratio is unchanged by collapsing over race, but it is not unchanged by collapsing over gender.

b. Suppose we remove the *GM* term from the model. Construct the independence graph and show that $\{G, R\}$ are separated from $\{C, M\}$ by A. Explain why all conditional associations among A, C, and M are then identical to those in model (AC, AM, CM), collapsing over G and R.

7.13 Table 7.17 comes from a General Social Survey. Subjects were asked whether methods of birth control should be available to teenagers and how often they attend religious services.

a. Fit the independence model. Describe the lack of fit.

b. Using equally spaced scores, fit the linear-by-linear association model. Describe the association. Test goodness of fit. Test independence by using the ordinality, and interpret.

c. Fit the $L \times L$ model using column scores $\{1, 2, 4, 5\}$. Explain why a fitted local log odds ratio using columns 2 and 3 is double a fitted local log odds ratio using columns 1 and 2 or columns 3 and 4. What is the relation between the odds ratios?

Table 7.17 Data for Exercise 7.13 on religion and birth control.

	Teenage Birth Control			
Religious Attendance	Strongly Agree	Agree	Disagree	Strongly Disagree
Never	49	49	19	9
Less than once a year	31	27	11	11
Once or twice a year	46	55	25	8
Several times a year	34	37	19	7
About once a month	21	22	14	16
2–3 times a month	26	36	16	16
Nearly every week	8	16	15	11
Every week	32	65	57	61
Several times a week	4	17	16	20

7.14 True or false?

a. With a single categorical response variable, logistic regression models are more appropriate than loglinear models.

b. To model the association and interaction structure among several categorical response variables, logistic regression models are more appropriate than loglinear models.

c. The logistic model is a GLM assuming a binomial random component whereas the loglinear model is a GLM assuming a Poisson random component. Hence, when both are fitted to a contingency table having 50 cells with a binary response, the logistic model treats the cell counts as 25 binomial observations whereas the loglinear model treats the cell counts as 50 Poisson observations.

7.15 Consider Table 7.11 on survival of lung cancer patients.

a. Fit the more complex model that allows interaction between stage and time. Analyze whether it provides a significantly improved fit.

b. Fit the simpler model with main effects of stage and time and no histology effects. Check the fit of the model and interpret the estimated effects of stage of disease.

7.16 For the `Crabs` data file at the text website, let y = number of satellites and x = weight. Fit the Poisson and negative binomial (NB) loglinear models. For each model, report the prediction equation and SE of the weight effect, and construct a 95% confidence interval for β. Explain why the interval is wider with the NB model. Which model is more appropriate? Why?

7.17 A recent General Social Survey asked subjects how many times they had sexual intercourse in the previous month. The sample means were 5.9 for males and 4.3 for females; the sample variances were 54.8 and 34.4.

a. Does an ordinary Poisson GLM seem appropriate for comparing the means? Explain.

b. The GLM with log link and an indicator variable for gender (1 = males, 0 = females) has gender estimate 0.308. The *SE* is 0.038 assuming a Poisson distribution and 0.127 assuming a negative binomial model. 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?