AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS, 2nd ed.

SOLUTIONS TO SELECTED PROBLEMS for STA 4504/5503

These solutions are solely for the use of students in STA 4504/5503 and are not to be distributed elsewhere. Please report any errors in the solutions to Alan Agresti, e-mail aa@stat.ufl.edu.

Chapter 1

1. Response variables are (a) Attitude toward gun control, (b) Heart disease, (c) Vote for President, (d) Quality of life.

2.a. nominal, b. ordinal, c. ordinal, d. nominal, e. nominal, f. ordinal

3.a. Binomial, $n = 100$, $\pi = 0.25$.
b. The mean is $n\pi = 25$ and the standard deviation is $\sqrt{n\pi(1-\pi)} = 4.33$. 50 correct responses would be surprising, since 50 is $z = (50 - 25)/4.33 = 5.8$ standard deviations above the mean of a distribution that is approximately normal.

4.a. Y is binomial for $n = 2$ and $\pi = 0.50$. Thus, $Y = 0$ with probability 0.25, $Y = 1$ with probability 0.50, and $Y = 2$ with probability 0.25. The mean is $2(0.50) = 1.0$ and the standard deviation is $\sqrt{2(0.50)(0.50)} = 0.71$.
b.(i) $P(Y = 0) = 0.16, P(Y = 1) = 0.48, P(Y = 2) = 0.36$;
(ii) $P(Y = 0) = 0.36, P(Y = 1) = 0.48, P(Y = 2) = 0.16$.
c. $\ell(\pi) = 2\pi(1-\pi)$.
d. From the plot or using calculus by taking the derivative and setting it equal to 0, the function $\ell(\pi) = 2\pi(1-\pi)$ takes its maximum value at $\pi = 0.50$.

8.a. 0.294. b. $z = -14.9$, P-value < 0.0001. Conclude that minority of population would say 'yes.'
c. $p \pm 1.96\sqrt{p(1-p)/n}$ is $0.294 \pm 2.58(0.0133)$, or (0.26, 0.33).

12.a. $SE = 0$, and the z statistic equals $-\infty$.
b. CI is (0, 0); no, in the population we expect some vegetarians, even if the proportion is small.
c. $z = (0 - 0.50)/\sqrt{0.50(0.50)/25} = -5.0$, P-value < 0.0001.
d. Note $z = (0 - 0.133)/\sqrt{0.133(0.133)/25} = -1.96$, so 0.133 is the null value that has a P-value of 0.05.

15.a. $\sigma(p)$ equals the binomial standard deviation $\sqrt{n\pi(1-\pi)}$ divided by the sample size n.
b. $\sigma(p)$ takes its maximum value at $\pi = 0.50$ and its minimum at $\pi = 0$ and 1. If $\pi = 1$, for instance, every observation must be a success, and the sample proportion p equals π with probability 1.

Chapter 2

2.a. Sensitivity $= P(Y=1|X=1) = \pi_1$, specificity $= P(Y=2|X=2) = 1 - P(Y=1|X=2) = 1 - \pi_2$.
b.

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

c. $0.86(0.01)/[0.86(0.01) + 0.12(0.99)] = 0.0675$.
d.

		Test diagnosis		
		+	−	Total
True	disease	0.0086	0.0014	0.01
	no disease	0.1188	0.8712	0.99

Nearly all (99%) subjects do not have breast cancer. The 12% errors for them swamp (in frequency) the 86% correct cases for the relatively few subjects who truly have it. In the column corresponding to a positive test result, we see that a much higher proportion are in the 'no disease' category than the 'disease' category.

3.a. (i) $0.0000624 - 0.0000013 = 0.000061$, (ii) $62.4/1.3 = 48$, so the estimated probability of a gun-related death in U.S. was 48 times that in Britain.
b. Relative risk, as difference of proportions makes it misleadingly seem as if there is no effect.

5.a. Relative risk.
b. (i) $\pi_1 = 0.55\pi_2$, so $\pi_1/\pi_2 = 0.55$. (ii) $1/0.55 = 1.82$.

6.a. 0.0012, 10.78; relative risk, since difference of proportions makes it appear there is no association.
b. $(0.001304/0.998696)/(0.000121/0.999879) = 10.79$; this happens when the proportion in the first category is close to zero for each group.

7.a. The quoted interpretation is that of the relative risk. Should substitute *odds* for *probability*. It would be approximately correct if the probability of survival were close to 0 for females and for males.
b. For females, proportion $= 2.9/(1 + 2.9) = 0.744$. Odds for males $= 2.9/11.4 = 0.254$, so proportion $= 0.254/(1 + 0.254) = 0.203$.
c. $R = 0.744/0.203 = 3.7$.

8.a. $(0.847/0.153)/(0.906/0.094) = 0.574$.
b. This is interpretation for relative risk, not the odds ratio. The actual relative risk $= 0.847/0.906 = 0.935$; i.e., 60% should have been 93.5%.

12.a.

	Heart attack		
Group	Yes	No	Total
Placebo	193	19,749	19,942
Aspirin	198	19,736	19,934

b. 0.974. The sample odds of a heart attack were actually a bit less for the placebo group.
c. CI for log odds ratio is $-0.0262 \pm 1.96(0.1017)$, or $(-0.225, 0.173)$. CI for odds ratio is (0.80, 1.19). It is plausible that there is no effect. If there is an effect, it is relatively weak.

17.a. $X^2 = 25.0$, $df = 1, P < 0.0001$. b. $G^2 = 25.4$, $df = 1$; for each statistic, very strong evidence that incidence of heart attacks depends on aspirin intake.

18.a. $35.8 = (290)(168)/n$, where $n = 1362$.
b. $df = 4$, P-value < 0.0001, extremely strong evidence of an association .
c. Strong evidence that fewer people are in those cells in the population than if the variables were independent. e.g., in this sample the number in the first cell is 2.973 standard errors smaller than the estimated expected frequency.
d. Strong evidence that more people are in those cells in the population than if the variables were independent.

19.a. $G^2 = 187.6, X^2 = 167.8, df = 2$; very strong evidence of association ($P < 0.0001$).
b. The large negative standardized residuals of -11.85 for white Democrats and -11.77 for black Republicans show extremely strong evidence of fewer people in these cells than we'd expect if party ID were independent of race. The large positive standardized residuals of 11.85 for black Democrats and 11.77 for white Republicans show extremely strong evidence of more people in these cells than we'd expect if party ID were independent of race.
c. $G^2 = 24.1$ for comparing races on (Democrat, Independent) choice, and $G^2 = 163.5$ for comparing races on (Dem. + Indep., Republican) choice; extremely strong evidence that whites are more likely than blacks to be Republicans. (To get independent components, combine the two groups compared in the first analysis and compare them to the other group in the second analysis.)

21.a. No, the samples in the different columns are dependent, because subjects can select as many columns as they wish.
b.

```
---------------------
              A
Gender     Yes No
Men        60  40
Women      75  25
---------------------
```

24. For any "reasonable" significance test, whenever H_0 is false, the test statistic tends to be larger and the P-value tends to be smaller as the sample size increases. Even if H_0 is just slightly false, the P-value will be small if the sample size is large enough. Most statisticians feel we learn more by estimating parameters using confidence intervals than by conducting significance tests.

25.a. Total of the estimated expected frequencies in row i equals
$\sum_j (n_{i+}n_{+j}/n) = (n_{i+}/n) \sum_j n_{+j} = n_{i+}$.
b. Their odds ratio equals $(n_{1+}n_{+1}/n)(n_{2+}n_{+2}/n)/(n_{1+}n_{+2}/n)(n_{2+}n_{+1}/n) = 1$.

26.a. Chi-squared with $df = 1$.
b. Note that $Y_1 + Y_2$ can be expressed as the sum of squares of $df_1 + df_2$ standard normal variates.

29. Table has entries (7,8) in row 1 and (0,15) in row 2. $P = \binom{15}{7}\binom{15}{0}/\binom{30}{7} = 15!23!/8!30! =$ 0.003; strong evidence of better results for treatment than control.

33.b. 0.67 for white victims and 0.79 for black victims.
c. 1.18; yes, Simpson's paradox occurs, because marginal association is in different direction than partial associations; reason for switch in association is same as in the text example.

34. Yes, this would be an occurrence of Simpson's paradox. One could display the data as a $2 \times 2 \times K$ table, where rows = (Smith, Jones), columns = (hit, out) response for each time at bat, layers = (year 1,...,year K). This could happen if Jones tends to have relatively more observations (i.e., "at bats") for years in which his average is high. Here's an example for $K = 2$. Suppose in year 1 that Jones had 80 hits in 200 at bats (batting average = 0.400), and in year 2 Jones had 20 hits in 100 at bats (average = 0.200), so the overall average is (80+20)/(200+100) = 0.333, whereas suppose in year 1 Smith had 45 hits in 100 at bats (average = 0.450), and in year 2 Smith had 50 hits in 200 at bats (average = 0.250), so the overall average is (45+50)/(100+200) = 0.317. Then, Jones is lower in each year but higher overall.

35. The age distribution is relatively higher in Maine.

36. X = whether eat ice cream today (yes, no), Y = whether go to beach today (yes, no), Z = high temperature today. As Z increases, X and Y are more likely to be in the 'yes' category.

37.a. 0.18 for males and 0.32 for females; e.g., for male children, the odds that a white was a murder victim. were 0.18 times the odds that a nonwhite was a murder victim.
b. 0.21.

38.a. If X and Y are conditionally independent, the true XY conditional odds ratio equals 1 at each level of Z. Since the odds ratios are identical at all levels of Z, there is homogeneous association.

b. If there is not homogeneous association, then the odds ratios between X and Y are not identical at all levels of Z, so they cannot equal 1.0 at all levels of Z, so X and Y cannot be conditionally independent.

39.a. T, b. T, c. F, d. T, e. F.

Chapter 3

1. The link function determines the function of the mean that is predicted by the linear predictor in a GLM. The identity link models the binomial probability directly as a linear function of the predictors. It is not often used, because probabilities must fall between 0 and 1, whereas straight lines provide predictions that can be any real number. When the probability is near 0 or 1 for some predictor values or when there are several predictors, it is not unusual to get predicted probabilities below 0 or above 1. With the logit link, any real number predicted value for the linear model corresponds to a probability between 0 and 1.

2.a. $P = 3$.
b. Estimated proportion $\hat{\pi} = -0.0003 + 0.0304(0.0774) = 0.0021$. The actual value is $\pi_i/\hat{\pi}_i = 3.8$ times the predicted value, which together with Fig. 3.8 suggests it is an outlier.

5. The fit of the linear probability model is (i) 0.018 + 0.018(snoring), (ii) 0.018 + 0.036(snoring), (iii) −0.019 + 0.036(snoring). Slope depends on distance between scores; doubling the distance halves the slope estimate. The fitted values are identical for any linear transformation.

6. The least squares fit is the ML fit assuming a normal distribution, so it is not the same as the ML fit assuming a binomial distribution.

9.a. $\text{logit}(\hat{\pi}) = -3.556 + 0.0532x$, where x = income and $\hat{\pi}$ is the estimated probability of possessing a travel credit card.
b. Since $\hat{\beta} = 0.0532 > 0$, the estimated probability of possessing a travel credit card increases as annual income increases.
c. Substituting $x = 66.84$ gives an estimated logit of 0 and thus an estimated probability of 0.50.

11.a. $\log(\hat{\mu}_B) - \log(\hat{\mu}_A) = [\alpha + \beta(1)] - \alpha = \beta$.
b. $\log(\hat{\mu}) = 1.609 + 0.588x$. Since $\beta = \log(\mu_B/\mu_A)$, $\exp(\hat{\beta}) = \hat{\mu}_B/\hat{\mu}_A = 1.80$; i.e., the mean is estimated to be 80% higher for treatment B. (In fact, this estimate is simply the ratio of sample means.)
c. Wald test gives $z = 0.588/0.176 = 3.33, z^2 = 11.1(df = 1), P < 0.001$. Likelihood-ratio statistic equals $27.86 - 16.27 = 11.6$ with $df = 1, P < 0.001$; higher defect rate for treatment B.
d. Exponentiate 95% CI for β of 0.588 ± 1.96(0.176) to get (1.27, 2.54).

12. Model with main effects and no interaction has fit $\log(\hat{\mu}) = 1.72 + 0.59x - 0.23z$; This shows some tendency for a lower rate of imperfections at the high thickness level ($z = 1$), although the SE of -0.23 equals 0.17. Adding an interaction (cross-product) term does not help, as the coefficient of the cross product of 0.27 has a SE of 0.36.

13.a. $\log(\hat{\mu}) = -0.428 + 0.589(\text{weight})$; b. 2.74.
c. $0.589 \pm 1.96(0.065) = 0.589 \pm 0.127$, or (0.462, 0.717); the CI for the multiplicative effect on the mean is (1.59, 2.05).
d. $z^2 = (0.589/0.065)^2 = 82.2$; extremely strong evidence that weight has a positive effect.
e. Likelihood-ratio statistic = 71.9, df = 1; extremely strong evidence that weight has a positive effect.

14.a. $\log(\hat{\mu}) = -0.865 + 0.760(\text{weight})$. Estimated dispersion parameter = 1.07 has $SE = 0.19$, so there is strong evidence of overdispersion. The model gives a better fit than the Poisson model.
b. $-0.760 \pm 1.96(0.177)$, or (0.41, 1.11), compared to (0.46, 0.72) for the Poisson model. The Poisson CI is unrealistically narrow because it does not take into account the overdispersion.

16.a. No, because there is clear evidence of substantial overdispersion.
b. Because the Poisson model does not take into account the overdispersion, hence giving an unrealistically small SE.
c. The negative binomial CI is more appropriate. The Poisson CI is overoptimistic, because it does not take into account the overdispersion.

17. CI for log rate is $2.549 \pm 1.96(0.04495)$, so CI for rate is (11.7, 14.0).

18.a. It is sensible to expect the number of arrests to be proportional to the attendance.
b. Model fit is $\log[\hat{E}(Y)/t] = -0.910$, where the estimate -0.910 has $SE = 0.022$; $\hat{\mu} = \exp(-0.910) = 0.402$.
c. Much larger: Aston Villa, Bradford City, Bournemouth, West Brom, Huddersfield, Birmingham, Shrewsbury. Much smaller: Sheffield Utd., Stoke City, Millwall, Hull City, Manchester City, Plymouth, Reading, Oldham.
d. Model fit is $\log[\hat{E}(Y)/t] = -0.905$, where the estimate -0.905 has $SE = 0.120$. The much larger SE, and the estimate of 0.32 of the dispersion parameter ($SE = 0.09$) suggests that the Poisson model does not permit sufficient variability; that is, there is overdispersion for that model.

20.a. The ratio of the rate for smokers to nonsmokers decreases markedly as age increases.
b. No, because (a) shows that the sample ratio is far from constant.
c. A quantitative interaction allows for a linear trend over the age values in the log of the ratio of death rates.
d. For main effects model, deviance $= 12.1, df = 4$. For age scores (1,2,3,4,5), model with interaction term has deviance $= 1.5, df = 3$. The difference of deviances $12.1 - 1.5 = 10.6$ ($df = 1$) is the

likelihood-ratio statistic comparing the models. The interaction model fits significantly better. The estimated interaction parameter = -0.309, with $SE = 0.097$, so the estimated ratio of rates is multiplied by $\exp(-0.309) = 0.73$ for each successive increase of one age category.

21. $\mu = \alpha t + \beta(tx)$, which has the form of a GLM with identity link, predictors t and tx, and no intercept term.

22.a. T, b. F. c. F.

Chapter 4

1.. The prediction equation gives $\hat{\pi} = e^{-3.7771+0.1449(8)}/[1 + e^{-3.7771+0.1449(8)}] = 0.068$.
b. $\hat{\pi} = 0.50$ at $-\hat{\alpha}/\hat{\beta} = 3.7771/0.1449 = 26$.
c. At LI = 8, $\hat{\pi} = 0.068$, so rate of change is $\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = 0.1449(0.068)(0.932) = 0.009$.
d. At LI = 14, $\hat{\pi} = e^{-3.7771+0.1449(14)}/[1 + e^{-3.7771+0.1449(14)}] = 0.15$.
e. $e^{\hat{\beta}} = e^{.1449} = 1.16$.

2.a. Wald statistic = $(0.1449/0.0593)^2 = 5.96$, $df = 1$, P-value = 0.0146 for H_a:$\beta \neq 0$.
b. CI is (1.03, 1.30), so the odds of remission at $LI = x + 1$ are estimated to fall between 1.03 and 1.30 times the odds of remission at $LI = x$.
c. Likelihood-ratio statistic = $34.37 - 26.07 = 8.30$, $df = 1$, P-value = 0.004.
d. Exponentiating (0.0425, 0.2846) from the table gives (1.04, 1.33). The odds of remission at $LI = x+1$ are estimated to fall between 1.04 and 1.33 times the odds of remission at $LI = x$.

4.a. The estimated probability of a heart attack increases as the level of snoring increases.
b. Estimated probabilities are 0.021, 0.132.
c. Multiplicative effect on odds equals $\exp(0.397) = 1.49$ for one-unit change in snoring, and 2.21 for two-unit change.

8.a. $\text{logit}(\hat{\pi}) = -3.695 + 1.815(\text{weight})$.
b. At 1.20 the estimated probability = 0.18, at 2.44 the estimated probability = 0.68, and at 5.20 the estimated probability = 0.997.
c. Estimated probability = 0.50 at weight = $3.695/1.815 = 2.04$.
d. (i) $\hat{\beta}/4 = 0.45$; (ii) for a 0.10kg increase, the estimated increase in the probability is 0.045.
e. CI for β of $1.815 \pm 1.96(0.3767)$, or (1.077, 2.5535), gives one (exponentiating) for the effect $\exp(\beta)$ on the odds of (2.9, 12.9).
f. Wald statistic $z^2 = 23.4$, likelihood-ratio statistic = 30.0, both based on $df = 1$, provide extremely strong evidence of a weight effect (P < 0.0001).

11. Odds ratio for spouse vs. others = $2.02/1.71 = 1.18$; odds ratio for \$10,000-24,999 vs. \$25,000+

equal $0.72/0.41 = 1.76$.

16.a. Using indicator variables that are 1 for E, S, T, J and 0 for the I, N, F, P ends of the scales, $\text{logit}(\hat{\pi}) = -2.47 + 0.55EI - 0.43SN + 0.69TF - 0.20JP$.
b. $e^{-1.86}/(1 + e^{-1.86}) = 0.14$.
c. Negative estimates for the S/N and J/P scales indicate higher probabilities of drinking frequently at the second category of each (N and P). The estimated probability is $e^{-2.47+0.55+0.69}/(1+e^{--2.47+0.55+0.69}) = 0.23$.

17.a. $e^{-2.83}/(1 + e^{-2.83}) = 0.056$.
b. $e^{0.5805} = 1.79$. Given the T/F category, the estimated odds an extroverted person drinks frequently is 1.79 times the estimated odds an introverted person drinks frequently.
c. $(e^{0.159}, e^{1.008}) = (1.17, 2.74)$.
d. $1/1.79 = 0.56$, CI $(1/2.74, 1/1.17) = (0.36, 0.85)$.
e. H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$, likelihood-ratio statistic $= 7.28$, $df = 1$, P-value $= 0.007$, strong evidence of an effect.

19.a. Controlling for religion and political party, the estimated difference in logits between females and males is $\hat{\beta}_1^G - \hat{\beta}_2^G = 0.16$. The odds of females supporting legalized abortion are estimated to be $\exp(0.16) = 1.17$ times the odds for males (i.e., odds ratio between opinion and gender is 1.17, controlling for R and P).
b. Using $\hat{\pi} = \exp(\text{logit})/[1+\exp(\text{logit})]$, we get (i) 0.08, (ii) 0.71. (Note: solutions at back of text have error)
c. Now $\hat{\beta}_2^G = -0.16$, but the estimated odds ratio is still $\exp(0.16) = 1.17$. (Note: solutions at back of text have error)
d. $\hat{\beta}_1^G = 0.08, \hat{\beta}_2^G = -0.08$.

21.a. Odds of obtaining condoms for the educated group estimated to be 4.04 times the odds for the non-educated group.
b. $\text{logit}(\hat{\pi}) = \hat{\alpha} + 1.40x_1 + 0.32x_2 + 1.76x_3 + 1.17x_4$, where $x_1 = 1$ for educated and 0 for non-educated, $x_2 = 1$ for males and 0 for females, $x_3 = 1$ for high SES and 0 for low SES, and $x_4 =$ lifetime no. of partners. The log odds ratio of 1.40 has confidence interval (0.16, 2.63). The interval is $1.40 \pm 1.96(SE)$. So the interval has width $3.92(SE)$, and $SE = 0.63$.
c. The confidence interval corresponds to one for the log odds ratio of (0.207, 2.556); 1.38 is the midpoint of this interval, suggesting that it may be the estimated log odds ratio, in which case $\exp(1.38) = 3.98$ is the estimated odds ratio.

23.a. $R = 1$: $\text{logit}(\hat{\pi}) = -6.7 + 0.1A + 1.4S$. $R = 0$: $\text{logit}(\hat{\pi}) = -7.0 + .1A + 1.2S$.
The YS conditional odds ratio is $\exp(1.4) = 4.1$ for blacks and $\exp(1.2) = 3.3$ for whites. Note that 0.2, the coefficient of the cross- product term, is the difference between the log odds ratios 1.4 and 1.2.

b. The coefficient of S of 1.2 is the log odds ratio between Y and S when $R = 0$ (whites), in which case the RS interaction does not enter the equation. The P-value of $P < 0.01$ for smoking represents the result of the test that the log odds ratio between Y and S for whites is 0.

27.a. -0.41 and 0.97 are the coefficients for standardized versions of the predictors for which the standard deviation is 1.0. That is, they refer to the change in the estimated logit for a one standard deviation increase in the predictor, controlling for the other predictor.
b. For $c = 4$ (dark crabs), $\text{logit}(\hat{\pi}) = -12.11 + 0.458x$. The estimated probability changes from 0.33 to 0.64 when x changes from 24.9 to 27.7.

28. $\hat{\pi}$ increases from $e^{-1.89}/(1 + e^{-1.89}) = 0.13$ to 0.95, which is quite a substantial cumulative effect.

35.a. The exponential term is maximized when the exponent equals 0, which happens when $x = -\alpha/\beta$.
b. 24.8.
c. $0.40(0.302) = 0.12$.

36.a. $\mu = 12.351/0.497 = 24.85$ and $\sigma = 1.814/0.497 = 3.65$.
b. $24.85 \pm 2(3.65)$ equals (17.5, 32.2).

37. a. T, b. F, c. T, d. F, e. T.

Chapter 5

4.a. Deviance = 11.1, *df* = 11, so no evidence of lack of fit. Model is adequate.
b. Take out JP term, as it is the least significant (Wald statistic = 0.80, *df* = 1, which has P-value = 0.37).
c. Likelihood-ratio statistic = $11.15 - 3.74 = 7.4$, *df* = 6, for which P-value = 0.28. The simpler model without interaction terms is adequate.

15. Logit model with additive factor effects for age and gender has $G^2 = 0.1$ and $X^2 = 0.1$ with *df* = 2. Estimated odds of females still being missing are exp(0.38) = 1.46 times those for males, given age. Estimated odds are considerably higher for those aged at least 19 than for other two age groups, given gender.

19.a. $\text{logit}(\pi) = \alpha + \beta_1 d_1 + \cdots + \beta_6 d_6$, where $d_i = 1$ for observations from department i and $d_i = 0$ otherwise.
b. Models fits poorly.
c. The only lack of fit of the model is in Department 1, where more females were admitted than one

would expect if the model lacking a gender effect truly holds.
d. -4.15, so fewer males were admitted than one would expect if the model lacking a gender effect truly holds. The residuals for males and females have the same absolute value but differ in sign.
e. Males apply in relatively greater numbers to the departments that have relatively higher proportions of acceptances.

20.a. $\text{logit}(\hat{\pi}) = -5.96 + 0.32(\text{alcohol})$. Deviance $= 1.95$ and Pearson statistic $= 2.05$, with $df = 3$, so model fits adequately.
b. Likelihood-ratio test stat. $= 6.20 - 1.95 = 4.25$, $df = 1$, $P = 0.04$, so relatively strong evidence of an alcohol effect.
c. Likelihood ratio test stat. $= 0.88$, $P = 0.35$, so now there is not much evidence of an alcohol effect.
d. $\text{logit}(\hat{\pi}) = -5.98 + 0.228(\text{alcohol})$. Likelihood-ratio test statistic $= 1.75$ ($df = 1$), $P = 0.19$, so there is not much evidence of an alcohol effect.

22.a. The "best fit" would have predicted value $\hat{\pi} = 0$ for $x \leq 40$ and $\hat{\pi} = 1$ for $x \geq 60$. But such a curve is a limit of a sequence of logistic regression curves in which $\hat{\beta}$ increases toward ∞. (The "separation theorem" states that if a plane can be passed through the space of values of the predictor variables in such a way that $y = 1$ for all data on one side of the plane and $y = 0$ for all data on the other side of the plane, then an infinite parameter estimate occurs.)
b. PROC GENMOD in SAS reports $\hat{\alpha} = -192.2$ and $\hat{\beta} = 3.84$. c. Still, $\hat{\beta} = \infty$, but software may not recognize this. e.g., PROC GENMOD in SAS reports $\hat{\alpha} = -132.3$ and $\hat{\beta} = 2.65$.
d. Yes, now there is no longer separation of x-values between those with $y = 0$ and those with $y = 1$, and the ML estimate is no longer infinite. Now, $\hat{\alpha} = -26.39$ and $\hat{\beta} = 0.53$.

30.a. F, b. T, c. T

Chapter 6

1. a. $\log(\hat{\pi}_R/\hat{\pi}_D) = -2.3 + 0.5x$. Since $\exp(0.5) = 1.65$, the estimated odds of preferring Republicans over Democrats increase by 65% for every \$10,000 increase in annual income.
b. $\hat{\pi}_R > \hat{\pi}_D$ when annual income $>$ \$46,000. (The logit equals 0, and hence the two estimated probabilities are the same, when $x = 4.6$ in the equation in (a).)
c. $\hat{\pi}_I = 1/[1 + \exp(3.3 - 0.2x) + \exp(1 + 0.3x)]$.

5.a. Job satisfaction tends to increase at higher levels of x_1 and lower levels of x_2 and x_3.
b. $x_1 = 4$ and $x_2 = x_3 = 1$.

6.a. $\log(\hat{\pi}_1)/\hat{\pi}_3) = -2.555 - 0.2275x$, $\log(\hat{\pi}_2)/\hat{\pi}_3) = -0.351 - 0.0962x$.
b. The estimated odds of being in the lower category (less happy) decreases as income increases.

c. Wald statistic $= 0.94$, $df = 2$, P-value $= 0.62$, plausible that income has no effect on marital happiness.
d. Deviance $= 3.19$, $df = 2$, P-value $= 0.20$, model fits adequately.
e. $1/[1 + \exp(-2.555 - 0.2275(2)) + \exp(-0.351 - 0.0962(2))] = 0.61$.

7.a. With 3 response categories, there are two cumulative probabilities to model, and hence 2 intercept parameters. The proportional odds form of model has the same predictor effects for each cumulative probability (i.e., the curves have the same shape), so only one effect is reported for income.
b. The estimated odds of being at the low end of the scale (less happy) decrease as income increases.
c. The likelihood-ratio statistic equals 0.89 with $df = 1$, and a P-value of 0.35. It is plausible that income has no effect on marital happiness.
d. Deviance $= 3.25$, $df = 3$, P-value $= 0.36$, so model fits adequately.
e. $1 -$ cumulative probability estimate for category 2, which is $1 - \exp(-0.2378 - 2(0.1117))/[1 + \exp(-0.2378 - 2(0.1117))] = 0.61$

12.a. The cumulative logit model of proportional odds form, using scores (1,2,3) for religious attendance, has ML effect estimate $\hat{\beta} = -0.384$. The estimated probability of being relatively unhappy decreases as religious attendance increases. The effect is strongly significant, with Wald statistic $= 71.3$ ($df = 1$) for testing H_0: $\beta = 0$.
b. The deviance statistic (reported by PROC LOGISTIC in SAS) is 0.62 ($df = 3$), so the fit is adequate.

22.a. T, b. T, c. F, d. T.

Chapter 7

5.a. 0.42, b. 1.45, conditional odds ratio less than 1.0, but marginal odds ratio greater than 1.0.
c. $G^2 = 0.38, df = 1, P = 0.54$, the model fits adequately.
d. Logit model with main effects of defendant race and victim race, using indicator variable for each.

6.a. Deviance $G^2 = 135.9$ and Pearson $X^2 = 145.1$ ($df = 11$), so model fits poorly.
b. Deviance $G^2 = 10.16$ and Pearson $X^2 = 10.10$ ($df = 5$), so model fits much better. The parameter estimate for the conditional log odds ratio between the S/N and J/P scales is 1.222, larger than any of the others in absolute value.
c. The parameter estimate for the conditional log odds ratio between the E/I and T/F scale is -0.194 ($SE = 0.131$), and the parameter estimate for the conditional log odds ratio between the E/I and J/P scales is 0.018 ($SE = 0.132$). The corresponding Wald chi-squared statistics are 2.20 and 0.02, each with $df = 1$.

7.a. Difference in deviances $= 12.37 - 10.16 = 2.21, df = 7 - 5 = 2$, so the simpler model is ade-

quate.
b. $\exp(-1.507, -0.938) = (0.22, 0.39)$.
c. $e^{1.220} = 3.39$, $\exp(0.938, 1.507)$ is $(1/0.39, 1/0.22)$, which is $(2.55, 4.51)$.

8.a. 1 intercept, 4 main effects, 6 two-factor association terms, 4 three-factor interaction terms, so numbers of parameters are 1+4, 1+4+6, 1+4+6+4.
b. AIC = -2(log likelihood - number parameters) minimized for model of homogeneous association.
27.a. T, b. F, c. T,

Chapter 8

2.a. z = (125 - 2)/$\sqrt{125+2}$ = 10.9 (or chi-squared = z^2 = 119.1), which has two-sided P-value < 0.0001. There is extremely strong evidence that more people believe in heaven than in hell.
b. Sample proportions equal 0.855 for heaven and 0.746 for hell. 90% CI is $0.110 \pm 1.645\sqrt{(125+2) - (125-2)^2/1120}/1120$, or 0.110 ± 0.016, or (0.094, 0.125).

4. The matched-pairs t test compares means for dependent samples, and McNemar‘s test compares proportions for dependent samples. The t test is valid for interval-scale data (with normally-distributed differences, for small samples) whereas McNemar’s test is valid for binary data.

7. $(0.314 - 0.292) \pm 1.96\sqrt{[0.314(0.686)/1144] + [0.292(0.708)/1144]}$, which is 0.022 ±0.038, or $(-0.016, 0.060)$, wider than for dependent samples.
8.a. Ignoring order, (A+,B-) occurred 45 times and (A-,B+)) occurred 22 times. The McNemar $z = 2.81$, which has a two-tail P-value of 0.005 and provides strong evidence that the response rate of successes is higher for drug A.

10.b. $z^2 = (3-1)^2/(3+1) = 1.0$ = CMH statistic.
c. When a partial table has a row or column total of 0, then conditional on the margins, that is the only possible table. So, in the first cell, the observed count = expected count with probability 1.0, the variance is 0 (as seen also from the variance formula in Section 4.3.4), and the stratum makes no contribution to the CMH statistic (4.9).
d. The ordinary P-value equals the binomial probability of 3 or more successes out of 4 trials when the success probability equals 0.50, which equals 5/16.

20.a. Using standardized residuals (Section 2.4.5), the neurologists both made ratings of 1 or both made ratings of 4 more often than the model of independence would predict.

Chapter 9

3.a. Marijuana: When let $S_1 = S_2 = 0$, the linear predictor takes greatest value when $R = 1$ and $G = 0$

(white males). For alcohol, let $S_1 = 1, S_2 = 0$, and then linear predictor takes greatest value when $R = 1$ and $G = 1$ (white females).
b. Race does not interact with gender or substance type, so the estimated odds for white subjects are $\exp(0.38) = 1.46$ times the estimated odds for black subjects.
c. For alcohol, estimated odds ratio $= \exp(-.20 + 0.37) = 1.19$; for cigarettes, $\exp(-.20 + .22) = 1.02$; for marijuana, $\exp(-.20) = .82$.
d. Estimated odds ratio $= \exp(1.93 + .37) = 9.97$.
e. Estimated odds ratio $= \exp(1.93) = 6.89$. The effect of alcohol use on marijuana use, controlling for cigarette use, seems to be a bit stronger for females than for males.

4. Similar substantive results occur. The GEE estimates, based on an exchangeable working correlation, show a linear time effect of 0.318 for the standard drug and 0.318 + 0.708 for the new one. The difference in slopes of 0.708 has a GEE empirical $SE = 0.136$, so the Wald statistic equals $(0.708/0.136)^2 = 5.2$ ($df = 1$, $P < 0.0001$) for comparing the slopes.

7. a. Subjects can select any number of sources, so given subject could have anywhere from 0 to 5 observations in the 40 cells of this table. Multinomial distribution (and ML fitting based on it) applies when each subject occurs in a single cell of the table.
b. Estimated correlation is weak, so results not much different from treating 5 responses by a subject as if from 5 independent subjects. The table gives the results for the five equations, one for each source, such as $-0.4875 - 0.2206s$ for the logit for source D. For source A, estimated size effect is 1.08 and highly significant (Wald statistic $= 6.46$, $df = 1$, $P < .0001$). For sources C, D, and E size effect estimates are all roughly -0.2.
18. True.