Section 5.2: Inference for Logistic Regression

5.6. Albert and Anderson (1984), Berkson (1951, 1953, 1955), Cox (1958a), Hodges (1958), and Walker and Duncan (1967) discussed ML estimation for logistic regression. For adjustments with complex sample surveys, see Hosmer and Lemeshow (2000, Sec. 6.4) and LaVange et al. (2001). Scott and Wild (2001) discussed the analyses of case–control studies with complex sampling designs.

5.7. Tsiatis (1980) suggested an alternative goodness-of-fit test that partitions values for the explanatory variables into a set of regions and adds a dummy variable to the model for each region. The test statistic compares the fit of this model to the simpler one, testing that the extra parameters are not needed. The idea of grouping values to check model fit by comparing observed and fitted counts extends to any GLM (Pregibon 1982). Hosmer et al. (1997) compared various ways of doing this.

Section 5.3: Logit Models with Categorical Predictors

5.8. The Cochran–Armitage trend test is locally asymptotically efficient for both linear and logistic alternatives for $P(Y = 1)$. Its efficiency against linear alternatives follows from the approximate normality of the sample proportions, with constant Bernoulli variance when $\beta = 0$. For the linear logit model (5.5), its efficiency follows from its equivalence with the score test. See Problem 9.35 and Cox (1958a) for related remarks. Tarone and Gart (1980) showed that the score test for a binary linear trend model does not depend on the link function. Gross (1981) noted that for the linear logit model, the local asymptotic relative efficiency for testing independence using the statistic with an incorrect set of scores equals the square of the Pearson correlation between the true and incorrect scores. Simon (1978) gave related asymptotic results. Corcoran et al. (2001), Mantel (1963), and Podgor et al. (1996) extended the trend test.

Section 5.4: Multiple Logistic Regression

5.9. Since the standardized logistic cdf has standard deviation $\pi/\sqrt{3}$, some software (e.g., PROC LOGISTIC in SAS) defines a standardized estimate by multiplying the unstandardized estimate by $s_{x_j}\sqrt{3}/\pi$.

PROBLEMS

Applications

5.1 For a study using logistic regression to determine characteristics associated with remission in cancer patients, Table 5.10 shows the most important explanatory variable, a labeling index (LI). This index measures proliferative activity of cells after a patient receives an injection of tritiated thymidine, representing the percentage of cells that are "labeled." The response Y measured whether the patient achieved remission (1 = yes). Software reports Table 5.11 for a logistic regression model using LI to predict the probability of remission.

TABLE 5.10 Data for Problem 5.1

LI	Number of Cases	Number of Remissions	LI	Number of Cases	Number of Remissions	LI	Number of Cases	Number of Remissions
8	2	0	18	1	1	28	1	1
10	2	0	20	3	2	32	1	0
12	3	0	22	2	1	34	1	1
14	3	0	24	1	0	38	3	2
16	3	0	26	1	1			

Source: Data reprinted with permission from E. T. Lee, *Comput. Prog. Biomed.* **4**: 80–92 (1974).

TABLE 5.11 Computer Output for Problem 5.1

```
                         Intercept     Intercept and
           Criterion       Only         Covariates
            -2 Log L      34.372          26.073

            Testing Global Null Hypothesis: BETA = 0
     Test                  Chi-Square     DF     Pr > ChiSq
     Likelihood Ratio        8.2988        1        0.0040
     Score                   7.9311        1        0.0049
     Wald                    5.9594        1        0.0146

 Parameter    Estimate   Standard Error   Chi-Square    Pr > ChiSq
 Intercept    -3.7771        1.3786         7.5064        0.0061
 li            0.1449        0.0593         5.9594        0.0146

                       Odds Ratio Estimates
     Effect     Point Estimate     95% Wald Confidence Limits
     li             1.156                  1.029    1.298

                    Estimated Covariance Matrix
           Variable       Intercept          li
           Intercept       1.900616      -0.07653
           li             -0.07653        0.003521

  Obs    li      remiss      n      pi_hat      lower       upper
   1      8        0         2     0.06797     0.01121     0.31925
   2     10        0         2     0.08879     0.01809     0.34010
```

a. Show how software obtained $\hat{\pi} = 0.068$ when LI $= 8$.

b. Show that $\hat{\pi} = 0.5$ when LI $= 26.0$.

c. Show that the rate of change in $\hat{\pi}$ is 0.009 when LI $= 8$ and 0.036 when LI $= 26$.

d. The lower quartile and upper quartile for LI are 14 and 28. Show that $\hat{\pi}$ increases by 0.42, from 0.15 to 0.57, between those values.

e. For a unit change in LI, show that the estimated odds of remission multiply by 1.16.

f. Explain how to obtain the confidence interval reported for the odds ratio. Interpret.

g. Construct a Wald test for the effect. Interpret.

h. Conduct a likelihood-ratio test for the effect, showing how to construct the test statistic using the $-2 \log L$ values reported.

i. Show how software obtained the confidence interval for π reported at LI = 8. (*Hint:* Use the reported covariance matrix.)

TABLE 5.12 Data for Problem 5.2[a]

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	2	70	1	3	69	0	4	68	0	5	67	0
6	72	0	7	73	0	8	70	0	9	57	1	10	63	1
11	70	1	12	78	0	13	67	0	14	53	1	15	67	0
16	75	0	17	70	0	18	81	0	19	76	0	20	79	0
21	75	1	22	76	0	23	58	1						

[a] Ft, flight number; Temp, temperature (°F); TD, thermal distress (1, yes; 0, no).
Source: Data based on Table 1 in *J. Amer. Statist. Assoc.*, **84:** 945–957, (1989), by S. R. Dalal, E. B. Fowlkes, and B. Hoadley. Reprinted with permission from the *Journal of the American Statistical Association.*

5.2 For the 23 space shuttle flights before the *Challenger* mission disaster in 1986, Table 5.12 shows the temperature at the time of the flight and whether at least one primary O-ring suffered thermal distress.

a. Use logistic regression to model the effect of temperature on the probability of thermal distress. Plot a figure of the fitted model, and interpret.

b. Estimate the probability of thermal distress at 31°F, the temperature at the place and time of the *Challenger* flight.

c. Construct a confidence interval for the effect of temperature on the odds of thermal distress, and test the statistical significance of the effect.

d. Check the model fit by comparing it to a more complex model.

5.3 Refer to Table 4.2. Using scores $\{0, 2, 4, 5\}$ for snoring, fit the logistic regression model. Interpret using fitted probabilities, linear approximations, and effects on the odds. Analyze the goodness of fit.

5.4 Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, severe forward flexion of the spine following corrective spinal surgery. The age in months at the time of the operation for the 18 subjects for whom kyphosis was present were 12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 139, 157