obtain a certain power is difficult to determine, because it depends on the distribution of the x values.

Under the assumption that X is random and has a normal distribution, the required sample size depends on the probability of success $\bar{\pi}$ at the mean of X and the odds ratio θ comparing $\bar{\pi}$ to the probability of success one standard deviation above the mean. Let $\lambda = \log(\theta)$. For a one-sided test,[11]

$$n = [z_\alpha + z_\beta \exp(-\lambda^2/4)]^2(1 + 2\bar{\pi}\delta)/(\bar{\pi}\lambda^2),$$

where

$$\delta = [1 + (1 + \lambda^2)\exp(5\lambda^2/4)]/[1 + \exp(-\lambda^2/4)].$$

The value n decreases as $\bar{\pi}$ gets closer to 0.50 and as $|\lambda|$ gets farther from the null value of 0.

A multiple logistic regression model requires larger n to detect a partial effect of the same size. Let R denote the multiple correlation between the explanatory variable x_j of primary interest and the others in the model. In this formula for n, we divide by $(1 - R^2)$, with $\bar{\pi}$ being the probability at the mean value of all the explanatory variables and θ being the effect of x_j at the mean of the others. However, this result is of limited use, because even if $R = 0$, an effect in a multiple logistic regression model changes in magnitude when variables are added to a model.[12]

These formulas provide, at best, rough ballpark indications of sample size. In most applications, we have only a crude guess for $\bar{\pi}$, θ, and R, and the explanatory variable of main interest may be far from normally distributed.

5.6.3 Example: Modeling the Probability of Heart Disease

A research study plans to model how the probability of severe heart disease depends on $x =$ cholesterol level for a middle-aged population. Previous studies have suggested that $\bar{\pi}$ is about 0.08. Suppose the investigators want the test of H_0: $\beta_1 = 0$ against H_a: $\beta_1 > 0$ to be sensitive to a 50% increase, for a standard deviation increase in cholesterol. The odds of severe heart disease at the mean cholesterol level equal 0.08/0.92 = 0.087, and the odds one standard deviation above the mean equal 0.12/0.88 = 0.136. The odds ratio equals $\theta = 0.136/0.087 = 1.57$, from which $\lambda = \log(1.57) = 0.450$ and $\delta = 1.306$. For $\beta = P$(Type II error) $= 0.10$ in an $\alpha = 0.05$-level test, $z_{.05} = 1.645$, $z_{.10} = 1.28$, and the study needs $n = 612$.

EXERCISES

5.1 For the horseshoe crabs data file, fit a model using weight and width as explanatory variables for the probability of a satellite.

a. Conduct a likelihood-ratio test of H_0: $\beta_1 = \beta_2 = 0$. Interpret.

[11] Due to F.Y. Hsieh, *Statist. Medic.*, **8**: 795–802 (1989).

[12] For example, see the article by L. Robinson and N. Jewell in *Intern. Statist. Rev.* **58**: 227–240 (1991).

b. Conduct separate likelihood-ratio tests for the partial effects of each variable. Why does neither test show evidence of an effect when the test in (**a**) shows very strong evidence?

c. Use purposeful selection or AIC to build a model when weight and the spine condition and color factors are the potential explanatory variables.

5.2 Table 7.8 in Chapter 7 shows data from the `Substance2` data file at the text website. Create a new data file from which you can build a logistic regression model for these data, treating marijuana use as the response variable and alcohol use, cigarette use, gender, and race as explanatory variables. Prepare a short report summarizing a model selection process, with edited software output as an appendix.

5.3 The `Crabs2` data file at the text website shows several variables that may be associated with y = whether a female horseshoe crab is monandrous (eggs fertilized by a single male crab) or polyandrous (eggs fertilized by multiple males). Using *year* of observation, *Fcolor* = the female crab's color (1 = dark, 3 = medium, 5 = light), *Fsurf* = her surface condition (values 1, 2, 3, 4, 5 with lower values representing worse), *FCW* = female's carapace width, *AMCW* = attached male's carapace width, *AMcolor* = attached male's color, and *AMsurf* = attached male's surface condition, conduct a logistic model-building process. Prepare a report summarizing this process, with edited software output as an appendix. Interpret results for your chosen model.

5.4 The `Students` data file at the text website shows responses of a class of social science graduate students at the University of Florida to a questionnaire that asked about *gender* (1 = female, 0 = male), *age*, *hsgpa* = high school GPA (on a four-point scale), *cogpa* = college GPA, *dhome* = distance (in miles) of the campus from your home town, *dres* = distance (in miles) of the classroom from your current residence, *tv* = average number of hours per week that you watch TV, *sport* = average number of hours per week that you participate in sports or have other physical exercise, *news* = number of times a week you read a newspaper, *aids* = number of people you know who have died from AIDS or who are HIV+, *veg* = whether you are a vegetarian (1 = yes, 0 = no), *affil* = political affiliation (1 = Democrat, 2 = Republican, 3 = Independent), *ideol* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), *relig* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week), *abor* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *affirm* = support affirmative action (1 = yes, 0 = no), and *life* = belief in life after death (1 = yes, 2 = no, 3 = undecided).

a. Show all steps of a model-selection method such as purposeful selection for choosing a model for predicting *abor*, when the potential explanatory variables are *ideol*, *relig*, *news*, *hsgpa*, and *gender*.

b. Using an automated tool such as the `stepAIC` or `bestglm` function in `R`, construct a model to predict *abor*, selecting from the 14 binary and quantitative variables in the data file as explanatory variables.

c. With y = *veg* and the 14 binary and quantitative variables in the data file as explanatory variables, show that the likelihood-ratio test of H_0: $\beta_1 = \cdots =$

$\beta_{14} = 0$ has P-value < 0.001, yet forward selection using Wald tests with 0.05 criterion selects the null model. Explain how this could happen.

5.5 Exercise 4.12 introduced four scales of the Myers–Briggs personality test. Table 5.6 shows SAS output for fitting a model using the four scales as predictors of whether a subject drinks alcohol frequently.

a. Conduct a model goodness-of-fit test, and interpret. If you were to simplify the model by removing a predictor, which would you remove? Why?

b. Software reports AIC values of 642.1 for the model with the four main effects and the six interaction terms, 637.5 for the model with only the four binary main effect terms, and 648.8 for the model with no predictors. According to this criterion, which model is preferred? Explain the rational for using AIC.

c. Using the `MBTI` data file at the website `www.stat.ufl.edu/~aa/intro-cda/data`, use model-building methods to select a model for this alcohol response variable.

Table 5.6 SAS output for fitting model to Myers–Briggs personality scales data of Exercise 4.12.

```
-----------------------------------------------------------------------------
     Criterion                     DF          Value
     Deviance                      11        11.1491
                                 Standard   Like-ratio 95%         Chi-
 Parameter           Estimate      Error     Conf Limits           Square
 Intercept           -2.4668     0.2429   -2.9617    -2.0078      103.10
 EI            e      0.5550     0.2170    0.1314     0.9843        6.54
 SN            s     -0.4292     0.2340   -0.8843     0.0353        3.36
 TF            t      0.6873     0.2206    0.2549     1.1219        9.71
 JP            j     -0.2022     0.2266   -0.6477     0.2426        0.80
-----------------------------------------------------------------------------
```

5.6 Refer to the previous exercise. The data file also shows responses on whether a person smokes frequently. Software reports model -2 log-likelihood values of 1130.23 with only an intercept term, 1124.86 with also the main effect predictors, and 1119.87 with also all the two-factor interactions.

a. Write the model for each case and show that the numbers of parameters are 1, 5, and 11.

b. Find AIC values. Which of the three models is preferable?

5.7 For data introduced in Exercise 4.10 about AIDS symptoms, AZT use, and race, here is some `R` output:

```
-------------------------------------------------------------------------
> fit <- glm(yes/(yes+no) ~ azt + race, weights=yes+no, family=binomial,
+                 data=AIDS)
> summary(fit)
Deviance Residuals:
          1           2          3          4
-0.5547       0.4253     0.7035    -0.6326
```

```
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    -1.07357     0.26294   -4.083  4.45e-05
aztyes         -0.71946     0.27898   -2.579   0.00991
racewhite       0.05548     0.28861    0.192   0.84755
---
    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom
> 1 - pchisq(1.3835, 1)
[1] 0.23950
> cbind(AIDS$azt, AIDS$race, fitted(fit), rstandard(fit,type="pearson"),
       residuals(fit,type="pearson"), residuals(fit,type="deviance"))
  [,1] [,2]    [,3]     [,4]     [,5]     [,6]
1    2    2  0.1496  -1.1794  -0.5447  -0.5547 # azt=yes, race=white
2    1    2  0.2654   1.1794   0.4282   0.4253 # azt=no,  race=white
3    2    1  0.1427   1.1794   0.7239   0.7035 # azt=yes, race=black
4    1    1  0.2547  -1.1794  -0.6220  -0.6326 # azt=no,  race=black
-------------------------------------------------------------------------
```

a. Test the model goodness of fit and interpret the result.

b. Explain how the relative sizes of the fitted values reflect the results of the individual tests for the AZT effect and the race effect.

c. The display shows the standardized residuals, Pearson residuals, and deviance residuals. Explain advantages of using standardized residuals rather than the others.

5.8 Refer to Table 2.9 on death penalty decisions. Fit a logistic model with the two race predictors. Conduct a residual analysis and interpret.

5.9 Table 5.7 shows a $2 \times 2 \times 6$ contingency table for y = whether admitted to graduate school at the University of California, Berkeley, for fall 1973, by gender of applicant for the six largest graduate departments.

a. Fit the logistic model that has department as the sole explanatory variable for y. Use the standardized residuals to describe the lack of fit.

Table 5.7 Data for Exercise 5.9 on admissions to Berkeley.

	Admitted, Male		Admitted, 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: Based on data in P. Bickel *et al.*, *Science* **187**: 398–403 (1975).

```
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -1.07357     0.26294   -4.083  4.45e-05
aztyes        -0.71946     0.27898   -2.579   0.00991
racewhite      0.05548     0.28861    0.192   0.84755
---
    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom
> 1 - pchisq(1.3835, 1)
[1] 0.23950
> cbind(AIDS$azt, AIDS$race, fitted(fit), rstandard(fit,type="pearson"),
       residuals(fit,type="pearson"), residuals(fit,type="deviance"))
  [,1] [,2]    [,3]     [,4]     [,5]     [,6]
1    2    2  0.1496  -1.1794  -0.5447  -0.5547 # azt=yes, race=white
2    1    2  0.2654   1.1794   0.4282   0.4253 # azt=no,  race=white
3    2    1  0.1427   1.1794   0.7239   0.7035 # azt=yes, race=black
4    1    1  0.2547  -1.1794  -0.6220  -0.6326 # azt=no,  race=black
--------------------------------------------------------------------------
```

a. Test the model goodness of fit and interpret the result.

b. Explain how the relative sizes of the fitted values reflect the results of the individual tests for the AZT effect and the race effect.

c. The display shows the standardized residuals, Pearson residuals, and deviance residuals. Explain advantages of using standardized residuals rather than the others.

5.8 Refer to Table 2.9 on death penalty decisions. Fit a logistic model with the two race predictors. Conduct a residual analysis and interpret.

5.9 Table 5.7 shows a $2 \times 2 \times 6$ contingency table for $y =$ whether admitted to graduate school at the University of California, Berkeley, for fall 1973, by gender of applicant for the six largest graduate departments.

a. Fit the logistic model that has department as the sole explanatory variable for y. Use the standardized residuals to describe the lack of fit.

Table 5.7 Data for Exercise 5.9 on admissions to Berkeley.

	Admitted, Male		Admitted, 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: Based on data in P. Bickel *et al.*, *Science* **187**: 398–403 (1975).

b. When we add a gender effect, the estimated conditional odds ratio between admissions and gender (1 = male, 0 = female) is 0.90. The marginal table, collapsed over department, has odds ratio 1.84. Explain what causes these associations to differ so much.

5.10 The `Lungs` data file at the text website[13] summarizes eight studies in China about smoking and lung cancer. Analyze these data and prepare a short report that summarizes your analyses and interpretations.

5.11 Refer to the model you selected in part (**a**) of Exercise 5.4. Check goodness of fit. Can you conduct a residual analysis with this data file? Explain.

5.12 Suppose $y = 0$ at $x = 0, 10, 20, 30$ and $y = 1$ at $x = 70, 80, 90, 100$.

a. Explain intuitively why $\hat{\beta} = \infty$ for the model, $\text{logit}[P(Y = 1)] = \alpha + \beta x$. Report $\hat{\beta}$ and its SE for the software you use.

b. Add two observations at $x = 50$, $y = 1$ for one and $y = 0$ for the other. Report $\hat{\beta}$ and its SE. Do you think these are correct? Why? What happens if you replace the two observations by $y = 1$ at $x = 49.9$ and $y = 0$ at $x = 50.1$?

5.13 Refer to Exercise 5.4. With *veg* as the response variable, find a logistic model for which at least one ML effect estimate is infinite. Explain the aspect of the data file that causes this. Report and interpret results from fitting the model using either Firth's penalized logistic regression or Bayesian inference.

5.14 Table 5.8 is from a study of nonmetastatic osteosarcoma described in the *LogXact 7* manual (Cytel Software, 2005, p. 171). The response is whether the subject achieved a three-year disease-free interval.

a. Show that each explanatory variable has a significant effect when it is used as the sole predictor in logistic regression. Try to fit a main-effects model containing all three predictors. Explain why the ML estimate for the effect of lymphocytic infiltration is actually infinite.

b. Report and interpret results from fitting the main-effects model using either Firth's penalized logistic regression or Bayesian inference.

Table 5.8 Data for Exercise 5.14.

Lymphocytic Infiltration	Sex	Osteoblastic Pathology	Disease-Free Yes	Disease-Free No
High	Female	No	3	0
High	Female	Yes	2	0
High	Male	No	4	0
High	Male	Yes	1	0
Low	Female	No	5	0
Low	Female	Yes	3	2
Low	Male	No	5	4
Low	Male	Yes	6	11

[13] Based on data in *Intern. J. Epidemiol.*, **21**: 197–201 (1992) by Z. Liu.

Table 5.9 Clinical trial relating treatment to response for five centers.

		Response	
Center	Treatment	Success	Failure
1	Active drug	0	5
	Placebo	0	9
2	Active drug	1	12
	Placebo	0	10
3	Active drug	0	7
	Placebo	0	5
4	Active drug	6	3
	Placebo	2	6
5	Active drug	5	9
	Placebo	2	12

Source: Diane Connell, Sandoz Pharmaceuticals Corp.

5.15 Table 5.9 shows results of a randomized clinical trial conducted at five centers. The purpose was to compare an active drug to placebo for treating fungal infections (1 = success, 0 = failure). For these data, let y = response, x = treatment (1 = active, 0 = placebo), and z = center.

a. For the model logit$[P(Y = 1) = \alpha + \beta x + \beta_k^z$, explain why quasi-complete separation occurs in terms of the effects of center.

b. Using a "no intercept" option so that $\{\beta_k^z\}$ refer to the individual centers rather than contrasts with a baseline center, fit the model and report $\hat{\beta}_1^z$ and $\hat{\beta}_3^z$ and their standard errors. What are the actual ML estimates?

c. The counts in the 2×2 marginal table relating treatment to response are all positive, so the empty cells do not affect the treatment estimate. Report the estimated treatment log odds ratio and show that it does not change when you delete Centers 1 and 3 from the analysis. (When a center has outcomes of only one type, it provides no information about the treatment effect.)

5.16 Refer to Exercise 4.1 on cancer remission. Table 5.10 shows output for fitting a probit model. Interpret the parameter estimates (**a**) finding the remission value at which the estimated probability of remission equals 0.50, (**b**) finding the difference between the estimated probabilities of remission at the upper and lower quartiles of the labeling index, 14 and 28, (**c**) using a corresponding normal latent variable model, (**d**) using characteristics of the normal *cdf* response curve.

Table 5.10 Table for Exercise 5.16 on probit model for cancer remission.

```
-----------------------------------------------------------------
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -2.31777     0.76060   -3.047   0.00231
LI             0.08785     0.03293    2.668   0.00763
-----------------------------------------------------------------
```

5.17 Refer to the `SoreThroat` data file introduced in Exercise 4.16. Fit and interpret the main effects (**a**) linear probability model, (**b**) probit model.

5.18 For the `Crabs` data file, fit the linear probability model to the probability that a female horseshoe crab with shell width x has a satellite. Is the fit adequate for large x values?

5.19 We expect success probabilities for two groups to be about 0.20 and 0.30, and we want an 80% chance of detecting a difference using a 90% confidence interval.

a. Assuming equal sample sizes, how large should they be?

b. Compare results to the sample sizes required for (i) a 90% interval with power 90%, (ii) a 95% interval with power 80%.

5.20 The width values in the `Crabs` data file have a mean of 26.3 and standard deviation of 2.1. If the true relationship is the fitted equation reported in Section 4.1.3, $\text{logit}[\pi(x)] = -12.351 + 0.497x$, about how large a sample yields P(Type II error) $= 0.10$ in an $\alpha = 0.05$-level test of H_0: $\beta = 0$ against H_a: $\beta > 0$? What assumption does this result require?

5.21 The following are true–false questions.

a. A model for a binary y has a continuous explanatory variable. If the model truly holds, the residual deviance has a distribution approaching chi-squared as n increases. It can be used to test model goodness of fit.

b. When x_1 or x_2 is the sole predictor for binary y, the likelihood-ratio test of the effect has P-value < 0.0001. When both x_1 and x_2 are in the model, it is possible that the likelihood-ratio tests for H_0: $\beta_1 = 0$ and for H_0: $\beta_2 = 0$ could both have P-values larger than 0.05.