those predictors. An advantage of R compared to R^2 is that it uses the original scale and it has value approximately proportional to the effect size; for instance, with a single quantitative explanatory variable, the correlation is the slope multiplied by the ratio of standard deviations of the two variables.

For a binary regression model, R is the correlation between the n binary $\{y_i\}$ observations (1 or 0 for each) and the fitted probabilities $\{\hat{\pi}_i\}$. The highly discrete nature of y can suppress the range of possible R values, more so when the imbalance between the frequencies of 0 and 1 values is greater. Also, like any correlation measure, the value of R depends on the range of values observed for the explanatory variables. Nevertheless, R is useful for comparing fits of different models for the same data.

Although this measure is not routinely provided by GLM software, it is simple to obtain, as shown here in R code:

```
> fit <- glm(y ~ width + factor(color), family=binomial, data=Crabs)
> cor(Crabs$y, fitted(fit))
[1] 0.45221
```

The simpler model that uses width and a dark-color indicator does essentially as well, with $R = 0.447$. Using width alone has $R = 0.402$.

The square of this measure does not have the proportional reduction in variation interpretation that it has for ordinary (least squares) regression. Various measures have been proposed in an attempt to do this for binary data. For example, one such measure approximates R^2 for an ordinary regression model presented in Section 5.5.3 for an underlying continuous variable for y. We present this approach for ordinal responses in Section 6.3.7, but it applies also for binary responses.

EXERCISES

4.1 A study[8] investigated characteristics associated with $y =$ whether a cancer patient achieved remission ($1 =$ yes, $0 =$ no). An important explanatory variable was a labeling index ($LI =$ percentage of "labeled" cells) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. Table 4.5 shows the data and R output for a logistic regression model.

a. Show that $\hat{P}(Y = 1) = 0.50$ when $LI = 26.0$.

b. When LI increases by 1, show that the estimated odds of remission multiply by 1.16.

c. Summarize the LI effect by how $\hat{P}(Y = 1)$ changes over the range or interquartile range of LI values.

d. Show that the rate of change in $\hat{P}(Y = 1)$ is 0.009 when $LI = 8$.

e. Summarize the LI effect by the estimated average marginal effect.

[8] Article by E.T. Lee, *Computer Prog. Biomed.* **4**: 80–92 (1974).

Table 4.5 Software output for Exercise 4.1 on cancer remission.

```
------------------------------------------------------------------------------
> LI <- c(8,8,10,10,12,12,12,14,14,14,16,16,16,18,20,20,20,22,22,24,26,28,32,34,
+         38,38,38)
> y <- c(0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)
> summary(glm(y ~ LI, family=binomial))
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -3.77714     1.37862   -2.740   0.00615
LI            0.14486     0.05934    2.441   0.01464
---
    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom

> confint(glm(y ~ LI, family=binomial))
                2.5 %     97.5 %
LI            0.04252    0.28467
------------------------------------------------------------------------------
```

4.2 Refer to the previous exercise. Using information from Table 4.5:

a. Conduct a Wald test for the LI effect and construct a 95% Wald confidence interval for the odds ratio corresponding to a 1-unit increase in LI. Interpret.

b. Conduct a likelihood-ratio test and construct a 95% profile likelihood interval. Interpret.

4.3 Refer to the previous two exercises. Set up the data file as 14 observations in grouped-data format. Compare to the `Remissions` data file at the text website. Fit the model with this data file. Are the ML model parameter estimates the same as with the ungrouped data file? Is the deviance the same? Why or why not?

4.4 For the snoring and heart disease data of Table 3.1 (Section 3.2.3) with snoring-level scores (0, 2, 4, 5), the logistic regression ML fit is $\text{logit}[\hat{P}(Y = 1)] = -3.866 + 0.397x$. Interpret the effect of snoring on the odds of heart disease.

4.5 For the 23 space shuttle flights before the Challenger mission disaster in 1986, Table 4.6 and the `Shuttle` data file at the text website shows the temperature (°F)

Table 4.6 Data for Exercise 4.5 on space shuttle.

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						

Note: Ft = flight no., Temp = temperature, 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 *J. Amer. Statist. Assoc.*

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. Interpret the effect.

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

c. At what temperature does the estimated probability equal 0.50? At that temperature, give a linear approximation for the change in the estimated probability per degree increase in temperature.

d. Interpret the effect of temperature on the odds of thermal distress.

e. Test the hypothesis that temperature has no effect.

4.6 For Exercise 3.9 on travel credit cards, use the logistic output there to (**a**) interpret the effect of income on the odds of possessing a travel credit card, and conduct (**b**) a significance test and (**c**) a confidence interval for that effect.

4.7 Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, which is severe forward flexion of the spine following corrective spinal surgery. The `Kyphosis` data file at the text website shows the 40 observations on $y =$ whether kyphosis is present (1 = yes), with $x =$ age as the explanatory variable.

a. Fit a logistic regression model. Test the effect of age.

b. Plot the data. Note the difference in dispersion of age at the two levels of kyphosis.

c. Fit the model $\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 x^2$. Test the significance of the squared age term, plot the fit, and interpret. (The final paragraph of Section 4.1.5 is relevant to these results.)

4.8 For the `Crabs` data file at the text website, fit the logistic regression model for the probability of a satellite ($y = 1$) using $x =$ weight as the sole explanatory variable.

a. Report the ML prediction equation. At the mean weight value of 2.437 kg, give a linear approximation for the estimated effect of (i) a 1-kg increase in weight. This represents a relatively large increase, so convert this to the effect of (ii) a 0.10-kg increase, (iii) a standard deviation increase in weight (0.58 kg).

b. Find and interpret the average marginal effect of weight per 0.10-kg increase.

c. Construct the classification table using the sample proportion of $y = 1$ as the cutoff. Report the sensitivity and specificity. Interpret.

d. Construct an ROC curve, and report and interpret the area under it.

4.9 For the `Crabs` data file, fit a logistic regression model for the probability of a satellite, using color alone as the predictor.

a. Treat color as a nominal-scale factor. Report the prediction equation and explain how to use it to compare the first and fourth colors.

b. For the model in (**a**), conduct a likelihood-ratio test of the hypothesis that color has no effect. Interpret.

c. Treating color in a quantitative manner (scores 1, 2, 3, 4), obtain a prediction equation. Interpret the coefficient of color and test the hypothesis that color has no effect.

d. When we treat color as quantitative instead of qualitative, state a potential advantage relating to power and a potential disadvantage relating to model lack of fit.

e. Using weight and quantitative color as explanatory variables, find standardized coefficients, and interpret.

4.10 In a study[9] on the effects of AZT in slowing the development of AIDS symptoms, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. Output follows of modeling the $2 \times 2 \times 2$ cross-classification of race, whether AZT was given immediately, and whether AIDS symptoms developed during the three-year study.

```
------------------------------------------------------------------------------
> AIDS # Data file at text website
   race azt yes no # yes and no are categories of AIDS symptoms response
1 white yes  14 93
2 white  no  32 81
3 black yes  11 52
4 black  no  12 43
> fit <- glm(yes/(yes+no) ~ azt + race, weights=yes+no, family=binomial,
+            data=AIDS)
> summary(fit)
                 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
------------------------------------------------------------------------------
```

a. What null hypothesis is tested by the difference between the null deviance and the residual deviance? Interpret.

b. Explain how to set up indicator variables for azt and race to obtain the estimates shown.

4.11 For Table 2.9 on racial characteristics and the death penalty, create a data file and fit a logistic model for death penalty as the response, with defendant's race and victims' race as predictors.

a. Report the model fit and interpret the parameter estimates. Based on those estimates, which group is most likely to have the *yes* response?

b. Conduct inference about the effect of victims' race, controlling for defendant's race. Interpret.

4.12 At the website `www.stat.ufl.edu/~aa/intro-cda/data` for the 2nd edition of this book, the `MBTI` data file cross-classifies a sample of people from the MBTI Step II National Sample on whether they report drinking alcohol frequently and on the four

[9] Described in the *New York Times*, Feb. 15, 1991.

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). The 16 predictor combinations correspond to the 16 personality types: ESTJ, ESTP, ESFJ, ESFP, ENTJ, ENTP, ENFJ, ENFP, ISTJ, ISTP, ISFJ, ISFP, INTJ, INTP, INFJ, INFP. (e.g., of the 77 people of type ESTJ, 13 reported smoking frequently.) Fit a model using the four scales as predictors of the probability of drinking alcohol frequently. Report the prediction equation, specifying how you set up the indicator variables. Based on the model parameter estimates, explain why the personality type with the highest estimated probability of drinking alcohol is ENTP.

4.13 For first-degree murder convictions[10] in East Baton Rouge Parish, Louisiana, between 1990 and 2008, the death penalty was given in 3 out of 25 cases in which a white killed a white, in 0 out of 3 cases in which a white killed a black, in 9 out of 30 cases in which a black killed a white, and in 11 out of 132 cases in which a black killed a black. Table 4.7 shows software output for fitting a logistic regression model, where $d = 1$ ($d = 0$) for black (white) defendants and $v = 1$ ($v = 0$) for black (white) victims. Summarize in a short, nontechnical report what you learn from this output.

Table 4.7 Logistic regression fit of death penalty data for Exercise 4.13.

```
                Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)     -2.0232       0.6137    -3.297   0.000978
d                1.1886       0.7236     1.643   0.100461
v               -1.5713       0.5028    -3.125   0.001778
Residual deviance: 0.16676  on 1  degrees of freedom
```

4.14 Table 4.8 shows results of an eight-center clinical trial to compare a drug to placebo for curing an infection. At each center, subjects were randomly assigned to treatments.

Table 4.8 Clinical trial data for Exercise 4.14.

		Response				Response	
Center	Treatment	Success	Failure	Center	Treatment	Success	Failure
1	Drug	11	25	5	Drug	6	11
	control	10	27		control	0	12
2	Drug	16	4	6	Drug	1	10
	control	22	10		control	0	10
3	Drug	14	5	7	Drug	1	4
	control	7	12		control	1	8
4	Drug	2	14	8	Drug	4	2
	control	1	16		control	6	1

Source: P.J. Beitler and J.R. Landis, *Biometrics*, vol. 41, pp. 991–1000 (1985).

[10] From G. Pierce and M. Radelet, *Louisiana Law Review*, vol. 71 (2011), pp. 647–673.

Analyze these data, available in the `Infection` data file at the text website. Using logistic regression, describe and make inference about the treatment effect.

4.15 In a study designed to evaluate whether an educational program makes sexually active adolescents more likely to obtain condoms, adolescents were randomly assigned to two experimental groups. The educational program, involving a lecture and videotape about transmission of the HIV virus, was provided to one group but not the other. In logistic regression models, factors observed to influence a teenager to obtain condoms were gender, socioeconomic status (SES), lifetime number of partners, and the experimental group. Table 4.9 summarizes study results.

a. Find the parameter estimates for the fitted model, using (1, 0) indicator variables for the first three explanatory variables.

b. Explain why either the estimate of 1.38 for the odds ratio for gender or the corresponding confidence interval is incorrect. Show that if the reported interval is correct, then 1.38 is actually the *log* odds ratio, and the estimated odds ratio equals 3.98.

Table 4.9 Table for Exercise 4.15 on condom use.

Variables	Odds Ratio	95% Confidence Interval
Group (Education vs. None)	4.04	(1.17, 13.9)
Gender (Males vs. Females)	1.38	(1.23, 12.88)
SES (High vs. Low)	5.82	(1.87, 18.28)
Lifetime no. of Partners	3.22	(1.08, 11.31)

Source: Rickert *et al., Clin. Pediatrics*, 205–210 (1992).

4.16 Table 4.10, which is the `SoreThroat` data file at the text website, shows results of a study about $y =$ whether a patient having surgery with general anesthesia experienced a sore throat on waking ($1 =$ yes, $0 =$ no) as a function of $d =$ duration of the surgery

Table 4.10 Data for Exercise 4.16 on sore throat after surgery.

Patient	d	t	y	Patient	d	t	y	Patient	d	t	y
1	45	0	0	2	15	0	0	3	40	0	1
4	83	1	1	5	90	1	1	6	25	1	1
7	35	0	1	8	65	0	1	9	95	0	1
10	35	0	1	11	75	0	1	12	45	1	1
13	50	1	0	14	75	1	1	15	30	0	0
16	25	0	1	17	20	1	0	18	60	1	1
19	70	1	1	20	30	0	1	21	60	0	1
22	61	0	0	23	65	0	1	24	15	1	0
25	20	1	0	26	45	0	1	27	15	1	0
28	25	0	1	29	15	1	0	30	30	0	1
31	40	0	1	32	15	1	0	33	135	1	1
34	20	1	0	35	40	1	0				

Source: Data from D. Collett, pp. 350–358 in *Encyclopedia of Biostatistics* (Wiley: 1998). Predictors are $d =$ duration of surgery, $t =$ type of device.

(in minutes) and t = type of device used to secure the airway (0 = laryngeal mask airway, 1 = tracheal tube).

a. Fit a model permitting interaction between the explanatory variables. Report and interpret the prediction equation for the effect of d when (i) $t = 1$, (ii) $t = 0$. Conduct inference about whether you need the interaction term.

b. Compare the predictive power of models with and without the interaction term by finding the correlation R between the observed and fitted values for each model.

4.17 Table 4.11 shows estimated effects for a logistic regression model for y = presence of squamous cell esophageal cancer (1 = yes, 0 = no). Smoking status (s) equals 1 for at least one pack per day and 0 otherwise, alcohol consumption (a) equals the average number of alcoholic drinks consumed per day, and race (r) equals 1 for blacks and 0 for whites.

a. To describe the race-by-smoking interaction, construct the prediction equation when $r = 1$ and again when $r = 0$. Find the fitted conditional odds ratio for the smoking effect for each case. Similarly, construct the prediction equation when $s = 1$ and again when $s = 0$. Find the fitted conditional odds ratio for the race effect for each case. (For each association, the coefficient of the cross-product term is the difference between the log odds ratios at the two levels for the other variable.)

b. In Table 4.11, what do the coefficients of smoking and race represent? What hypotheses do their P-values refer to?

Table 4.11 Table for Exercise 4.17 on effects on esophageal cancer.

Variable	Effect	P-value
Intercept	−7.00	<0.01
Alcohol use	0.10	0.03
Smoking	1.20	<0.01
Race	0.30	0.02
Race × Smoking	0.20	0.04

4.18 For Table 4.12 from the 2016 General Social Survey, create a data file and analyze the data using logistic regression. Summarize your analyses in a short report, including edited output in an appendix.

Table 4.12 Data on belief in afterlife for Exercise 4.18.

		Belief in Afterlife	
Race	Religion	Yes	No or Undecided
White	Protestant	817	250
	Catholic	519	194
	Other	48	9
Black	Protestant	298	86
	Catholic	39	13
	Other	119	38

4.19 For model (4.3) for the horseshoe crabs with color and width predictors, add three terms to permit interaction between color and width.

a. Report the prediction equations relating width to the probability of a satellite, for each color. Plot or sketch them, and interpret.

b. Test whether the interaction model fits better than the simpler model without interaction terms. Interpret. Compare their predictive power by finding the correlation R between the observed and fitted values for each model.

4.20 Refer to Exercise 4.12 about MBTI and alcohol drinking.

a. When the sample proportion of 0.092 who reported drinking alcohol frequently is the cutpoint for forming a classification table, sensitivity $= 0.53$ and specificity $= 0.66$. Explain what these mean, and show that the sample proportion of correct classifications was 0.65.

b. The `MBTI` data file also shows responses on whether a person smokes frequently. When a classification table for the model containing the four main effect terms to predict smoking uses the sample proportion of frequent smokers of 0.23 as the cutoff, sensitivity $= 0.48$ and specificity $= 0.55$. The area under the ROC curve is 0.55. Does knowledge of personality type help you predict well whether someone is a frequent smoker? Explain.

4.21 Explain how the classification table in Table 4.4 with $\pi_0 = 0.50$ was constructed. Estimate the sensitivity and specificity, and interpret.

4.22 You plan to study the relation between $x =$ age and $y =$ whether a member of Facebook ($1 =$ yes, $0 =$ no). A priori, you predict that $P(Y = 1)$ is currently about 0.80 at $x = 18$ and about 0.20 at $x = 65$. Assuming that the logistic regression model describes this relation well, approximate the value for the effect β of x in the model.

4.23 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 use logistic regression to analyze 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 model-based descriptive and inferential results.

4.24 Refer to the following artificial data:

x	number of trials	number of successes
0	4	1
1	4	2
2	4	4

Denote by M_0 the logistic null model and by M_1 the model that also has x as a predictor. Denote the maximized log-likelihood values by L_0 for M_0, L_1 for M_1, and L_s for the saturated model. Create a data file in two ways, entering the data as (i) ungrouped data: 12 individual binary observations, (ii) grouped data: 3 summary binomial observations each with sample size $= 4$.

a. Fit M_0 and M_1 for each data file. Report L_0 and L_1 (or $-2L_0$ and $-2L_1$) in each case. Do they depend on the form of data entry?

b. Show that the deviances for M_0 and M_1 depend on the form of data entry. Why is this? (*Hint*: The saturated model has 12 parameters for data file (i) but 3 parameters for data file (ii).)

c. Show that the difference between the deviances does not depend on the form of data file. Thus, for testing the effect of x, it does not matter how you enter the data.