Table 3.3 Generalized linear models for statistical analysis.

Random Component	Link Function	Explanatory Variables	Model	Chapter
Normal	Identity	Continuous	Regression	
Normal	Identity	Categorical	Analysis of variance	
Normal	Identity	Mixed	Analysis of covariance	
Binomial	Logit	Mixed	Logistic regression	4–5, 8–10
Multinomial	Logits	Mixed	Multinomial logit	6, 8–10
Poisson	Log	Mixed	Loglinear	7

EXERCISES

3.1 Describe the purpose of the link function of a GLM. Define the identity link and explain why it is not often used with a binomial parameter.

3.2 In the years 1904, 1914, 1924, . . . , 2014, the percentage of times the starting pitcher pitched a complete game were[7]: 87.6, 55.0, 48.7, 43.4, 45.2, 34.0, 24.5, 28.0, 15.0, 8.0, 3.1, 2.4.

a. The linear probability model has least squares fit $\hat{P}(Y = 1) = 0.6930 - 0.0662x$, where $x =$ number of decades since 1904. Interpret -0.0662.

b. Substituting $x = 12$ in the linear prediction equation, predict the proportion of complete games for 2024. The ML fit of the logistic regression model yields $\hat{P}(Y = 1) = 0.034$ at $x = 12$. Which prediction is more plausible? Why?

3.3 For Table 2.6 on $x =$ mother's alcohol consumption and $Y =$ whether a baby has sex organ malformation, ML fitting of the linear probability model with x scores 0, 0.5, 1.5, 4.0, 7.0 has output:

```
------------------------------------------
Parameter   Estimate  Std Error
Intercept   0.00255    0.0003
alcohol     0.00109    0.0007
------------------------------------------
```

a. State the prediction equation and interpret the intercept and slope.

b. Use the model fit to estimate the (i) probabilities of malformation for alcohol levels 0 and 7.0, (ii) relative risk comparing those levels.

c. Is the result sensitive to the choice of scores? Re-fit the linear probability model using scores 0, 1, 2, 3, 4, and re-evaluate fitted probabilities at alcohol levels 0 and 7 and the relative risk.

d. The sample proportion of malformations is much higher in the highest alcohol category than the others because, although it has only one malformation, its sample size is only 38. Are results sensitive to this single malformation? Fit the logistic regression or linear probability model with and without that observation, and evaluate fitted probabilities at alcohol levels 0 and 7 and the relative risk.

[7] *Source*: `https://en.wikipedia.org/wiki/Complete_game`.

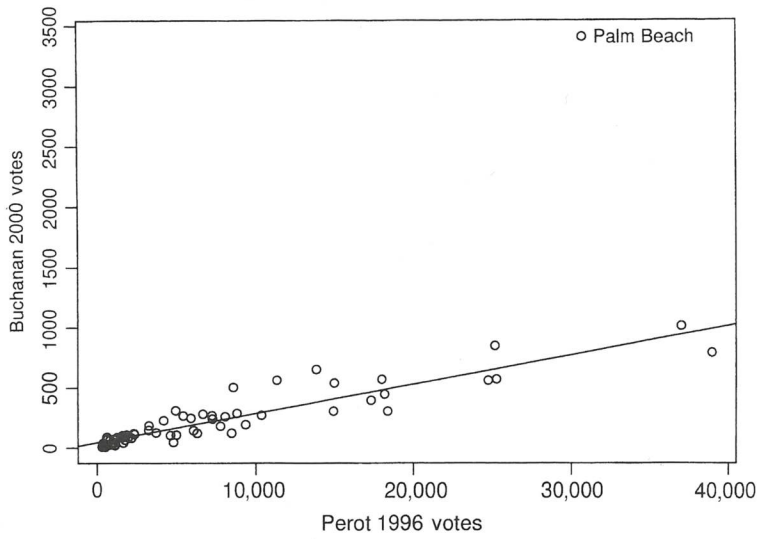

Figure 3.6 Total vote, by county in Florida, for Reform Party candidates Buchanan in 2000 and Perot in 1996.

3.4 In the 2000 U.S. Presidential election, Palm Beach County in Florida was the focus of unusual voting patterns apparently caused by a confusing "butterfly ballot." Many voters claimed they voted mistakenly for the Reform Party candidate, Pat Buchanan, when they intended to vote for Al Gore. Figure 3.6 shows the total number of votes for Buchanan plotted against the number of votes for the Reform Party candidate in 1996 (Ross Perot), by county in Florida.[8]

a. In county i, let π_i denote the proportion of the vote for Buchanan and let x_i denote the proportion of the vote for Perot in 1996. For the linear probability model fitted to all counties except Palm Beach County, $\hat{\pi}_i = -0.0003 + 0.0304x_i$. Give the value of P in the interpretation. The estimated proportion vote for Buchanan in 2000 was roughly $P\%$ of that for Perot in 1996.

b. For Palm Beach County, $\pi_i = 0.0079$ and $x_i = 0.0774$. Does this result appear to be an outlier for the model? Investigate, by finding $\pi_i/\hat{\pi}_i$ and $\pi_i - \hat{\pi}_i$. (Statistical analyses predicted that fewer than 900 votes were truly intended for Buchanan, compared to the 3407 he received. George W. Bush won the state by 537 votes and, with it, the Electoral College and the election. Other ballot design problems played a role in 110,000 disqualified overvote ballots, in which people mistakenly voted for more than one candidate, with Gore marked on 84,197 ballots and Bush on 37,731.)

3.5 Access the horseshoe crab data file (shown partly in Table 3.2) at `www.stat.ufl.edu/~aa/cat/data`. In the data file, $y = 1$ if a crab has at least one satellite and $y = 0$ otherwise.

a. Using weight as the predictor, fit the linear probability model to $P(Y = 1)$. If your software cannot use the identity link with the binomial or fails to converge, use

[8] For details, see A. Agresti and B. Presnell, *Statistical Science* **17**: 436–440 (2002).

ordinary least squares by treating Y as normally distributed. Interpret the parameter estimates. Find $\hat{P}(Y = 1)$ at the highest observed weight of 5.20 kg. Comment.

b. Fit the logistic regression model. Show that at a weight of 5.20 kg, $\hat{P}(Y = 1) = 0.9968$.

3.6 From the 2016 General Social Survey, when we cross-classify political ideology (with 1 being most liberal and 7 being most conservative) by political party affiliation for subjects of ages 18–27, we get:

	1	2	3	4	5	6	7
Democrat	5	18	19	25	7	7	2
Republican	1	3	1	11	10	11	1

When we use R to model the effect of political ideology on the probability of being a Democrat, we get the results:

```
> y <- c(5,18,19,25,7,7,2); n <- c(6,21,20,36,17,18,3)
> x <- c(1,2,3,4,5,6,7)
> fit <- glm(y/n ~ x, family=binomial(link=logit), weights=n)
> summary(fit)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  3.1870      0.7002    4.552  5.33e-06
x           -0.5901      0.1564   -3.772  0.000162
---
    Null deviance: 24.7983  on 6  degrees of freedom
Residual deviance:  7.7894  on 5  degrees of freedom
Number of Fisher Scoring iterations: 4
> confint(fit)
                2.5 %    97.5 %
(Intercept)   1.90180   4.66484
x            -0.91587  -0.29832
```

a. Report the prediction equation and interpret the direction of the estimated effect.

b. Construct the 95% Wald confidence interval for the effect of political ideology. Interpret and compare to the profile likelihood interval shown.

c. Conduct the Wald test for the effect of x. Report the test statistic, P-value, and interpret.

d. Conduct the likelihood-ratio test for the effect of x. Report the test statistic, find the P-value, and interpret.

e. Explain the output about the number of Fisher scoring iterations.

3.7 Consider Table 3.1 on snoring and heart disease.

a. Re-fit the logistic regression model using the scores (i) (0, 2, 4, 6), (ii) (0, 1, 2, 3), (iii). (1, 2, 3, 4). Compare the model parameter estimates under the three choices.

Compare the fitted values. What can you conclude about the effect of *linear* transformations of scores that preserve relative sizes of spacings between scores?

b. Fit the logistic regression model using the scores (0, 2, 6, 7), approximating the number of days in a week that the subject snores. Compare fitted values to those with the scores (0, 2, 4, 5) used in the text example. Do results seem to be sensitive to the choice of scores?

3.8 Fit the logistic regression model of Section 3.2.3 to Table 3.1 on snoring and heart disease. Show results of significance tests and confidence intervals for the effect of snoring.

3.9 Table 3.4, the `Credit` data file at the text website, shows data for a sample of 100 adults randomly selected for an Italian study on the relation between x = annual income and y = whether you have a travel credit card (1 = yes, 0 = no). At each level of x (in thousands of euros), the table indicates the number of subjects in the sample and the number of those having at least one travel credit card. Software provides the following results of using logistic regression:

```
------------------------------------------------
             Estimate  Std. Error
(Intercept)  -3.5179      0.7103
x             0.1054      0.0262
------------------------------------------------
```

a. Report the prediction equation and interpret the sign of $\hat{\beta}$.

b. When $\hat{P}(Y = 1) = 0.50$, show that the estimated logit value is 0. Based on this, for these data explain why the estimated probability of a travel credit card is 0.50 at income = 33.4 thousand euros.

c. Show how to apply software to the `Credit` data file at the text website to obtain the logistic regression fit.

Table 3.4 Data on travel credit cards and income for exercise 3.9.

Income	No. of Cases	Credit Cards	Income	No. of Cases	Credit Cards	Income	No. of Cases	Credit Cards
12	1	0	21	2	0	34	3	3
13	1	0	22	1	1	35	5	3
14	8	2	24	2	0	39	1	0
15	14	2	25	10	2	40	1	0
16	9	0	26	1	0	42	1	0
17	8	2	29	1	0	47	1	0
19	5	1	30	5	2	60	6	6
20	7	0	32	6	6	65	1	1

Source: Thanks to R. Piccarreta, Bocconi University, Milan, for original form of data.

3.10 A recent General Social Survey asked "How many people at your work place are close friends?" The 756 responses had a mean of 2.76, standard deviation of 3.65, and a mode of 0. Would the Poisson distribution describe these data well? Why or why not?

3.11 An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the numbers of imperfections are 8, 7, 6, 6, 3, 4, 7, 2, 3, 4. Treatment B applied to 10 other wafers has 9, 9, 8, 14, 8, 13, 11, 5, 7, 6 imperfections. Treat the counts as independent Poisson variates having means μ_A and μ_B. Consider the model $\log \mu = \alpha + \beta x$, where $x = 1$ for treatment B and $x = 0$ for treatment A, for which $\beta = \log \mu_B - \log \mu_A = \log(\mu_B/\mu_A)$ and $e^\beta = \mu_B/\mu_A$. Fit the model. Report the prediction equation and interpret $\hat{\beta}$.

3.12 Refer to the previous exercise.

a. Test H_0: $\mu_A = \mu_B$ by conducting the Wald or likelihood-ratio test of H_0: $\beta = 0$. Interpret.

b. Construct a 95% confidence interval for μ_B/μ_A. (*Hint*: Construct one for $\beta = \log(\mu_B/\mu_A)$ and then exponentiate.)

3.13 For the `Crabs` data file (partially shown in Table 3.2) at `www.stat.ufl.edu/~aa/cat/data`, fit the Poisson loglinear model to use weight to predict the number of satellites.

a. Report the prediction equation, and estimate the mean response for female crabs of average weight, 2.44 kg.

b. Use $\hat{\beta}$ to describe the weight effect. Construct a 95% confidence interval for β and for the multiplicative effect of a 1-kg increase.

c. Conduct Wald and likelihood-ratio tests of the hypothesis that the mean response is independent of weight. Interpret.

3.14 If you are modeling count data, explain why it is not sufficient to analyze ordinary raw residuals, $(y_i - \hat{\mu}_i)$, as you would for ordinary linear models.

3.15 True or false?

a. An ordinary regression model that treats Y as normally distributed is a special case of a GLM, with a normal random component and identity link function.

b. With a GLM, Y does not need to have a normal distribution and one can model a function of the mean of Y instead of just the mean itself, but to get ML estimates the variance of Y must be constant at all values of explanatory variables.