Model	Variance
Poisson	μ
Poisson (quasi)	$\psi\mu$
Negative binomial	$\mu(1 + \phi\mu)$

Because the negative binomial distribution is thought to be a natural model for extra-Poisson variation and because maximum likelihood can be used for regression inference based on it, negative binomial log-linear regression has always been an attractive alternative to Poisson log-linear regression. Until recently, however, easy-to-use computer programs for its use have not been available. Fortunately, statistical computer programs have now incorporated modules for negative binomial regression and these can be used in the same way as the GLM routines for Poisson log-linear regression. The strategies for data analysis remain the same, except there is no need to investigate extra-Poisson variation because it is automatically included in the model. Tests and confidence intervals are based on Wald and drop-in-deviance (also known as likelihood ratio) theory, with the same issues as for Poisson. There is no guarantee that the negative binomial model is appropriate for all nonnegative integer response variables, but it offers greater flexibility than Poisson and has proven itself useful in a wide variety of applications.

22.7 SUMMARY

Poisson regression should be considered whenever the response variable is a count of events that occur over some observation period or spatial area. The main feature of the Poisson distribution is that the variance is equal to the mean. Consequently, if ordinary regression is used, the residual plot will exhibit a funnel shape. In the Poisson log-linear regression model, the distribution of the response variables for each combination of the explanatory variables is thought to be Poisson, and the logarithms of the means of these distributions are modeled by formulas that are linear functions of regression coefficients. The method of maximum likelihood is used to estimate the unknown regression coefficients.

Elephant Mating Data

A scatterplot of the number of successful matings versus age for the 41 male elephants indicates the inadequacy of a straight-line regression and of a constant variance assumption for these data. Linear regression after a log transformation of number of matings is possible, but the zeros in the data set are a nuisance in this approach. Poisson log-linear regression provides an attractive alternative. The counts are too small to justify the deviance goodness-of-fit test, but plots and informal testing of extra terms imply that the model with the log of the mean count as a straight-line function of age is adequate. One question of interest is whether the mean count of successful matings peaks at some age. This can be investigated by testing the significance of an age-squared term. The drop-in-deviance test does not provide any evidence of a quadratic effect.

Salamander Data

Residuals from the fit of a rich model were larger than expected, strongly suggesting extra-Poisson variability. Therefore, quasi-likelihood methods were used in model selection and in reporting results. There were two distinct canopy cover regimes. A quasi-likelihood F-test indicated that forest age was not associated with mean counts, after taking into account canopy cover. The fit of the reduced model included a highly significant difference between quadratic effects of canopy, so the descriptive model was selected as having separate quadratics in low cover and high cover situations.

22.8 EXERCISES

Conceptual Exercises

1. Elephant Mating. Both the binomial and the Poisson distributions provide probability models for random counts. Is the binomial distribution appropriate for the number of successful matings of the male African elephants?

2. What is the difference between a log-linear model and a linear model after the log transformation of the response?

3. If a confidence interval is obtained for the coefficient of an indicator variable in a log-linear regression model, do the antilogarithms of the endpoints describe a confidence interval for a ratio of means or for a ratio of medians?

4. Elephant Mating. In Display 22.2 the spread of the responses is larger for larger values of the mean response. Is this something to be concerned about if Poisson log-linear regression is used?

5. Elephant Mating. From the estimated log-linear regression of elephants' successful matings on age (Display 22.8), what are the mean and the variance of the distribution of counts of successful matings (in 8 years) for elephants who are aged 25 years at the beginning of the observation period? What are the mean and the variance for elephants who are aged 45 years?

6. (a) Why are ordinary residuals—$(Y_i - \hat{\mu}_i)$—not particularly useful for Poisson regression? (b) How are the Pearson residuals designed to deal with this problem?

7. Consider the deviance goodness-of-fit test. (a) Under what conditions is it valid for Poisson regression? (b) When it is valid, what possibilities are suggested by a small p-value? (c) When it is valid, what possibilities are suggested by a large p-value?

8. (a) Why is it more difficult to check the adequacy of a Poisson log-linear regression model when the counts are small than when they are large? (b) What tools are available in this situation?

9. (a) How does the drop-in-deviance test for Poisson log-linear regression resemble the extra-sum-of-squares test in ordinary regression? (b) How does it differ?

10. (a) How does the drop-in-deviance test for the quasi-likelihood version of the log-linear regression model resemble the extra-sum-of-squares test in ordinary regression? (b) How does it differ?

11. How does the quasi-likelihood version of the log-linear regression model allow for more variation than would be expected if the responses were Poisson?

12. If responses follow the Poisson log-linear regression model, the Pearson residuals should have variance approximately equal to 1. If, instead, the quasi-likelihood model with dispersion parameter ψ is appropriate, what is the approximate variance of the Pearson residuals?

13. Is it acceptable to use the quasi-likelihood model when the data actually follow the Poisson model?

14. Consider a table that categorizes 1,000 subjects into 5 rows and 10 columns. (a) If Poisson log-linear regression is used to analyze the data, what is the sample size? (That is, how many Poisson counts are there?) (b) How would one test for independence of row and column factors?

Computational Exercises

15. Elephant Mating and Age. Give an estimated model for describing the number of successful matings as a function of age, using (a) simple linear regression after transforming the number of successful matings to the square root; (b) simple linear regression after a logarithmic transformation (after adding 1); (c) log-linear regression. (d) Do the models used in parts (a) or (b) exhibit obvious inadequacies?

16. Murder–Suicides by Deliberate Plane Crash. Some sociologists suspect that highly publicized suicides may trigger additional suicides. In one investigation of this hypothesis, D. P. Phillips collected information about 17 airplane crashes that were known (because of notes left behind) to be murder–suicides. For each of these crashes, Phillips reported an index of the news coverage (circulation in nine newspapers devoting space to the crash multiplied by length of coverage) and the number of multiple-fatality plane crashes during the week following the publicized crash. The data are exhibited in Display 22.12. (Data from D. P. Phillips, "Airplane Accident Fatalities Increase Just After Newspaper Stories About Murder and Suicide," *Science* 201 (1978): 748–50.) Is there evidence that the mean number of crashes increases with increasing levels of publicity of a murder–suicide?

DISPLAY 22.12 Multiple-fatality plane crashes in the week following a murder–suicide by plane crash, and the amount of newspaper coverage given the murder–suicide

Index of coverage	Number of crashes	Index of coverage	Number of crashes
376	8	63	2
347	5	44	7
322	8	40	4
104	4	5	3
103	6	5	2
98	4	0	4
96	8	0	3
85	6	0	2
82	4		

17. Obesity and Heart Disease. Analyze the table of counts in Display 22.11 as suggested there. What is the p-value for testing independence of obesity and CVD death outcome?

18. Galapagos Islands. Reanalyze the data in Exercise 12.20 with number of native species as the response, but using log-linear regression. (a) Fit the model with log area, log elevation, log of distance from nearest island, and log area of nearest island as explanatory variables; and then check for extra-Poisson variation. (b) Use backward elimination to eliminate insignificant explanatory variables. (c) Describe the effects of the remaining explanatory variables.

19. Galapagos Islands. Repeat the previous exercise, but use the number of nonnative species as the response variable (total number of species minus the number of native species).

20. Cancer Deaths of Atomic Bomb Survivors. The data in Display 22.13 are the number of cancer deaths among survivors of the atomic bombs dropped on Japan during World War II, categorized by time (years) after the bomb that death occurred and the amount of radiation exposure that the survivors received from the blast. (Data from D. A. Pierce, personal communication.) Also listed in each cell is the *person-years at risk*, in 100's. This is the sum total of all years spent by all persons in the category. Suppose that the mean number of cancer deaths in each cell is Poisson with mean $\mu = risk \times rate$, where *risk* is the person-years at risk and *rate* is the rate of cancer deaths per person per year. It is desired to describe this rate in terms of the amount of radiation, adjusting for the effects of time after exposure. (a) Using log(*risk*) as an offset, fit the Poisson log-linear regression model with time after blast treated as a factor (with seven levels) and with *rads* and *rads*-squared treated as covariates. Look at the deviance statistic and the deviance residuals. Does extra-Poisson variation seem to be present? Is the *rads*-squared term necessary? (b) Try the same model as in part (a); but instead of treating time after bomb as a factor with seven levels, compute the midpoint of each interval and include log(*time*) as a numerical explanatory variable. Is the deviance statistic substantially larger in this model, or does it appear that time can adequately be represented through this single term? (c) Try fitting a model that includes the interaction of log(*time*) and exposure. Is the interaction significant? (d) Based on a good-fitting model, make a statement about the effect of radiation exposure on the number of cancer deaths per person per year (and include a confidence interval if you supply an estimate of a parameter).

DISPLAY 22.13 Cancer deaths among Japanese atomic bomb survivors, categorized by estimated exposure to radiation (in rads) and years after exposure; below the number of cancer deaths are the person-years (in 100's) at risk

		Years after exposure						
exposure (rads)		0–7	8–11	12–15	16–19	20–23	24–27	28–31
0	*deaths:*	10	12	19	31	35	48	73
	risk:	262	243	240	237	233	227	220
25	*deaths:*	17	17	17	47	50	65	71
	risk:	313	290	285	280	275	269	262
75	*deaths:*	0	2	1	5	8	7	12
	risk:	38	36	35	34	34	33	32
150	*deaths:*	1	0	4	1	6	12	11
	risk:	28	26	25	25	24	24	23
250	*deaths:*	1	1	0	4	3	7	13
	risk:	13	12	12	12	11	11	10
400	*deaths:*	0	2	5	3	2	3	5
	risk:	15	14	14	14	13	13	13

21. El Niño and Hurricanes. Reconsider the El Niño and Hurricane data set from Exercise 10.28. Use poisson log-linear regression to describe the distribution of (a) number of storms and (b) number of hurricanes as a function of El Niño temperature and West African wetness.

Data Problems

22. Emulating Jane Austen's Writing Style. When she died in 1817, the English novelist Jane Austen had not yet finished the novel *Sanditon*, but she did leave notes on how she intended to conclude the book. The novel was completed by a ghost writer, who attempted to emulate Austen's style. In 1978, a researcher reported counts of some words found in chapters of books written by Austen and in chapters written by the emulator. These are reproduced in Display 22.14. (Data from A. Q. Morton, *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*, New York: Charles Scribner's Sons, 1978.) Was Jane Austen consistent in the three books in her relative uses of these words? Did the emulator do a good job in terms of matching the relative rates of occurrence of these six words? In particular, did the emulator match the relative rates that Austen used the words in the first part of *Sanditon*?

DISPLAY 22.14 Occurrences of six words in various chapters of books written by Jane Austen (*Sense and Sensibility*, *Emma*, and the first part of *Sanditon*) and in some chapters written by the writer who completed *Sanditon* (Sanditon II)

Word	Book			
	Sense and Sensibility	*Emma*	*Sanditon* I	*Sanditon* II
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4

23. Space Shuttle O-Ring Failures. On January 27, 1986, the night before the space shuttle *Challenger* exploded, an engineer recommended to the National Aeronautics and Space Administration (NASA) that the shuttle not be launched in the cold weather. The forecasted temperature for the *Challenger* launch was 31°F—the coldest launch ever. After an intense 3-hour telephone conference, officials decided to proceed with the launch. Shown in Display 22.15 are the launch temperatures and the number of O-ring problems in 24 shuttle launches prior to the *Challenger* (Chapter 4). Do these data offer evidence that the number of incidents increases with decreasing temperature?

DISPLAY 22.15 Launch temperatures (°F) and numbers of O-ring incidents in 24 space shuttle flights

Launch temperature (°F)	Number of incidents	Launch temperature (°F)	Number of incidents
53	3	70	1
56	1	70	1
57	1	72	0
63	0	73	0
66	0	75	0
67	0	75	2
67	0	76	0
67	0	76	0
68	0	78	0
69	0	79	0
70	0	80	0
70	1	81	0

24. Valve Failure in Nuclear Reactors. Display 22.16 shows characteristics and numbers of *failures* observed in valve types from one pressurized water reactor. There are five explanatory factors: *system* (1 = containment, 2 = nuclear, 3 = power conversion, 4 = safety, 5 = process auxiliary); *operator* type (1 = air, 2 = solenoid, 3 = motor-driven, 4 = manual); *valve* type (1 = ball, 2 = butterfly, 3 = diaphragm, 4 = gate, 5 = globe, 6 = directional control); head *size* (1 = less than 2 inches, 2 = 2–10 inches, 3 = 10–30 inches); and operation *mode* (1 = normally closed, 2 = normally open). The lengths of observation periods are quite different, as indicated in the last column, *time*. Using an offset for log of observation time, identify the factors associated with large numbers of valve failures. (Data from L. M. Moore and R. J. Beckman, "Appropriate One-Sided Tolerance Bounds on the Number of Failures Using Poisson Regression," *Technometrics* 30 (1988): 283–90.)

DISPLAY 22.16 Valve characteristics and numbers of failures in a nuclear reactor; first 6 of 91 rows

System	*Operator*	*Valve*	*Size*	*Mode*	*Failures*	*Time*
1	3	4	3	1	2	4
1	3	4	3	2	2	4
1	3	5	1	1	1	2
2	1	2	2	2	0	2
2	1	3	2	1	0	2
2	1	3	2	2	0	1

25. Body Size and Reproductive Success in a Population of Male Bullfrogs. As an example of field observation in evidence of theories of sexual selection, S. J. Arnold and M. J. Wade ("On the Measurement of Natural and Sexual Selection: Applications," *Evolution* 38 (1984): 720–34) presented the following data set on size and number of mates observed in 38 male bullfrogs (Display 22.17). Is there evidence that the distribution of number of mates in this population is related to body size? If so, supply a quantitative description of that relationship, along with an appropriate measure of uncertainty. Write a brief summary of statistical findings.

DISPLAY 22.17 Body size (mm) and number of mates for 38 male bullfrogs (*Rana catesbeiana*); first 5 of 38 rows

Body size	Mates
144	3
150	2
144	1
154	1
132	1

26. Number of Moons. Display 22.18 shows some characteristics of terrestrial planets, gas giants, and dwarf planets in our solar system, including the known number of moons. Apparently, larger planets have more moons, but is it the volume (as indicated by diameter) or mass that are more relevant, or is it both? Answer this question and arrive at a model for describing the mean number of moons as a function of planet size. Use negative binomial regression if you have access to a computer package that does the computations. Otherwise, use quasi-likelihood analysis. (Data from Wikipedia: http://en.wikipedia.org/ wiki/Planet#Solar_System, August 10, 2011.)

Answers to Conceptual Exercises

1. No. A binomial distribution is for a count that has a precisely-defined upper limit, such as the number of heads in 10 flips of a coin (with upper limit 10) or the number of elephant legs with defects (with upper limit 4). Although there is surely some practical limit to the number of successful elephant matings, there is no precisely-defined maximum to be used as the binomial index. The Poisson distribution is more sensible.

2. In a log-linear model, the mean of Y is μ and the model is $\log(\mu) = \beta_0 + \beta_1 X_1$. Y is not transformed. If simple linear regression is used after a log transformation, the model is expressed in terms of the mean of the logarithm of Y.

3. Ratio of means. The model states that the log of the mean—not the median—has the regression form. It is not necessary to introduce the median of the Poisson distribution here.

4. No. The nonconstant variance is anticipated by the Poisson model. The maximum likelihood procedure correctly uses the information in the data to estimate the regression coefficients.

DISPLAY 22.18 Average distance from the sun, diameter, and mass (all scaled so that the values for earth are 1); and number of moons of 13 planets and dwarf planets in our solar system

Name	Distance	Diameter	Mass	Moons
Mercury	0.39	0.382	0.06	0
Venus	0.72	0.949	0.82	0
Earth	1	1	1	1
Mars	1.52	0.532	0.11	2
Ceres	2.75	0.08	0.0002	0
Jupiter	5.2	11.209	317.8	64
Saturn	9.54	9.449	95.2	62
Uranus	19.22	4.007	14.6	27
Neptune	30.06	3.883	17.2	13
Pluto	39.5	0.18	0.0022	4
Haumea	43.35	0.15	0.0007	2
Makemake	45.8	0.12	0.0007	0
Eris	67.7	0.19	0.0025	1

5. For 25-year-old elephants: Mean = 1.15; Variance = 1.15. For 45-year-old elephants: Mean = 4.53; Variance = 4.53.

6. (a) The residuals with larger means will have larger variances. Thus, if an observation has a large residual it is difficult to know whether it is an outlier or an observation from a distribution with larger variance than the others. (b) The residuals are scaled to have the same variance.

7. (a) Large Poisson means. (b) The Poisson distribution is an inadequate model, the regression model terms are inadequate, or there are a few contaminating observations. (c) Either the model is correct, or there is insufficient data to detect any inadequacies.

8. (a) Scatterplots are uninformative, the deviance goodness-of-fit test cannot be used, and the approximate normality of residuals is not guaranteed. (b) One may try to add extra terms, such as a squared term in some explanatory variable, to model a simple departure from linearity and to test its significance with a drop-in-deviance or Wald's test.

9. (a) It is based on a comparison of the magnitudes of residuals from a full to a reduced model. (b) The residuals used are deviance residuals, and a chi-squared statistic is used rather than an F-statistic.

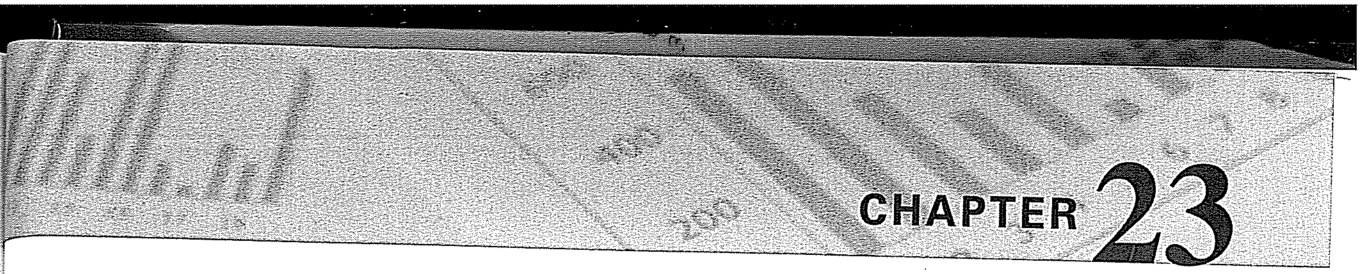
10. (a) Same as 10(a). In this case, an F-statistic is formed just as with the usual extra-sum-of-squares test. (b) It is based on deviance residuals.

11. The variance of the responses is $\psi\mu$, where ψ is unknown. A value of ψ greater than 1 allows for more variation than anticipated by a Poisson distribution.

12. ψ.

13. Yes; the Poisson mean–variance relationship is a special case of the quasi-likelihood model (when ψ is 1). (More powerful comparisons result, however, if the Poisson model is used when it definitely applies.)

14. (a) 50. (b) Test for the significance of the 36 interaction terms in the log-linear regression with row, column, and row-by-column effects. This may be accomplished by fitting the model without interaction and comparing the Pearson statistic to a chi-squared distribution on 36 degrees of freedom.



CHAPTER 23

Elements of Research Design

Throughout this book, case studies have highlighted fruitful designs. It is now time to take a closer look at study design, with the objective of setting down the basic principles and illustrating the steps leading to a well-designed study. The discussion of research design takes up this and the following chapter. This chapter lays out the general principles. It provides practical suggestions on the steps to follow, including the important task of determining sample size. The second design chapter (Chapter 24) presents some standard statistical design structures that allow straightforward analysis and interpretation of results.