TI 2002-119/4
Tinbergen Institute Discussion Paper

The Origins of Logistic Regression

J.S. Cramer

Faculty of Economics and Econometrics, University of Amsterdam, and Tinbergen Institute.

The Origins of Logistic Regression

J.S. Cramer *

November 2002

Abstract

This paper describes the origins of the logistic function, its adoption in bio-assay, and its wider acceptance in statistics. Its roots spread far back to the early 19th century; the survival of the term *logistic* and the wide application of the device have been determined decisively by the personal histories and individual actions of a few scholars.

This is a much extended version of Chapter 9 of
Logit Models from Economics and Other Fields,
forthcoming at Cambridge University Press

*University of Amsterdam and Tinbergen Institute, Amsterdam; postal adress Baambrugse Zuwe 194, 3645 AM Vinkeveen, the Netherlands; e-mail address mars.cram@worldonline.nl. I had the benefit of comments of J. Sandee on an earlier version.

1 Introduction

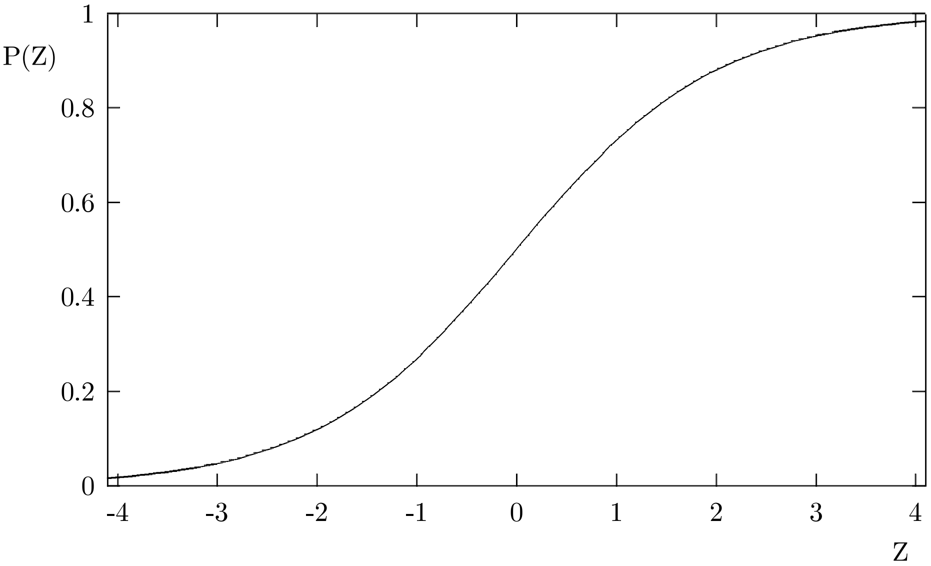


Figure 1. The logistic curve P(Z)

The sigmoid curve of Figure 1 is traced by the logistic function

$$P(Z) = \frac{expZ}{1 + expZ}. \tag{1}$$

P behaves like the distribution function of a symmetrical density, with midpoint zero; as Z moves through the real number axis, P rises monotonically between the bounds of zero and 1. The meaning of this function varies according to the the definition of the variables. In the logit version of bio-assay P is the probability of a binary outcome, and $Z = \alpha + \beta X$, with X a stimulus or exposure variable; α determines the location of the curve on the X-axis, and β its slope. In logistic regression there are several determinants of P, and $Z = x^T\beta$, with x a vector of covariates (including a unit constant) and β their coefficients. But the logistic function originally describes the course of a *proportion* P over *time* t, with $Z = \alpha + \beta t$; since $P(t)$ rises monotonically with t) this is a *growth curve.*

Over a fairly wide central range, for values of P from .3 to .7, the logistic curve closely resembles in shape as the normal probability distribution function. The two functions

$$P_l(x) = \frac{exp(\beta x)}{1 + exp(\beta x)}. \tag{2}$$

and

$$P_n(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\inf}^{x} \exp\{-1/2(u/\sigma)^2\} du. \tag{3}$$

both pass through the point $(0, .5)$, and they can be made almost to coincide upon a suitable adjustment of β and σ. This is a sheer algebraic coincidence, for there appears to be no intrinsic relation between the two forms.

2 The origins of the logistic function

The logistic function was invented in the 19th century for the description of the growth of populations and the course of autocatalytic chemical reactions. In either case we consider the time path of a quantity $W(t)$ and its growth rate

$$\dot{W}(t) = \mathrm{d}W(t)/\mathrm{d}t. \tag{4}$$

The simplest assumption is that $\dot{W}(t)$ is proportional to $W(t)$

$$\dot{W}(t) = \beta W(t),\ \beta = \dot{W}(t)/W(t), \tag{5}$$

with β the constant rate of growth. This leads of course to exponential growth

$$W(t) = A \exp \beta t,$$

where A is sometimes replaced by the initial value $W(0)$. This is a reasonable model for unopposed population growth in a young country like the United States in its early years; as Malthus (1789) put it, a human population, left to itself, will increase *in geometric progression.*[1] But Alphonse Quetelet (1795–1874), the Belgian astronomer turned statistician, was well aware that the indiscriminate extrapolation of exponential growth must lead to impossible values. He experimented with several adjustments of (5) and also asked his pupil Pierre-François Verhulst (1804–1849) to look into the problem.

Like Quetelet, Verhulst approached the problem by adding an extra term to (5) to represent the increasing resistance to further growth, as in

$$\dot{W}(t) = \beta W(t) - \phi(W(t)). \tag{6}$$

[1] Two hundred years later exponential growth played a major part in the *Report to the Club of Rome* of Meadows et al (1972), and it still lies at the basis of many economic analyses.

and then experimenting with various forms of ϕ. The logistic appears when this is a quadratic, for then we may rewrite (6) as

$$\dot{W}(t) = \beta W(t)(\Omega - W(t)) \tag{7}$$

where Ω denotes the upper limit or *saturation level* of W. Growth is now proportional both to the population already attained $W(t)$ and to the remaining room for further expansion $\Omega - W(t)$. If we express $W(t)$ as a proportion $P(t) = W(t)/\Omega$ this gives

$$P(t) = \beta P(t)\{1 - P(t)\}, \tag{8}$$

and the solution of this differential equation is

$$P(t) = \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)}, \tag{9}$$

which Verhulst named the *logistic* function. The population $W(t)$ then follows

$$W(t) = \Omega \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)}. \tag{10}$$

Verhulst published his suggestions between 1838 and 1847 in three papers. The first is a brief note in the *Correspondance Mathématique et Physique* edited by Quetelet in 1838. It contains the essence of the argument in four small pages, followed by a demonstration that the curve agrees very well with the actual course of the population of France, Belgium, Essex and Russia for periods up to 1833; Verhulst explains that he did his research a couple of years before, that he did not have the time for an update and that he publishes these notes only at the insistence of Quetelet. He does not say how he fitted the curves. The second paper, in the *Proceedings* of the Belgian Royal Academy of 1845, is a much fuller account of the function and its properties. Here Verhulst names it the logistic, without further explanation: in a neat diagram, the *courbe logistique* is drawn alongside the *courbe logarithmique*, which we would nowadays call the exponential. Verhulst also determines the three parameters Ω, α and β of (10) by making the curve pass through three observed points. With data for some twenty or thirty years only this is a hazardous method, as is borne out by the resulting estimates of the limiting population Ω of 6.6 millions for Belgium and 40 million for France: at present these populations number 10.2 and 58.7 million. In 1847 there followed a second paper in the *Proceedings*, which is chiefly notable for an adjustment of the correction term that leads to a much better estimate of 9.5 millions for the belgian Ω.

Verhulst was primarily a mathematician, but sensitive to social and political issues; he was in poor health and died young. Quetelet attributes his early death to overwork and, rather curiously, to his great stature, as Verhulst was 1.89 meters or six feet tall. His impulsive nature was demonstrated in the summer of 1830. Verhulst had gone to Italy for his health and was staying in Rome when the news of the revolution in Paris and of the Belgian secession from the Netherlands broke. These events moved him strongly and set him drafting a democratic constitution for the Papal State. He submitted this document to some cardinals he had met, who expressed great interest; still the police were called in, and Verhulst banished from Rome. He left under somewhat dramatic circumstances, having at first barricaded his apartment with the intention of withstanding a siege by the forces of law and order.

The logistic function was discovered anew in 1920 by Pearl and Reed in a study of the population growth of the United States. They were unaware of Verhulst's work (though not of the curves for autocatalytic reactions discussed presently), and they arrived independently at the logistic curve of (10). When this was fitted to Census figures, again by making the curve pass through three points, it gave a good fit for the period from 1790 to 1910. But the estimate of Ω of 197 millions again compares badly with the present value of about 270 millions. In spite of many other interests, Pearl and his collaborators in the next twenty years went on to apply the logistic growth curve to almost any living population from fruit flies to the human population of the French colonies in North Africa as well as to the growth of cantaloupes; we list a few of these studies in the bibliography.

In 1920, Raymond Pearl (1879–1940) had just been appointed Director of the Department of Biometry and Vital Statistics at Johns Hopkins University, and Lowell J. Reed (1886–1966) was his deputy (and his successor when a few years later Pearl was promoted to Professor of Biology). Pearl was trained as a biologist and acquired his statistics as a young man in 1905–1906 by spending a year in London with Karl Pearson (and later quarrelling with him). He became a prodigious investigator and a prolific writer on a wide variety of phenomena like longevity, fertility, contraception, and the effects of alcohol and tobacco consumption on health, all subsumed under the heading of human biology. During World War I Pearl worked in the U.S. Food Administration, and this may account for his preoccupation with the food needs of a growing population in the 1920 paper. Reed, who was trained as a mathematician, made a quiet career in biostatistics; he excelled as a teacher and as an administrator, and was brought back in 1953 from retirement to serve as President of Johns Hopkins. Among his publications in the aftermath of the 1920 paper with Pearl is an application of the logistic curve to autocatalytic reactions, Reed and Berkson (1929). We shall hear more about this co-author in the next section.

The term *logistic* was of course not used until Verhulst's work was rediscovered, which was soon after Pearl and Reed's first paper of 1920. The immediate sequel, Pearl and Reed (1922), does not mention Verhulst; his priority is first acknowledged in a footnote in Pearl (1922). In Pearl and Reed (1923) Verhulst is again named and references are given to his two papers of 1845 and 1847, but his terminology is not adopted. Pearl and Reed call Verhulst's papers 'long since forgotten', except for a single article by Du Pasquier (1918), and they then go out of their way to criticize that author for an *entirely unjustified and in practice usually incorrect modification* of Verhulst's formula, without substantiating this harsh judgment. In fact Du Pasquier's paper is a harmless reflection on four mathematical theories of population, of a very formal and abstract character to the point of inanity. The four theories are linked to Halley, de Moivre, Euler and Verhulst, and these authors are briefly introduced; Halley, for example, as *"the famous astronomer"*, and Verhulst as *"a Belgian who died in 1847"*. No references are given. It is not clear how Du Pasquier knew about Verhulst, nor how Pearl and Reed knew about Du Pasquier. On the first point, Du Pasquier (1876 – ?), who was a Professor of Mathematics at the University of Neuchatel, in Switzerland, was educated in mathematics and economics in Zürich and Paris, and he may have read about Verhulst in the French literature. On the second point, Du Pasquier may well have taken the initiative in establishing contact with Pearl or Reed; after all, Pearl and Reed had published their paper in the *Proceedings of the National Academy of Sciences*, which would reach Switzerland, and Du Pasquier in the *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich* which was unlikely to reach Baltimore. But all this is pure speculation.

The next important publication is Yule's Presidential Address to the Royal Statistical Society of 1925. Yule, who says he owes the reference to Pearl (1922), treats Verhulst much more handsomely than Pearl and Reed did, devoting an appendix to his work and reviving the name *logistic*. It would take until 1933 for Miner (a collaborator of Pearl) to pay tribute to

Verhulst, if in an indirect way: instead of reproducing at least one of Verhulst's papers, he published a translation of Verhulst's obituary by Quetelet, with the addition of an extract from the memoirs of Queen Hortense de Beauharnais, who records the episode of Verhulst's stay in Rome in 1830.

As we have indicated there is another early root of the logistic function in chemistry, where it was employed (again with some variations) to describe the course of autocatalytic reactions. These are chemical chain reactions where the product itself acts as a catalyst for the process while the supply of raw material is fixed. This leads naturally to a differential equation like (8) and hence to the logistic function for the time path of the amount of the reaction product. The review of the application of logistic curves to a number of such processes by Reed and Berkson (1929) quotes work of the German professor of chemistry Wilhelm Ostwald of 1883. Authors like Yule (1925) and Wilson (1925) were well aware of this strand of the literature.

The basic idea of logistic growth is simple and effective, and it is used to this day to model population growth and market penetration of new products and technologies. The introduction of mobile telephones is an autocatalytic process, and so is the spread of many new products and techniques in industry.

3 The invention of the probit and the advent of the logit

The invention of the probit model is usually credited to Gaddum (1933) and Bliss (1934a), (1934b), but one look at the historical section of Finney (1971) or indeed at Gaddum's paper and his references will show that this is too simple. The roots of the method and in particular the transformation of frequencies to equivalent normal deviates can be traced to the German scholar Fechner (1801–1887). Stigler (1986) recounts how Fechner was drawn to study human responses to external stimuli by experimental test of the ability to distinguish differences in weight. The issue of the variability of human responses had been raised by astronomers, who relied on human observers of celestial phenomena and found that their readings showed much unaccountable variation. Fechner recognized that human response to an identical stimulus is not uniform, and he was the first to transform observed differences to equivalent normal deviates. The historical sketches of Finney (1971), Ch. 3.6, and of Aitchison and Brown (1957), Ch. 1.2, record a long line of largely independent rediscoveries of this approach that spans the seventy years from Fechner (1860) to the early 1930's when Gaddum and Bliss published their contributions. Both authors regard the assumption of a normal distribution as commonplace, and attach more importance to the logarithmic transformation of the stimulus. Their papers contain no major innovations, but they mark the emergence of a standard paradigm of bio-assay. Gaddum wrote a comprehensive and authoritative report with the emphasis on practical aspects of the experiments and on the statistical interpretation of bio-assay, giving several worked examples from the medical and pharmaceutical literature. Bliss published two brief notes in *Science*, introducing the term *probit*; he followed this up with a series of articles setting out the maximum likelihood estimation of the probit curve, in one instance with assistance from R.A. Fisher (Bliss (1935)). Both Gaddum and Bliss set standards of estimation; until the 1930's this was largely a matter of ad hoc numerical and graphical adjustment of curves to categorical data.

John Henry Gaddum (1900–1965) studied medicine at Cambridge but failed in his final examinations. He turned to pharmacology and worked under Trevan at the Wellcome Laboratories, then transferred to the National Institute for Medical Research (where he wrote the 1933 report) before he embarked on an academic career of professorships in pharmacology in Cairo, London and Edinburgh. He was elected to the Royal Society in 1945 and knighted in 1964. To this day the British Pharmacological Society awards an annual Gaddum Memorial Prize for pharmaceutical research.

Charles Ittner Bliss (1899–1979) studied as an entomologist at Ohio State University and was a field worker with the U.S. Department of Agriculture until this employment was terminated in 1933. He then spent two years in London studying statistics with R.A. Fisher, and Fisher found him a job in Leningrad where he lived from 1936 and 1938. The political conditions were not propitious for serious work. Bliss returned to the Connecticut Agricultural Experiment Station, combining his work as a practising statistician with a Lecturership at Yale from 1942 until his retirement. He played an important role in the founding of the Biometric Society.

In their early writings on bio-assay both authors adhere firmly to the classical model of bio-assay, where the stimulus is determinate and responses are random because of the variability of individual tolerance levels. Bliss introduced the term *probit* (short for 'probability unit') originally as a convenient *scale* for normal deviates, but abandoned this within a year in favour of a different definition which was generally accepted. For any (relative) frequency f there is an equivalent normal deviate $\tilde{Z}$ such that the cumulative normal distribution at $\tilde{Z}$ equals f; $\tilde{Z}$ is the solution of

$$f = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{\tilde{Z}} \exp\{-1/2u^2\}du,$$

and can be read of from a table of the normal distribution. The probit is the equivalent normal deviate increased by 5. This ensures that the probit is almost always positive, which facilitates calculation; at the time such additive constants were a common device.

The acceptance of the probit method was aided by the articles of Bliss, who published regularly in this field until the 1950's, and by Finney and others (Gaddum returned to pharmacology). The full flowering of this school in bio-assay probably coincides with the first edition of Finney's monograph in 1947. Applications in other fields like economics and market research appear already in the 1950's: Farrell (1954) employed a probit model for the ownership of cars of different vintage as a function of household income, and Adam (1958) fitted lognormal demand curves to survey data of the willingness to buy cigarette lighters and the like at various prices. The classic monograph on the lognormal distribution of Aitchison and Brown (1957) brought probit analysis to the notice of a wider audience of economists.

As far as I can see the introduction of the logistic as an alternative to the normal probability function is the work of a single person, namely Joseph Berkson (1899–1982), Reed's co-author of the paper on autocatalytic functions of 1929. Berkson read physics at Columbia, then went to Johns Hopkins for his M.D. and a doctorate in statistics in 1928. He stayed on as an assistant for three years and this is when he collaborated with Reed on autocatalytic functions. Berkson then moved to the Mayo Clinic where he remained for the rest of his working life as chief statistician. In the 1930's he published numerous papers on medical and public health matters, but in 1944 he turned his attention to the statistical methodology of bio-assay and proposed the use of the logistic, coining the term 'logit' by analogy to the 'probit' of Bliss (for which he was initially much derided). The issue of logit versus probit was tangled by Berkson's simultaneous attacks on the method of maximum likelihood and his advocacy of minimum chi-squared estimation instead. Between 1944 and 1980 he wrote a large number of papers on both issues; examples are Berkson (1951) and Berkson (1980). He often adopted a somewhat provocative style, and much controversy ensued.

The close resemblance of the logistic to the normal distribution function must have been common knowledge among those who were familiar with the logistic; it had been demonstrated by Wilson (1925) and written up by Winsor (1932) (another collaborator of Pearl). Wilson was probably the first to publish an application of the logistic in bio-assay in Wilson and Worcester (1943), just before Berkson (1944). But it was Berkson who persisted and fought a long and spirited campaign which lasted for several decades.

Berkson's suggestion was not well received by the biometric establishment. In the first place, the logit was regarded as somewhat inferior and disreputable because unlike the probit it can not be related to an underlying (normal) distribution of tolerance levels. Aitchison and Brown (1957) dismiss the logit in a single sentence, because it *"lacks a well-recognized and manageable frequency distribution of tolerances which the probit curve does possess in a natural way"* (p.72). Berkson was aware of this defect and tried to remedy it by adapting the autocatalytic argument, in Berkson (1951), but this did not convince as the autocatalytic argument essentially deals with a process over time. In retrospect it is surprising that so much importance was attached to these somehwat ideological points of interpretation. At the time no one (not even Berkson) seems to have recognized the formidable power of the logistic's analytical properties. In the second place, Berkson's case for the logit was not helped by his simultaneous attacks on the established wisdom of maximum likelihood estimation and his advocacy of minimum chi-squared. The unpleasant atmosphere in which this discussion was conducted can be gauged from the acrimonious exchanges between R.A. Fisher and Berkson in Fisher (1954).

In the practical aspect of ease of computation the logit had a clear advantage over the probit, even with maximum likelihood estimation. To quote Cochran (from his comments on Fisher (1954), p.147.) *".. the speed with which a new technique becomes widely used is considerably influenced by the simplicity or otherwise of the calculations that it requires. Next door to the lecture room in which the probit method is expounded one may still find the laboratory in which the workers compute their LD 50s by the* [much less sophisticated] *Behrens (Reed–Muench) method .."*. On this count the logit spread much more quickly in workfloor practice than in the academic discourse. Until the advent of the computer and the pocket calculator, some trwenty years later, all numerical work was done by hand, that is with pencil and paper, sometimes aided by graphical inspection of 'freehand curves', 'fitted by eye'. For probit and logit analyses of grouped data or class frequencies there was graph paper with a special grid on which a probit or logit curve would appear as a straight line. Wilson (1925) had introduced the logistic (or 'autocatalytic') grid, and examples of lognormal paper can be found in Aitchison and Brown (1957) and Adam (1958);[2] Berkson himself had designed logistic graph paper as well as several nomograms.[3] Numerical work was supported rather feebly by the slide rule and by mechanical calculating machines, driven by hand or powered by a small electric motor, which were capable of addition and multiplication; punched card equipment was helpful if numerous data had to be analysed. Values of the normal distribution (and of exponentials and logarithms) were obtained from printed tables like Pearson's *Biometrika* Tables or the *Statistical Tables* of Fisher and Yates (1938). From the first edition the latter carried specially designed tables for probit analysis (with auxiliary tables contributed by Bliss and by Finney), but from the fifth edition of 1957 onwards they also included special tables for logit analysis.

In time, the ideological conflict over bio-assay abated. Finney, who had ignored the logit in the second edition of his textbook of 1952, made amends in the third edition of 1970, recognizing that *"what matters is the dependence of P on dose and the unknown parameters, and the tolerance distribution is merely a substructure leading to this"*. Between 1960 and 1970 the logit indeed gradually achieved an equal footing with the probit. By then it was also slowly recognized that its analytical properties permit much wider statistical applications, beyond bio-assay: it can be linked to discriminant analysis, it leads to loglinear models, it can be used with retrospective samples as in case-control studies, and so on. One of the first to recognize and exploit these avenues was Cox, in a series of articles in the 1960's, and in Cox (1969). This general development is illustrated in Table 1, which is drawn from the JSTOR electronic repertory of twelve major statistical journals in the english language. The table show the number of articles which contain the word "probit" or "logit". The number of statistical journals included in JSTOR increases over time, as does the number of articles in each journal; from 1935 to 1985 the total number of articles covered annually increases about eightfold. It is therefore the *relative* position of "probit" and "logit" that counts. By 1970 logit reaches parity, and thereafter soars ahead.

[2] Finney (1947) traces the invention of the probability grid to a French artilleryman of the late 1890's.

[3] A nomogram is a graph from which one can read off a transformations, as from a table; sophisticated nomograms may permit the quick solution of more complicated equations.

Table 1. Number of articles in statistical journals containing the word 'probit' or 'logit'.

	probit	logit
1935 – 39	6	-
1940 – 44	3	1
1945 – 49	22	6
1950 – 54	50	15
1955 – 59	53	23
1960 – 64	41	27
1965 – 69	43	41
1970 – 74	48	61
1975 – 79	45	72
1980 – 84	93	147
1985 – 89	98	215
1990 – 94	127	311

Both probit and logit were also adopted beyond bio-assay, in economics, in epidemiology and in the social sciences. The close link to tolerance levels or threshold values was dissolved and less stringent interpretations were admitted; the elegant but quite abstract model of the latent regression equation was probably first explicitly formulated by McKelvey and Zavoina (1975) for an ordered probit model of the voting behaviour of U.S. Congressmen, far removed from bio-assay. Analyses linking binary discrete responses to several covariates became known as logistic regression. This wider acceptance was greatly helped by the advent of the computer and by the introduction of package routines for the maximum likelihood estimation of both logit and probit models from individual data. The BMDP or BIOMEDICAL DATA PROCESSING computer package of 1977 was probably the first to offer this facility, which soon became a standard feature of most statistical packages. By the time the first comprehensive textbook of Hosmer and Lemeshow (1989) appeared the use of such routines was taken for granted.

Of the two causes Berkson advocated, minimum chi-squared was thus overtaken by the computer revolution, but the logit was there to stay. Its multinomial generalization was first mooted by Cox (1966) and then, independently, by Theil (1969) who immediately saw its potential as a general approach to the modelling of shares. The simple algebra of this generalisation opened up a very wide field of applications in economics and other social sciences, and interest in an interpretation in terms of an underlying process waned. But in 1973 McFadden, working as a consultant for a Californian public transportation project, first linked the multinomial logit to the theory of discrete choice from mathematical psychology. This provided a theoretical foundation of the logit model that is much more profound than any theory put forward for the use of the probit in bio-assay. It earned McFadden a Nobel prize in 2000.

References

Adam, Daniel (1958), *Les réactions du consommateur devant les prix.* Paris: Sedes.

Aitchison, J. and J.A. C. Brown (1957) *The Lognormal Distribution.* Cambridge: Cambridge University Press.

Berkson, J. (1944) Application of the Logistic Function to Bio-assay. *Journal of the American Statistical Association*, **9**, 357–365.

Berkson, J. (1951) Why I Prefer Logits to Probits. *Biometrics*, **9**, 357–365.

Berkson, J. (1980) Minimum Chi-Square, Not Maximum Likelihood! *Annals of Mathematical Statistics*, **8**, 457–487.

Bliss, C.I. (1934a) The Method of Probits. *Science*, **79**, 38–39.

Bliss, C.I. (1934b) The Method of Probits. *Science*, **79**, 409–410.

Bliss, C.I. (1935) The Calculation of the Dosage-Mortality Curve (with an appendix by R.A. Fisher). *Annals of Applied Biology*, **22**, 134–167.

Cox, David R. (1966) Some Procedures Connected with the Logistic Qualitative Response Curve. In: F.N. David (ed) *Research Papers in Statistics: Festschrift for J. Neyman.* London: Wiley.

Cox, David R. (1969) *Analysis of Binary Data.* London: Chapman and Hall.

Du Pasquier, Louis-Gustave (1918) Esquisse d'une nouvelle théorie de la population. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, **63**, 236–249.

Fechner, G.T. (1860) *Elemente der Psychophysik.* Leipzig: Breitkopf und Härtel.

Finney, D. (1971) *Probit Analysis.* Cambridge: Cambridge University Press. (third edition; first edition 1947)

Fisher, Sir Ronald (1954) The Analysis of Variance With Various Binomial Transformations (With Comments by M. S. Bartlett, F. J. Anscombe, W. G. Cochran and J. Berkson), *Biometrics*, **10**, 130–151.

Gaddum, J.H. (1933) *Report on Biological Standards III: Methods of Biological Assay Depending on Quantal Response.* Special Report Series of the Medical Research Council, no.183. London: Medical Research Council.

Hosmer, David W. and Stanley Lemeshow (1989) *Applied Logistic Regression* New York: Wiley.

Malthus, T.R. (1798) *An Essay on the Principle of Population.* London.

McFadden, Daniel (2001) Economic Choices. (Nobel prize acceptance speech.) *American Economic Review*, **91**, 352–370.

Meadows, D.H., Meadows, D.L., Randers J. and W.W. Behren (1972) *The Limits to Growth* New York: Universe Books.

Miner, J. R. (1933) Pierre-François Verhulst, the Discoverer of the Logistic Function. *Human Biology*, **5**, 673–689.

Pearl, Raymond (1922) *The Biology of Death.* Philadelphia: Lippincott.

Pearl, Raymond (1927) The Indigenous Population of Algeria in 1926. *Science* **66**, 593–594.

Pearl, Raymond (1939) *The Natural History of Population* Oxford: Oxford University Press.

Pearl, Raymond and Lowell J. Reed (1920) On the Rate of Growth of the Population of the United States and its Mathematical Representation. *Proceedings of the National Academy of Sciences*, **6**, 275–288.

Pearl, Raymond and Lowell J. Reed (1922) A Further Note on the Mathematical Theory of Population Growth. *Proceedings of the National Academy of Sciences*, **8**, 365–368.

Pearl, R., C.P. Winsor and F.B. White (1928) The Form of the Growth Curve of the Cantaloupe (Cucumis melo) under field conditions. *Proceedings of the National Academy of Sciences*, **14**, 895–901.

Pearl, R., L.J. Reed and J.F. Kish (1940) The Logistic Curve and the Census Count of 1940. *Science*, **92**, 486–488.

Reed, L.J., and J. Berkson (1929) The Application of the Logistic Function to Experimental Data. *Journal of Physical Chemistry*, **33**, 760–779.

Stigler, S.M. (1986) *The History of Statistics.* Cambridge, Mass.: Harvard University Press.

Theil, H. (1969) A multinomial extension of the linear logit model. *International Economic Review*, **10**, 251–259.

Verhulst, Pierre-François (1838) Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et Physique, publiée par A. Quetelet*, **10**, 113–120.

Verhulst, Pierre-François (1845) Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, **18**, 1–38.

Verhulst, Pierre-François (1847) Deuxième Mémoire sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, **20**, 1–32.

Wilson, Edwin B. (1925) The Logistic or Autocatalytic Grid. *Proceedings of the national Academy of Sciences*, **11**, 431–456.

Wilson, E.B. and Jane Worcester (1943) The Determination of L.D.50 and Its Sampling Error in Bio-assay. *Proceedings of the National Academy of Sciences*, **29**, 79–85.

Winsor, C.P. (1932) A Comparison of Certain Symmetrical Growth Curves. *Proceedings of the Washington Academy of Sciences*, **22**, 73–84.

Yule, G. Udney (1925) The Growth of Population and the Factors which Control It. *Journal of the Royal Statistical Society*, **38**, 1–59.

Other Biographical Sources

On Pearl:

Jennings, H.S. (1941) Raymond Pearl, 1879–1940. *Biographical Memoirs of the National Academy of Sciences of the United States*, bf 22, nr. 14, 295–347.

Miner, John R. and Joseph Berkson (1940) Raymond Pearl, 1879–1940. *The Scientific Monthly*, **52**, 1092–194.

On Reed:

Cochran, W.G. (1967) Lowell Jacob Reed. *Journal of the Royal Statistical Society, Series A*, **130**, 279-281.

On Gaddum:

Feldberg, W. (1967) John Henry Gaddum, 1900–1965. *Biographical Memoirs of Fellows of the Royal Society*, **13**, 57–77.

On Berkson:

Armitage, P., and T. Colton (eds) (1998) Joseph Berkson 1899–1982. *Encyclopedia of Biostatistics*, volume I, 290–300. New York: Wiley.

Taylor, W.F. (1983) Joseph Berkson, 1899-1982. *Journal of the Royal Statistical Society, Series A*, **146**, 413–419.

WORKING PAPER

Introduction to Logistic Regression Models

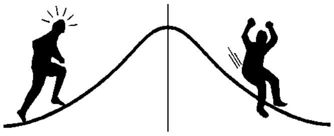
With Worked Forestry Examples

Biometrics Information Handbook No.7

26/1996

British Columbia

Ministry of Forests Research Program



Introduction to Logistic Regression Models

With Worked Forestry Examples

Biometrics Information Handbook No.7

Wendy A. Bergerud

British
Columbia

Ministry of Forests Research Program

The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the Government of British Columbia of any product or service to the exclusion of any others that may also be suitable. Contents of this report are presented for discussion purposes only. Funding assistance does not imply endorsement of any statements or information contained herein by the Government of British Columbia.

Citation:
Bergerud, W.A. 1996. Introduction to Regression Models: with worked forestry examples. Biom. Info. Hand. 7. Res. Br., B.C. Min. For., Victoria, B.C. Work. Pap. 26/1996.

Prepared by
Wendy A. Bergerud
for
B.C. Ministry of Forests
Research Branch
31 Bastion Square
Victoria, B.C. v8w 3e7

Copies of this report may be obtained, depending upon supply, from:
B.C. Ministry of Forests
Forestry Division Services Branch
595 Pandora Avenue
Victoria, BC v8w 3e7

The contents of this report may not be cited in whole or in part without the approval of the Director of Research, B.C. Ministry of Forests, Victoria, B.C.

ACKNOWLEDGEMENTS

I began my Masters Degree at the University of Waterloo in order to study logistic regression. Two professors in particular, Steven Brown and Jock McKay, were very helpful in my learning process. This handbook has been in the works for many years and I have received useful comments from many. (I apologize to those whose names should be here but who have been lost in the mists of my mind.) I would like to thank Les Peterson for his example in Appendix 1. Comments from Amanda F.L. Nemec helped clarify the notation and discussion. Linda Stordeur's request for fully worked examples doubled the handbook's size but, I believe, made it more useful. Susan Bannerman's comments on the writing and organization of the material were also very helpful and helped make the material easier to read. Other reviewers include John Stevenson, Rick Page, Chris Thompson, Vera Sit, Peter Ott, Lesley McKnight, and Bill Roland. I would also like to thank the Production Resources staff, Heather Strongitharm, David Izard, and Paul Nystedt, for guiding this through the publication stage, with special thanks to David Izard for tidying up the SAS/Graph figures that I produced. I would like to thank Neil West for typesetting this publication, and Paul Nystedt, David Izard and Stephen Stearns-Smith for helping me find publication funding and to Chris Hyde at the Pacific Forestry Centre (Canadian Forest Service) for approving and administering the FRDA II ETT funds for printing. I would particularly like to thank John Stevenson for believing in this project and pushing to get it completed—without his support, it would not have been published.

TABLES

FIGURES

1 INTRODUCTION

Logistic regression is a useful tool for analyzing data that includes categorical response variables, such as tree survival, presence or absence of a species in quadrats, and presence of disease or damage to seedlings. The models work by fitting the probability of response to the proportions of responses observed. For instance, the number of outplanted seedlings in 50-tree rows that die from frost damage is an observed response. These observed numbers are converted to proportions which are then fitted by models that determine the probability that a seedling will die from frost damage. Normal distribution approximations to the proportions and the consequent analytical methods (e.g., regression and analysis of variance) can be used if large sample sizes exist for each experimental unit. However, logistic regression does not require large sample sizes for the data analysis to be feasible. Furthermore, it is possible to analyze individual tree data.

Five forestry examples are used extensively in this handbook to illustrate possible study designs and the statistical aspects of logistic regression analysis. The first example, a simple regression, examines the relationship of the survival of caribou calves during their first year to the number of wolves in their vicinity. The second example studies the relationship between tree survival and age class in stands with root rot. Stands that are "similar" according to some criteria are selected from available stands resulting in an unbalanced one-way classification study. The third example is a controlled experiment, where rows of seedlings are treated with different amounts of fertilizer. With height as the response variable, this can be analyzed with a familiar one-way analysis of variance (ANOVA). We shall use logistic regression analysis to study the effect of fertilizer amounts on the probability of seedling survival. The fourth example is a traditional multiple regression situation, where two qualitative variables, tree size and amount of herbicide, are used to predict the probability that treated trees die. This example is based on a trial conducted by the British Columbia Ministry of Forests and uses real data. The fifth example would be a traditional analysis of covariance if seedling height were the response variable of interest. However, here the effectiveness of screefing around outplanted seedlings to reduce attack by the root collar weevil is examined. The attack of root collar weevil is quantified by counting the number of seedlings attacked in each plot. This example leads us into two interesting discussions: first, about whether traditional analysis of covariance models are always appropriate when the explanatory variables are both categorical and continuous; and second, about the shape of the logistic regression models. Since this handbook is an introduction to logistic regression models all of these examples are relatively simple. Nevertheless, these examples provide the necessary bulding blocks for understanding and interpreting more complicated study designs.

Chapter 2 discusses statistical models in general and the logistic regression model for two response categories in particular. Chapter 3 is considerably more technical. Methods of, and problems with, fitting logistic regression models, parameter estimation, and testing are discussed. Chapter 4 describes the study design examples more fully. Models are set up for each of the examples and contrasted with the corresponding normal distribution models with which most readers will be familiar. Each example is then explored in more depth by examining the specific data sets and fitting the models. Chapter 5 provides a detailed discussion of the SAS programs used to calculate and fit logistic regression models for each example. The last two chapters cover some advanced topics regarding indicator (or dummy) variables and other methods to fit the one-way classification models.

Readers should be familiar with common parametric statistical techniques, such as contingency tables, *t*-tests, simple and multiple linear regression, analysis of variance and covariance, and should know how to use statistical tables for the *t*-, *F*- and χ^2-distributions. Familiarity with SAS, particularly with `PROC GLM` and `PROC REG`, will also be helpful, but is not essential for understanding the example studies and their analysis. A good reference to use along with this handbook is Agresti (1996). Biometrics Handbook No. 1, "Pictures of Linear Models" (Bergerud 1991) may be useful to help develop a basic understanding of simple linear models.

2 STATISTICAL MODELS

Every statistical test has an associated statistical model. Whenever a test is used on data, it is assumed that the associated model fits the data reasonably well. If this is not the case, then the test is inappropriate and the results could be misleading. On the other hand, several statistical models may fit the same data reasonably well. Since different statistical models usually lead to different statistical tests, this can mean that the data could be tested in several different ways. The choice of test depends on the choice of statistical model. Although the fit of models can sometimes be tested, and even compared, the final choice of a model is, in general, a non-statistical one.

Statistical models have two components:[1]

1. a deterministic or systematic component, and
2. a stochastic or random component.

The deterministic component is a function that describes the expected or predicted value of the response variable and usually is of most interest to the scientist or researcher. It often has a parametric form. This means that it can be specified by one or more unknown parameters or constants which are estimated in the fitting procedure.[2] This report discusses linear parametric models in which the parameters appear as coefficients in a

[1] A third component, called the link function, is described by McCullagh and Nelder (1983), Dobson (1983), and Gilchrist (1984).

[2] Analyses of variance fit into this class of model. Because means are usually of interest, the estimated parameters are not automatically output by computer procedures. (See section 6.)

TABLE 1 *The deterministic component of common types of linear models*

Common name	Deterministic component of model
Mean or intercept only	$\mathrm{E}(Y_j) = \mu$
Simple regression	$\mathrm{E}(Y_j \mid x_j) = \mu + \beta x_j$
Multiple regression	$\mathrm{E}(Y_j \mid x_j, z_j) = \mu + \beta x_j + \gamma z_j$
One-way classification (one-way ANOVA)	$\mathrm{E}(Y_{ij}) = \mu + \alpha_i$
One-way classification with a covariate (one-way ANCOVA)	$\mathrm{E}(Y_{ij} \mid x_{ij}) = \mu + \alpha_i + \beta x_{ij}$

Where:

i, j are indices that uniquely identify specific observations or experimental units.

$\mathrm{E}(Y_j)$ or $\mathrm{E}(Y_{ij})$ are the expected responses, while $\mathrm{E}(Y_j \mid x_j)$ is the expected response for a given x_j, and $\mathrm{E}(Y_j \mid x_j, z_j)$ is the expected response given a pair of values for x_j and z_j (the | stands for "given").

μ, β, γ, and α_i are unknown quantities, referred to as *parameters*.

μ, β, and γ are constant for all observations being modelled.

β and γ are often called *slopes*, *regression parameters*, or *coefficients*.

α_i are constants that may be different for different levels of the classifying variables (where the different levels are denoted by i).

x_j, x_{ij}, and z_j are continuous-valued explanatory variables used to predict the responses.

sum of simple terms. Common types of linear models are summarized in Table 1.

The deterministic component is quite distinct from the stochastic component of a statistical model. The stochastic component describes the random variation of the response variable. It provides the basis for statistical tests by specifying a suitable probability distribution for the data. The development of these statistical tests is often mathematically complicated.

The t-test is a simple and familiar statistical test used to illustrate the roles played by the deterministic and stochastic components. The underlying model is:

$$Y_j = \mu + \varepsilon_j,$$

where: Y_j is the j^{th} response (e.g., height or diameter of the j^{th} tree),
μ is the mean or expected value of Y_j, and
ε_j is a random error or residual (i.e., it is the difference $Y_j - \mu$ between the expected value μ and the observed value Y_j).

Here, the deterministic component indicates that the expected value of each observed value Y_j is theoretically equal to some unknown constant μ. The stochastic component indicates that the differences, $\varepsilon_j = Y_j - \mu$, are independent and vary randomly around a mean value of zero according to the normal or gaussian distribution with an unknown variance of σ^2. Equivalently, Y_j has a normal distribution with mean μ and variance σ^2.

If the collected data fit the above model, then the ratio of the difference between the sample mean and μ to the estimated standard error of the sample mean will have a (central) t-distribution. This distribution describes the probability of obtaining a value at least as large as the absolute value of any particular observed t-value, given that the null hypothesis (H_0: $\mu = c$) is correct. This can be used to test whether the data are consistent with the null hypothesis by making the following calculation:

$$t_{obs} = \frac{\bar{y} - c}{SE(\bar{y})}$$

where: $\bar{y}$ is the sample mean, and

$SE(\bar{y}) = \sqrt{s^2/n}$ is its standard error calculated from the sample variance s^2 and sample size n.

Note that $SE(\bar{y})$ is a measure of precision for the estimate ($\bar{y}$) of the mean μ. The observed t-value, t_{obs}, is compared to the t-distribution with degrees of freedom, $df = n - 1$, since the shape of the distribution depends on the degrees of freedom. If the observed t-value is improbable (based on its associated probability value from the t-distribution) given the null hypothesis that the unknown constant μ is equal to c, then that hypothesis is rejected.

Thus, the simple t-test is based on a model with a fairly complicated stochastic component and a relatively simple deterministic component. Every time the simple t-test is used, this model is assumed true for the data analyzed.

2.1 General Linear Models (GLM)

The most commonly used statistical tests are developed from models that assume independent responses and follow a normal distribution with a constant variance. The mean is the deterministic component of these models and may be expected to vary according to the independent variables in the model. Some common examples are presented in Table 1. The variance and parameters of the deterministic component are usually unknown. The term, general linear model (GLM), refers to linear models having the normal distribution as the stochastic component.

Statistical tests of the model parameters are derived by comparing the ratio of two variances, called an F-ratio. For models with simple stochastic components, the denominator of this F-ratio is the sample variance of the residuals (i.e., the mean squares of the differences between the observed and predicted values). The numerator has the same expected value as the variance in the denominator, if certain parameters of the deterministic component are zero. For instance, in a one-way classification with equal sample sizes, the variance of the group means multiplied by the sample size is the numerator of the F-ratio and is expected to be similar in magnitude to the variance of the residuals if all means have the same value or, equivalently, if all the α_i's shown in Table 1 are zero. The variance of the residuals is often called the mean square error. This ratio of variances can be tested for unlikely values by examining the corresponding significance obtained by comparison with the F-distribution.[3]

[3] Since the square of a t-value has an F-distribution, the two distributions are equivalent for testing purposes in the case of two groups. See Biometrics Information Pamphlet No. 27.

2.2 Stochastic Components

This section briefly describes two distributions often used to describe or model the stochastic component of linear models. The normal or gaussian distribution is most familiar because it is used with general linear models. The binomial distribution[4] provides the stochastic component for logistic regression models.

2.2.1 The normal distribution The two parameters of the normal distribution are:

1. the mean (μ), and
2. the variance (σ^2).

These parameters are independent of each other in the sense that the mean does not determine the value of the variance. The mean can take on any real value, while the variance must be positive. The influence that μ and σ^2 have on the shape of the normal distribution are illustrated in Figure 1. Note that the distribution is smooth and is always symmetric about the mean.

This distribution is often used to approximate the mean of data generated from other distributions. It does this especially well if the sample size for the mean is large and the other distribution is symmetric.

2.2.2 The binomial distribution The two parameters of the binomial distribution are:

1. the sample size,[5] (m) which is the number of sampling units per experimental unit,[6] and
2. the probability of some specified event (π), which is often called a "success."[7]

The sample size must be a positive integer and the probability can only have a value between zero and one. Both parameters can have any value in their allowed range regardless of the value of the other. The response variable is either the proportion or the number (out of the m sampling units) of successes for a given experimental unit. The response variable, number of successes, has mean ($m\pi$) and variance [$m\pi(1 - \pi)$].

[4] Referred to as the multinomial distribution if there are more than two categories of response. Note that the discussion of three or more response categories is beyond the scope of this handbook.

[5] Notation for the sample size can be confusing. In this text, m is used for the number of sampling units per experimental unit, while n is used for the total number of sampling units.

[6] An experimental unit is a basic unit of experimental material to which one level of a treatment is applied. An experimental unit may be composed of many sampling units upon which actual measurements or responses are taken. See Biometrics Information Pamphlet Nos. 5, 17, and 55 for some discussion of this.

[7] Note that π is the Greek symbol for the letter "p." It is used here to represent the unknown, but true value, of the probability of the binomial distribution, just as μ is used to stand for the mean of the normal distribution. It should not be confused with the mathematical symbol π, which is used to represent the ratio of the circumference of a circle to its diameter.

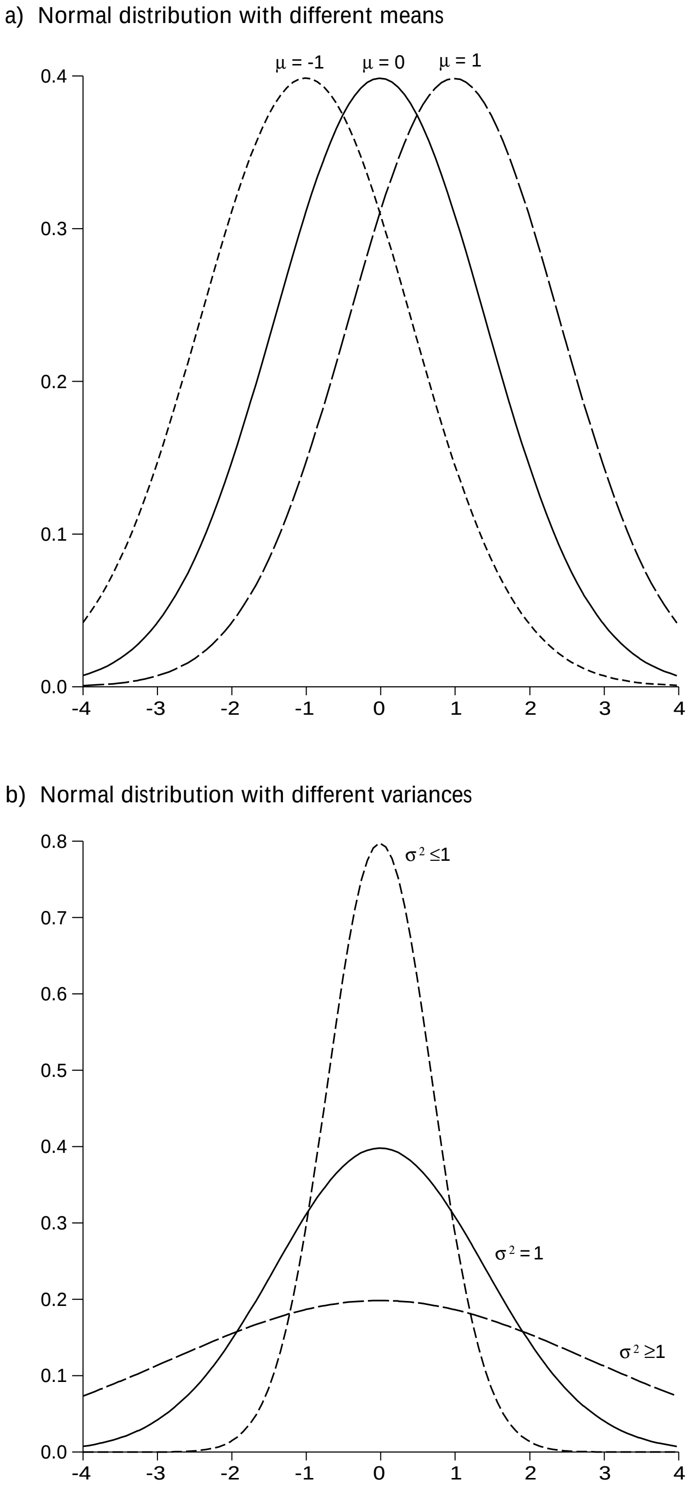


FIGURE 1 *The normal distribution with different values for (a) the mean, and (b) the variance.*

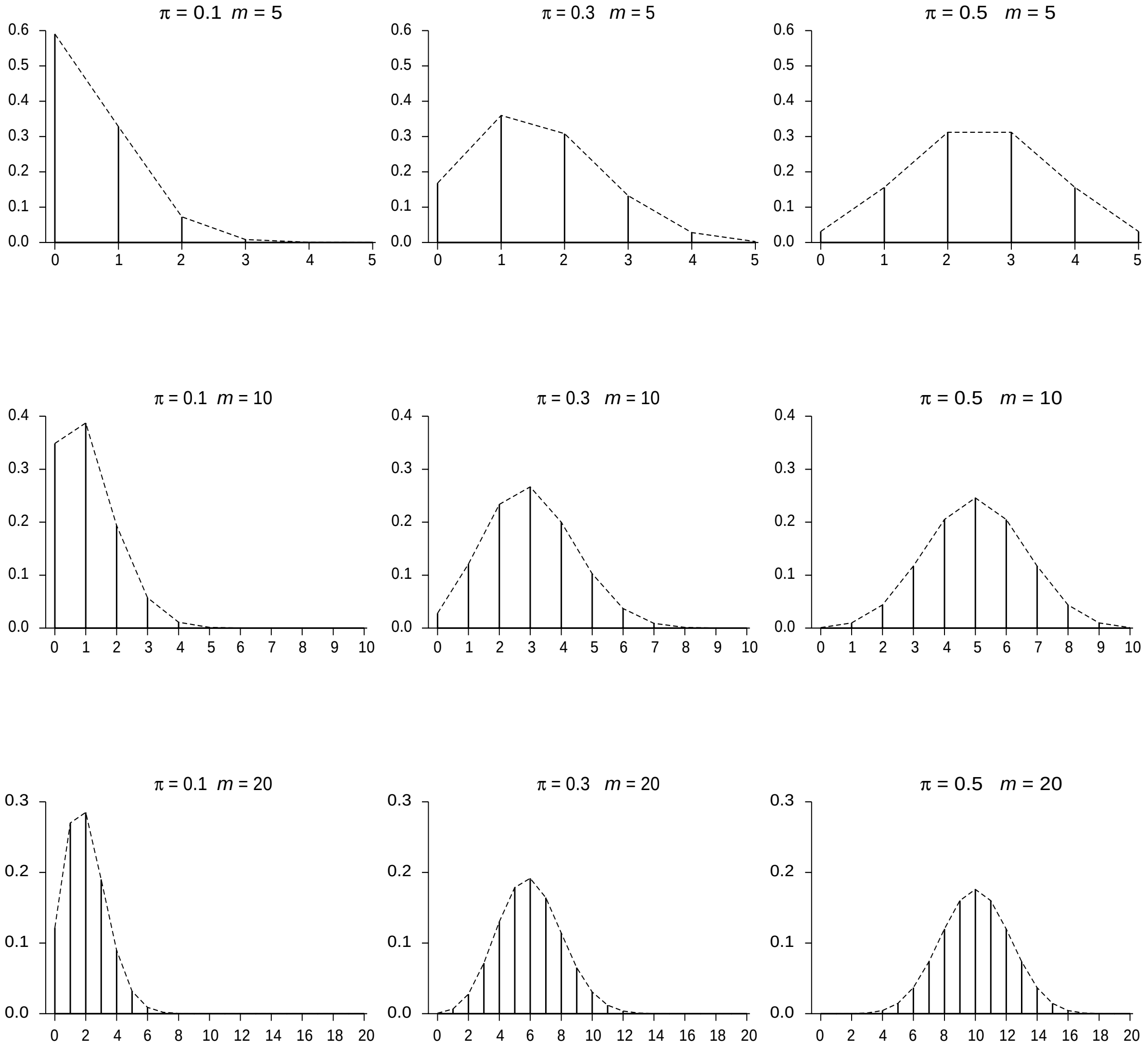


FIGURE 2 *The binomial distribution for various values of π and* m.

Note that the variance is a function of the mean. The sample size directly influences the form of the distribution by limiting the maximum number of successes possible. Some possible forms of the binomial distribution are shown in Figure 2.

2.2.3 Comparison of the normal and binomial distributions The normal distribution is continuous and symmetric with no restrictions on the possible values of the response variable. The binomial distribution is discontinuous, asymmetric unless $\pi = 0.5$, and the response variable is limited to the range of integer values between zero and the sample size inclusive. While sample size is an explicit parameter of the binomial distribution, it is also important for normal distributions because the variance of a mean depends on the sample size. For instance, the mean of normally distributed data is also normally distributed with the same mean, but with a variance reduced by the sample size (i.e., σ^2/m).

Binomial data are often approximated by a normal distribution. According to one rule, this is appropriate when both success and failure mean counts, $m\pi$ and $m(1 - \pi)$, are greater than five. The binomial distribution is reasonably symmetric and multi-valued when this is the case (see Figure 2 for m = 20, π = 0.3; m = 10, π = 0.5; and m = 20, π = 0.5). For various values of π, the corresponding minimum sample size required to use the normal approximation is shown in Table 2.

TABLE 2 *Minimum sample size required to maintain* $m\pi$ *= 5 and corresponding variance of the binomial distribution*

π	$(1 - \pi)$	Minimum sample size, m	Variance
0.5	0.5	10	2.5
0.4	0.6	13	3.1
0.3	0.7	17	3.6
0.2	0.8	25	4.0
0.1	0.9	50	4.5
0.05	0.95	100	4.75
0.01	0.99	500	4.95

If π is approximately 0.05 or 0.95, then experimental units with approximately 100 sampling units will be required. Therefore, 100 measurements are needed to determine a mean response for *each* experimental unit in a regression or ANOVA. However, the use of familiar statistical methods for data analysis is a substantial advantage, if such large experiments are feasible. These methods assume homogeneity of variance for all experimental unit means, which is clearly incorrect for data that are binomially distributed (since the variance depends on π). The angular transformation (i.e., arcsine square root) of percentage data is usually recommended to rectify this situation. For a constant sample size, this transformation will not make much difference, unless probabilities fall below 0.05 or exceed 0.95, and data with probabilities of around 0.5 are also present. Occasionally, the required sample size is so large that the study becomes impractical or the phenomenon of real interest can not be investigated. Logistic regression methods use the binomial distribution with its non-constant variance to model the data. This allows trials to be designed on a smaller scale. Effects that are only practical or meaningful with smaller sample sizes may then be studied.

2.3 Logistic Regression Models

Logistic regression models use the logistic function to fit models to data. This is an S-shaped function and an example curve is shown in Figure 3. This function can be used to fit data in three ways. Although each is distinct, these approaches can be called logistic regression and are briefly described in Table 3. They all fit a response variable, either y or y/m, to the S-shaped logistic function of the independent variable, x. The first model could fit growth data (y on any scale) versus time (x) with a logistic curve, while the next two fit proportional responses (with values restricted to the range between zero and one) with the logistic curve. The

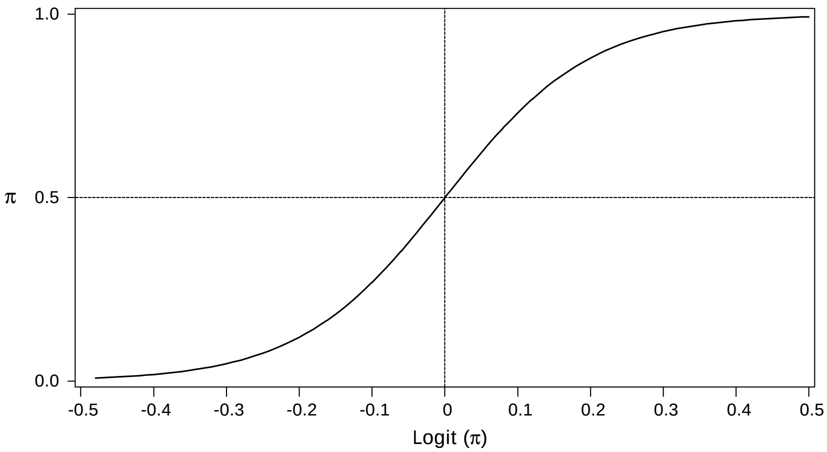


FIGURE 3 *The logistic curve with a maximum value of one (i.e., $\gamma = 1$): $\pi = \exp(\text{logit})/[1 + \exp(\text{logit})]$.*

TABLE 3 *Three types of logistic regression models*

Model	Deterministic component	Stochastic component
1	$E(y) = \gamma / [\gamma + \exp(\alpha + \beta x)]$	y is normal
2	$\pi = E(y/m) = \exp(\alpha + \beta x) / [1 + \exp(\alpha + \beta x)]$	$p = y/m$ is approximately normal
3	$\pi = E(y/m) = \exp(\alpha + \beta x) / [1 + \exp(\alpha + \beta x)]$	y is binomial

Where:
y is a continuous response variable in model 1 and y is a count (out of m) in models 2 and 3;
$E(y)$ is the expected value of y; $\exp(x)$ = exponential function of x;
p is the proportion of success when the response variable is recorded as a count;
x is a fixed independent variable; and
γ, α, and β are unknown parameters.

first two models assume that the data follow, at least approximately, a normal distribution, while the third assumes that the data are binomially distributed. This report is restricted to a discussion of the third type of model.

For this third type of logistic model, the parameter π of the binomial distribution represents the "true" but unknown probability of success. It is transformed by the logit function to create linear models in the independent variables. The logit transform, also called the log-odds of π, is:

$$\text{logit}(\pi) = \log[\pi/(1 - \pi)].$$

The inverse of this transform is $\pi = \exp(\text{logit})/[1 + \exp(\text{logit})]$ or $1 - \pi = 1/[1 + \exp(\text{logit})]$. These functions are variations of the logistic function and both can be used to fit data. They have the same S-shape as the example shown in Figure 3. The correspondence between some values of logits and probabilities is shown in Table 4. The deterministic components of logistic regression models are formed by equating this logit

TABLE 4 *Correspondence between logits and probabilities, π*

π	Logit(π)	$1 - \pi$	Logit(π)	π	$1 - \pi$
			–5.00	0.0067	0.9933
0.01	–4.60	0.99			
			–4.00	0.018	0.982
			–3.00	0.047	0.953
0.05	–2.94	0.95			
0.10	–2.20	0.90			
			–2.00	0.12	0.88
0.20	–1.39	0.80			
			–1.00	0.27	0.73
0.50	0.00	0.50	0.00	0.50	0.50
			1.00	0.73	0.27
0.80	1.39	0.20			
			2.00	0.88	0.12
0.90	2.20	0.10			
0.95	2.94	0.05			
			3.00	0.95	0.047
			4.00	0.98	0.018
0.99	4.60	0.01			
			5.00	0.9933	0.0067

to a general linear model, examples of which were shown in Table 1. For example, a simple linear logistic regression model is written as:

$$\text{logit}(\pi) = \log[\pi/(1 - \pi)] = \alpha + \beta x. \qquad (1)$$

Models that are linear after an appropriate transformation of the expected response, and whose stochastic component is not limited to the normal distribution, are often called *generalized linear models* (as opposed to general linear models).[8] The logistic regression models discussed here are a type of generalized linear model. They also belong to a group of models referred to as *log-linear models*, since a log transform of the response variable is used to give a linear form to the model.

The final fitted model will predict values on the logit scale. These can be back-transformed to probabilities by using the inverse functions:

$$\pi = \exp(\text{logit})/[1 + \exp(\text{logit})] \text{ or } 1 - \pi = 1 / [1 + \exp(\text{logit})] \qquad (2)$$

Discussions of the model-fitting process will include both the logit scale on which models are built and the probability scale on which interpretations are made (see Figure 4).

2.3.1 Suitable types of data Logistic regression models with a binomial stochastic component can provide adequate models for data with certain

[8] A more precise definition is found in many texts (see, for example, McCullagh and Nelder [1983, 1989]).

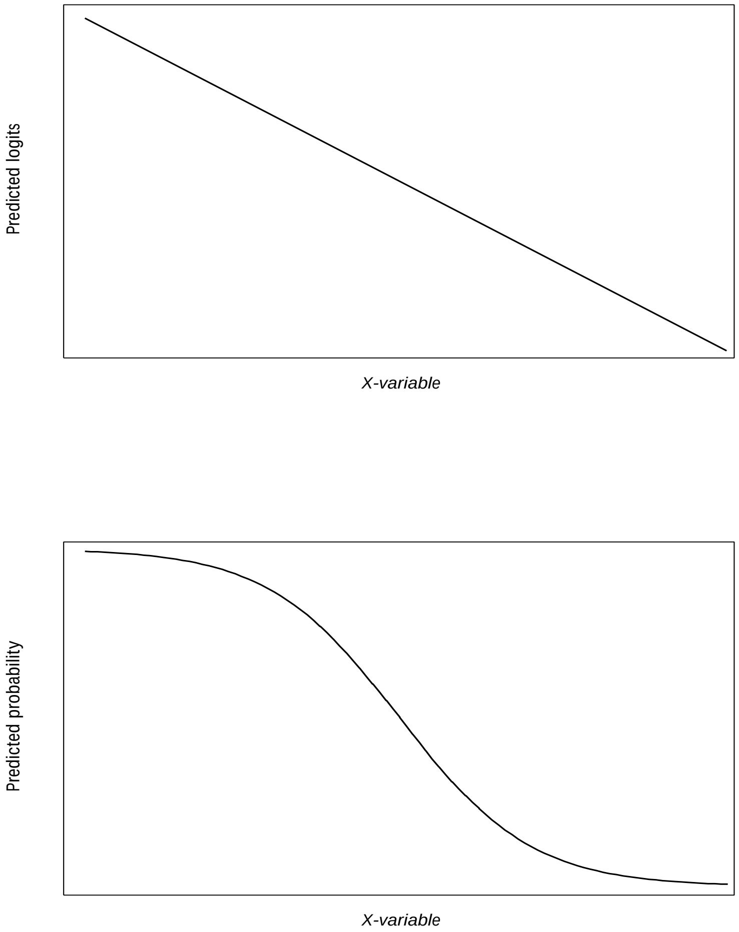


FIGURE 4 *Forms of the simple regression model on the logit and probability scales.*

characteristics and for which certain assumptions are valid. The first and most obvious characteristic of the data is that the response variable has only two responses of interest. This could include, for instance, whether trees attain a certain minimum height or whether brush reaches an undesirable level. It is possible to generalize logistic regression analysis for variables with three or more levels, but the required methods are not presented here (see McCullagh and Nelder [1989: Chapter 5], for instance).

Another important requirement, which can be ensured by appropriate study design, is that the response of each experimental unit must be independent of any other's response. For example, if the experimental unit is an individual tree and the fact that one dies implies that others must also die, then they are not responding independently. This might occur if a herbicide could travel through root grafts to kill neighbouring trees. This requirement does not preclude the possibility that many trees die because the probability of death is high. It also applies to sampling units. If experimental units are rows of 30 trees, then each tree within a row (the sampling unit) is also expected to respond independently and with constant probability.

Another requirement is that sample sizes must be known and fixed before the trial or experiment is conducted so that they will not be subject to random variation.[9] As well, both sampling and experimental unit responses are assumed to follow a binomial distribution. While this assumption is often made for binary response variables, it is not the only distribution available, nor does the data necessarily fit this distribution. When the number of sampling units for each experimental unit is large enough, a goodness-of-fit test is available to test if the data follow a binomial distribution. Table 5 outlines the different assumptions and requirements for fitting a simple linear regression model ($\alpha + \beta x$) to data with the normal and binomial stochastic components.

TABLE 5 *The assumptions and requirements for fitting a simple linear regression model ($\alpha + \beta x$) to data with either the normal distribution or the binomial distribution*

Stochastic component	
Normal distribution	Binomial distribution
Assumptions that apply to both models:	
Experimental and sampling units are randomly selected from an appropriate (and clearly defined) population to which results are to be generalized.	
Treatments are randomly assigned to experimental units.	
Responses are independent.	
Values of any independent variable (x) are chosen and known without error.	
The form of the model is correct.	
Differences between normal and binomial models:	
No restrictions on the values of the response variable.	Response counts (proportions) must be between zero and m (one).
Responses are normally distributed.	Response counts are binomially distributed.
Responses at a specific x-value have the same mean.	The probability of success at a specific x-value is constant.
Responses at any x-value have the same variance.	Response counts at any x-value have the binomial variance of $m\pi(1 - \pi)$, which depends on x, if π depends on x.

Basic principles of good study design remain the same regardless of the models used. One of the current challenges with standard logistic regression is that all treatments or factors in the model are assumed to have fixed levels—that is, the factor levels are specifically chosen by the experimenter and any inferences made to levels outside of the experimental situation are not intended. This is currently an area of active statistical research (see for example, Follmann and Lambert [1989], Breslow and Clayton [1993], and Zackin, et al. [1996]), but standard methods are not

[9] An advanced approach to the problem of random sample sizes is to condition on the observed sample sizes. The methodology for this approach is not presented here.

yet available. Therefore, it is currently unclear how to handle random factors such as seedlots, clones, or blocks in randomized block designs.

3 TESTING AND FITTING PROCEDURES

This chapter outlines methods of fitting logistic regression models to data, testing their fit and the usefulness of various terms in the deterministic component. Potential problems associated with testing and fitting are discussed. Two procedures in SAS (`PROC`'s `CATMOD` and `LOGISTIC`) that test and fit models are briefly introduced. Their use is more fully explored in Chapters 5, 6, and 7, where data for the examples of Chapter 4 are analyzed. The discussion in this section is necessarily terse and applied. More in-depth sources will be referred to in the following discussion.

3.1 Maximum Likelihood (ML)

Likelihood is a simple concept that motivates the maximum likelihood method of fitting and testing models. For discrete data, it is derived from a probability function, such as the binomial distribution, that predicts the probability of obtaining specific data values given known values of the parameters. In general, we do not know the values of the parameters, but we do have data. Thus, from a data analysis point of view, it makes sense to use this function to determine the likely values of the parameters. When we use the function in this way it is called the *likelihood*. The likelihood function for binomially distributed data is derived in the next paragraph and can be skipped by those less interested in the mathematical aspects.

For a specific observation y, the binomial probability function is:[10]

$$\mathrm{P}\ (y\,|\,m,\ \pi) = \begin{bmatrix} m \\ y \end{bmatrix} \pi^{y}\ (1 - \pi)^{(m-y)} \tag{3}$$

where the left side represents the probability of observing y given ($|$) the values of m and π. This equation can be used to calculate the probability of observing any y for one group, given known values of m and π. Some example values are plotted in Figure 2. The corresponding likelihood for one observation of y is:

$$\mathrm{L}\ (\pi\,|\,m,\ y) = \begin{bmatrix} m \\ y \end{bmatrix} \pi^{y}\ (1 - \pi)^{(m-y)} \tag{4}$$

where the left side represents the likelihood of observing π given values of m and y. The log-likelihood $l\,(\pi|\,m,\,y) = \log[L\,(\pi|\,m,\,y)] =$ $\log[\pi^{y}(1 - \pi_{y})^{m-y}] + \log\begin{bmatrix} m \\ y \end{bmatrix}$ and because the last term is constant

[10] In this equation, $\begin{bmatrix} m \\ y \end{bmatrix} = \dfrac{m!}{y!(m - y)!}$ and is called the binomial coefficient. It represents the number of combinations of y items taken from a total set of m items. For example, $\begin{bmatrix} 5 \\ 2 \end{bmatrix} = \dfrac{5!}{2!3!} = \dfrac{5\cdot4\cdot3\cdot2\cdot1}{2\cdot1\cdot3\cdot2\cdot1} = \dfrac{5\cdot4}{2\cdot1} = 10.$

when m is fixed, fitting procedures work with $l(\pi | m, y) = y \log(\pi) + (m - y) \log(1 - \pi)$. This is used during calculations since it is easier to work with and has useful statistical properties. The maximum likelihood estimate is the value of π that maximizes $L(\pi | m, y)$, or maximizes $l(\pi | m, y)$, or equivalently, minimizes $-2\log[L(\pi | m, y)]$ (referred to as –2LogL).

Some example values for various values of y and π are presented in Tables 6 and 7. The values of Table 7 are plotted in Figure 5. The maximum likelihood estimate or most likely value of π for a given y can be obtained from the table or figures. For a sample size of five and an observed count of $y = 0$, the maximum likelihood estimate of π, denoted by $\hat{p}$, is zero. When the observed count is one, $\hat{p} = 0.20$ and for $y = 3$, $\hat{p} = 0.60$. This concept of maximum likelihood analysis is extended further in sections 4 and 7.2 for data from several groups (a one-way classification study).

TABLE 6 *Probability of observing counts (y) given a sample size (m = 5) and success probabilities π = 0.1, 0.3, and 0.5*

Observed count	Known success probabilities		
y	$\pi = 0.1$	$\pi = 0.3$	$\pi = 0.5$
0	0.59	0.17	0.03
1	0.33	0.36	0.16
2	0.07	0.31	0.31
3	0.01	0.13	0.31
4	0.00	0.03	0.16
5	0.00	0.00	0.03

TABLE 7 *Likelihood (L[π]) and log-likelihood (l[π]) values for a fixed sample size (m = 5) and observed counts (y = 0, 1, and 3)*

Probability	Observed count					
value	$y = 0$		$y = 1$		$y = 3$	
π	$L(\pi)$	$l(\pi)$	$L(\pi)$	$l(\pi)$	$L(\pi)$	$l(\pi)$
0.0	[a]	[a]	0.00	[a]	0.00	[a]
0.005	0.98	–0.03	0.02	–3.71	0.00	–13.60
0.05	0.77	–0.26	0.20	–1.59	0.00	–6.79
0.1	0.59	–0.53	0.33	–1.11	0.01	–4.82
0.2	0.33	–1.12	0.41	–0.89	0.05	–2.97
0.3	0.17	–1.78	0.36	–1.02	0.13	–2.02
0.4	0.08	–2.55	0.26	–1.35	0.23	–1.47
0.5	0.03	–3.47	0.16	–1.86	0.31	–1.16
0.6	0.01	–4.58	0.08	–2.57	0.35	–1.06
0.7	0.00	–6.02	0.03	–3.56	0.31	–1.18
0.8	0.00	–8.05	0.01	–5.05	0.20	–1.59
0.9	0.00	–11.51	0.00	–7.71	0.07	–2.62
1.0	0.00	[a]	0.00	[a]	0.00	[a]

[a] Value undefined.

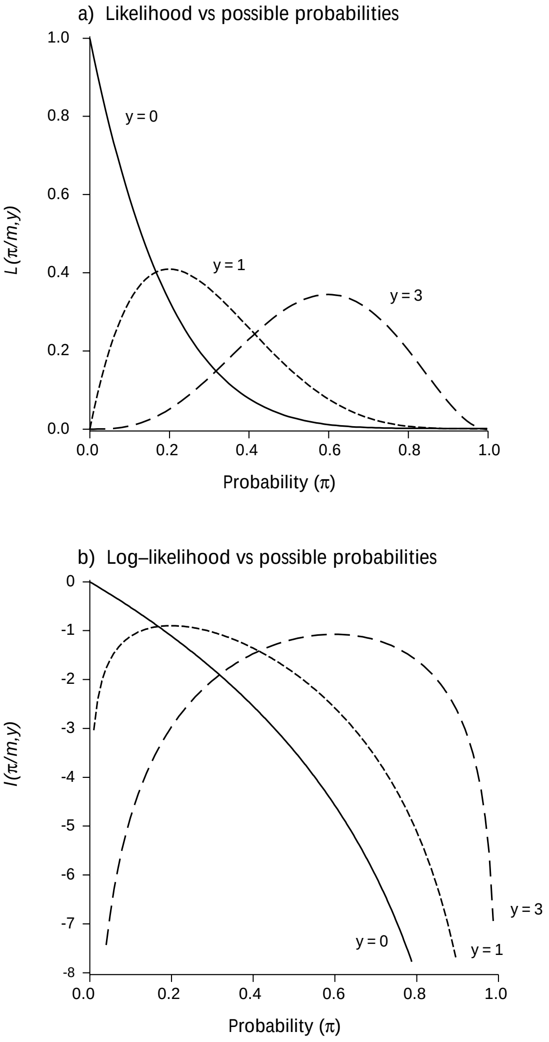


FIGURE 5 *Likelihood (L[π]), and Loglikelihood (l[π]) for a binomial sample of 5 and observed counts of 0, 1, and 3.*

3.2 The Logic of the Fitting and Testing Process

The logic used to test a model's independent variables or factors depends on comparing how different models fit a set of data. One measure of the fit is calculated by the likelihood, or more usefully by –2LogL. For general linear models, this function is provided by the residual sums of squares using the extra sums of squares principle.[11]

It is useful to divide possible models into three types—saturated, restricted, and the simplest model, the mean (also known as the intercept-only model). The saturated and mean models provide helpful extremes against which to compare various restricted or "ordinary" models. A model is *saturated* when it contains as many parameters as there are experimental units. This is the largest number of parameters any model

[11] See "Pictures of Linear Models" (Bergerud 1991) for another discussion of this topic.

can have. Although this model provides no simplification, it does have the largest likelihood any model can have and has the best fit. On the other hand, the simplest model, the mean or intercept-only model, has only one parameter and the smallest likelihood that a model can have.

All models, other than the saturated model, are called *restricted* because they provide some simplification when compared to the saturated model. Some relationships are postulated between the experimental units so that the number of parameters is reduced from that of the saturated model. For instance, the simple regression model used in the caribou calf survival example (section 4.1) has only two parameters, although the saturated model would have as many parameters as caribou herds included in this study. The model for the fertilizer study (section 4.3) has four parameters, while the saturated model would have 24, reflecting the 24 experimental units in this study. The mean model is the most restricted model because only one parameter is used for all the experimental units. Contrasts or planned compensations are restricted models with very specific patterns postulated for the parameters.

A restricted model that fits the data better than the simplest model, but also fits the data *almost* as well as the saturated model, is suitable from a statistical perspective. However, several restricted models may fit reasonably well. In this case, final decisions about the most suitable model must be based on the subject investigated.

3.3 Fitting Procedures

The likelihood of a model is obtained by fitting the model to the data. In general, this involves an iterative numerical search. This means that an initial guess is made about the parameters of the model, which the fitting procedure then uses to calculate better estimates. These are, in turn, fitted to the data. This process is repeated until some criterion is met which indicates an adequate fit. An obvious criterion would be to stop when the parameter estimates change very little from one iteration to the next. Problems can arise when a probability is zero or one because the corresponding logit value will be unbounded (i.e., negative or positive infinity) and the estimates never converge even though an adequate fit is obtained. Therefore, the change in the residual sums of squares or likelihood is used instead.

The Newton-Raphson method and the method of scoring are often used as numerical search techniques and are well described in Dobson (1983: 30–33). The likelihood for a simple one-way classification can be calculated by hand and is developed in Section 7.2. Straightforward discussions of maximum likelihood can be found in Gilchrist (1984) and Wetherill (1981). Other information about fitting methods can be found in Bishop et al. (1975), McCullagh and Nelder (1983 or 1989), and Hosmer and Lemeshow (1989).

3.4 Criteria For Assessing Fit

Three common criteria or measures of fit are used to compare models. The first one is –2LogL, which is often the convergence criterion during the fitting procedure. If the model has many variables (parameters) compared to the number of observations available, –2LogL can be smaller than if it was calculated with a sufficient amount of data. To account for this effect, –2LogL can be adjusted in two ways:

- Akaike Information Criterion: AIC = –2LogL + 2(1 + number of explanatory variables)
- Schwartz Criterion: SC = –2LogL + (1 + number of explanatory variables) × log(total number of sampling units)

Unlike the coefficient of determinations (R^2 and the adjusted R^2) of multiple regression, these variations of –2LogL are not constrained between zero and one. Because the absolute values of these criteria do not have any meaning, it is the difference between values from different models that is of interest. If all three, –2LogL, AIC, and SC are similar in size, then the sample size is adequate for the model being fit. The AIC and SC can be used to compare the fit of models that are not nested in one another. A generalized coefficient of determination has been added to version 6.11 of SAS, and might also be used in analysis.

3.5 Using Deviance to Compare Likelihoods

Each model that is fitted to the same set of data has a corresponding log-likelihood value that is calculated at the maximum likelihood estimates for that model. These values are used to compare and statistically test terms in the models. In general, suppose that model one has t parameters, while model two is a subset of model one with only r of the t parameters so that $r < t$. Model one will have a higher log-likelihood than model two. For large sample sizes, the difference between these two likelihoods, when multiplied by two, will behave like the chi-square distribution with $t - r$ degrees of freedom. This can be used to test the null hypothesis that the $t - r$ parameters that are not in both models are zero.

Computer printouts produce either the log-likelihoods (`LogL` are negative values) or `-2LogL` (which are positive values). These values can be used directly to calculate the differences for statistical tests. Differences between -2LogL's are called deviances, where:[12]

$$\begin{aligned} D &= -2[l(\text{model 2}) - l(\text{model 1})] \\ &= -2\text{LogL}(\text{model 2}) - -2\text{LogL}(\text{model 1}), \end{aligned} \tag{5}$$

which under certain conditions approximately follows a chi-square distribution with $t - r$ degrees of freedom. A deviance test labelled the `Likelihood Ratio` is printed at the bottom of the `Maximum Likelihood Analysis of Variance Table` output by `PROC CATMOD`. This test is the difference in deviance between the saturated model (model one with t parameters) and the restricted model which was just fit (model two with r parameters). Asymptotically (i.e., with large t), it is a goodness-of-fit test (with $df = t - r$) for the null hypothesis that the data are binomially distributed. `PROC LOGISTIC` also outputs this deviance test value under the column titled `Intercept and Covariates`, but does not provide a p-value for it. As well, `PROC LOGISTIC` produces a reliable test for the explanatory variables as a group (labelled `Chi-square for covariates`). This test compares the deviances between the simplest,

[12] The deviance is a likelihood ratio statistic.

intercept-only model, with the current restricted model that has just been fit to the data. In this case, model one is the restricted model with *r* parameters, while model two has just one parameter. Thus, this deviance test has $df = r - 1$.

Differences between deviances are also deviances which follow an approximate chi-square distribution, if both deviances were calculated on the same set of data and both either used the saturated model or the intercept-only model (and sample sizes are sufficiently large). Thus, differences between the Likelihood ratio output by `PROC CATMOD`, or the −2LogL output by `PROC LOGISTIC`, for restricted models nested in one another are also deviances with a chi-square distribution.

3.6 Wald Statistics

Asymptotically, each estimated parameter of a fitted model will have a normal distribution. Thus, each parameter can be tested with the simple *t*-test. Most computer programs will square the *t*-value (the ratio of the parameter estimate divided by its estimated standard error) and output it as a chi-square value (known as a Wald test). This statistic has an approximate chi-square distribution with one degree of freedom. Wald tests are particularly helpful when deciding which variables or terms should be dropped from the model at hand. They are considered a “last-in” test; that is, they test whether the current term, if it was the last term added to the model, would substantially reduce the −2LogL. However, Wald statistics are considered approximate and somewhat unreliable. Therefore, marginally significant results should be confirmed by fitting models with and without the terms of interest, and then conducting the corresponding deviance test. Problems can also arise with Wald statistics when complete success or failure occurs in one or more of the experimental units. See Appendix 1 for an example.

3.7 Model Checking

When first looking at the data, it is useful to plot it on both the probability and the logit scale. A problem can arise with the logit if an experimental unit has either complete success or failure. In this case, the logit is undefined because the numerator or the denominator will be zero. If a small value is added to both the top and bottom of the logit, then the logit function becomes, for instance, $\log[(p + 0.01)/(1 - p + 0.01)]$. This is known as an empirical logit and can be used to plot the observed data, since there will be points for each experimental unit. Values other than 0.01 may also be used when calculating empirical logits.

The various statistical tests outlined in sections 3.5 and 3.6 describe ways to check the overall fit of models. When a reasonable model is selected on the basis of these overall measures of adequacy, the next step is to look at individual observations to determine if they appear to be well fitted by the model. As in multiple regression, examining the residuals (differences between the observed and fitted data points) and various influence statistics is desirable. Influence statistics can be produced by `PROC LOGISTIC`. More information about influence statistics can be found in standard texts on multiple regression and in, for example, Hosmer and Lemeshow (1989: section 5.3:149–170), McCullagh and Nelder (1989, Chapter 12), and Agresti (1996, section 5.3.4).

The simplest residuals to define are the observed minus the predicted, namely $y_i - m_i\hat{p}_i$, where $\hat{p}_i$ are the values fitted by the model. PROC CATMOD produces two of these residuals for each observation: one for success and one for failure, while PROC LOGISTIC produces only one residual per observation. Plots of residuals output by PROC CATMOD will have a characteristic "mirror" pattern about a horizontal line at zero if all the residuals are plotted. While these are easily understood, they have the disadvantage of non-constant variance (since the variance depends on the proportions, π_i). Two other residuals produced by PROC LOGISTIC are the Pearson residual:

$$\text{reschi}_i = \frac{(y_i - m_i\hat{p}_i)}{\sqrt{m_i\hat{p}_i(1-\hat{p}_i)}},$$

and the deviance residual

$$\text{resdev}_i = \pm\sqrt{2y_i\log\left\{\frac{y_i}{m_i\hat{p}_i}\right\} + 2(m_i - y_i)\log\left\{\frac{m_i - y_i}{m_i(1-\hat{p}_i)}\right\}},$$

where the sign is positive if y_i/m_i is greater than $\hat{p}_i$ and negative otherwise. Given that the fitted model is correct, with suitably large m_i, these two types of residuals will follow an approximately normal distribution with mean zero and constant variance. Thus, they can be plotted against the predicted values and the various explanatory variables to look for unusual patterns and data points. They can also be examined for normality with a normality test and cumulative probability plots.

3.8 Possible Problems and Pitfalls

Logistic regression methods are useful because they allow analysis of binomially distributed data without requiring the large experimental unit sizes (m_i) necessary for normal distribution approximation. Interestingly, current statistical tests require large enough samples sizes so that the observed statistics behave like normally distributed statistics (the chi-square distribution is based on the normal distribution). If sample sizes are sufficiently large, the deviance and the Wald tests can both be compared to the chi-square distribution to develop probability values for the observed statistics.

The sample size must be adequate so that the statistics will behave asymptotically. One way to achieve this is called m-asymptotics. This occurs when the number of sampling units (m_i) for each or most experimental units is large enough. The required sample size depends on π (see Table 2). When the m_i are small, the likelihood ratio test used to compare a restricted model to the saturated model (output by PROC CATMOD) is unreliable and should be ignored (McCullagh and Nelder, 1989:118–121). Residuals also do not behave as normally distributed values do and must be examined with caution. When the m_i are large enough, then the likelihood ratio test, the deviance tests, and the residuals behave as expected. Another way of ensuring an adequate sample size is called n-asymptotics. This occurs when $n = \Sigma m_i$ is reasonably large, even if the number of sampling units is small. While the goodness-of-fit likelihood ratio test is not well behaved in this case, the deviance statistic (equation 5) behaves as expected and is a reliable test statistic. It is also more reliable under a variety of conditions compared to the Wald tests and the goodness-of-fit likelihood ratio test.

The most common problem associated with fitting models occurs when an observed sample or experimental unit has complete failure or success. The SAS Institute (SAS Institute 1989, 2:463) recommends the use of the maximum likelihood method (the default fitting method) in this case. Most importantly, though, is that the Wald statistics will be unreliable. An example of this type of problem is described in Appendix 1. Study designs with at least 25–30 experimental/observational units and with less than 20% of the treatment combinations having five or fewer experimental units are desirable (SAS Institute 1989, 2:462).

4 WORKING WITH LOGISTIC REGRESSION: SOME FORESTRY EXAMPLES

This chapter shows how to use logistic regression analysis methods by focusing on some forestry examples. The objective, design, and variables for each study are described and specific logistic regression models are developed. Each example is then explored in more depth by examining specific data sets and fitting models. This is done without substantial reference to the programs that were used to produce the results. Nevertheless, many of the figures presented are printer plots output by SAS. Chapter 5 provides a detailed discussion of the SAS programs used and can be referred to by the interested reader.

While sample sizes are kept small for illustrative purposes, some of these studies might not be powerful enough to detect effect sizes of interest. This could be determined using power analysis, a topic that is beyond the scope of this handbook (see Agresti [1996:130–131]).

4.1 Simple Regression: Caribou and Wolf Predation

The objective of this example is to study the relationship between wolf presence and the survival of caribou calves during their first year. It is hypothesized that fewer calves will survive if high numbers of wolves are in their vicinity.

This trial might be designed as follows. Several separate caribou herds are selected and a random sample of calves from each herd are radio collared in the spring soon after birth. The following winter a wolf census is conducted for each herd and in the following spring all radio-collars are located to determine the number of surviving calves (if calves have been killed, the collar can still be found). For this investigation to be successful, it is important that:

- wolf numbers range from low to high, so that any effects will be as large as possible, making them easier to identify;
- wolves have only one herd available for predation, to ensure independence of herd responses;
- the proportion of collared calves is small for each herd;
- the wolf census numbers have been determined with little error; and

- the effect of other predators is small and can be neglected, or alternatively, that their effect is constant for all the herds in the study and will not interact with wolf predation.

The data set will have the following variables:[13]

- j = herd number, $j = 1, 2, \ldots k$ (where there are k herds in all)
- m_j = number of collared calves in herd j
- y_j = number of surviving collared calves in herd j
- w_j = estimated number of wolves living in the vicinity of herd j
- $p_j = y_j/m_j$ = observed proportion of calf survival for herd j

A simple regression might be an appropriate model

$$\text{logit}\ (\pi_{\text{j}}) = \mu + \beta w_j, \tag{6}$$

where π_j = probability that a calf survives, if w_j wolves live in the vicinity of the herd, and the number of surviving calves is $y_j \sim \text{binomial}(\pi_j, m_j)$.[14]

Since the hypothesized relationship is of decreasing calf survival with greater wolf presence, it is expected that β will be negative. The general form of this model is shown in Figure 6. The line is straight on the logit scale, but has the characteristic S-shaped curve on the probability scale. The data (y_j/m_j versus w_j) should be graphed to determine if this model is appropriate.

4.1.1 Data collection For this study, a wildlife biologist collects data from the published and unpublished work of other scientists to generate an extensive data set on caribou herds scattered throughout the northern hemisphere. Suitable information exists for the survival of collared calves in nine herds during their first summer of life. A reliable estimate of wolf presence is also available for the following winter. The presence of other predators (e.g., grizzly bear and lynx) in the vicinity of these herds during this time is estimated to be low and therefore need not be included in the models. The herds are widely separated in space, so that no wolf pack has access to more that one of the nine herds.

The data are shown in Table 8 and the observed proportions are plotted against the wolf density estimates in Figure 7. It is apparent that the proportion of surviving calves decreases with increasing wolf density. This is an observational study because wolf density is an observed variable which is not controlled by the researcher. Therefore, statements of cause and effect are not valid on either statistical or study design grounds. If the levels of wolf density were under the control of the researcher and were randomly assigned to each herd, then the results could be clearly discussed in cause and effect terms.

[13] Note that for all of the example studies, the subscript j is used to number experimental units, while i is used for treatment levels.

[14] The symbol ~ means distributed as. So $y_j \sim \text{binomial}(\pi_j, m_j)$ means that y_j has a binomial distribution with parameters π_j for probability of success and m_j for sample size.

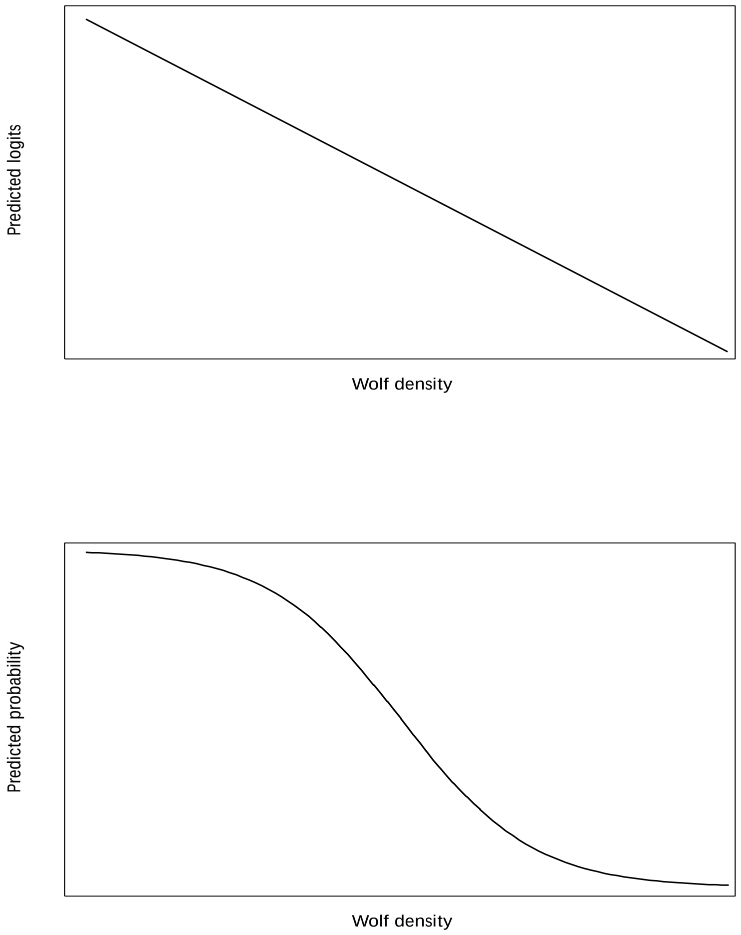


FIGURE 6 *Forms of the simple regression model on the logit and probability scales.*

TABLE 8 *Listing of the data for the caribou calf survival example*

Herd number	Wolf density (no. per 1000 km^2)	Total number	Alive	Proportion alive
1	9	15	14	0.933
2	10	7	7	1.000
3	12	4	3	0.750
4	13	5	5	1.000
5	15	10	9	0.900
6	23	10	9	0.900
7	31	15	9	0.600
8	34	13	4	0.308
9	38	13	1	0.077

This data is first analyzed using traditional simple and quadratic regression (section 4.1.2). Then a logistic regression model is fitted in section 4.1.3, and these two approaches are contrasted in section 4.1.4. The fit of the logistic regression model is discussed in section 4.1.5, while the odds ratio is discussed in section 4.1.6. SAS programs including detailed output

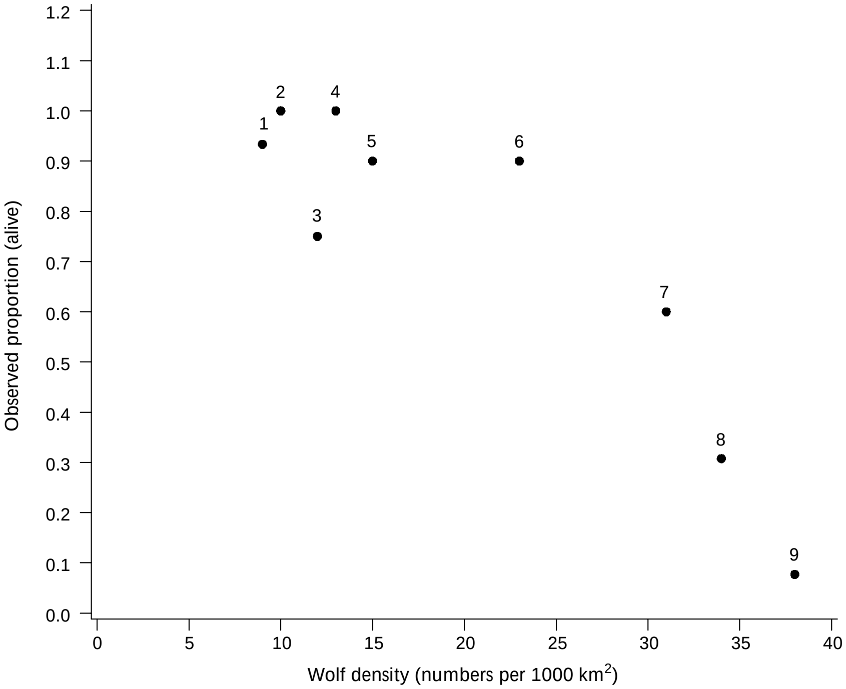


FIGURE 7 *Plot of observed proportions against estimated wolf density (numbers represent herd number).*

and comparison of the CATMOD and LOGISTIC procedures are presented in section 5.2.

4.1.2 Analysis of simple and quadratic regression models Judging from Table 8 and Figure 7, a regression of survival probability against wolf density is likely a reasonable model. A quadratic fit should also be considered given the apparent curve to the data in Figure 7. We could start the analysis by looking at a regression of the proportion of surviving calves against both the number of wolves (wolf) and the squared number of wolves (wolf^2). The fit of the simple and quadratic models is compared in the following table:

Model	*df*	Sums of squares	Mean square	*F*-value	*p*-value	Adjusted R^2
Linear	1	0.69095	0.69095	27.7	0.0012	0.769
Quadratic	2	0.80556	0.40278	40.0	0.0003	0.907

The simple regression model shows a strong and statistically significant correlation between the proportion of calves that survive the summer and wolf density (adjusted $R^2 = 0.77$ with $F = 27.6$ and p-value = 0.0012). The fit is improved though, when a quadratic model is used (adjusted $R^2 = 0.91$ with $F = 40.0$ and p-value = 0.0003).

4.1.3 Logistic regression analysis The results of fitting a simple logistic regression with wolf as the independent variable is shown in the following table:

		Model	
	Intercept	Intercept and	Chi-square
Criterion	only	covariates	for covariates
Akaike Information Criterion (AIC)	119.575	79.920	—
Schwartz Criterion (SC)	122.097	84.963	—
Log-likelihood	117.575	75.920	41.656[a]

[a] With 1 *df* ($p = 0.001$).

The log-likelihood results are the final values from the iterative fitting process. The first log-likelihood value is for an intercept-only model (i.e., a model with only one mean for all the data), while the second is for the full model, which includes the intercept and model variables. The difference between the first two values provides an overall test for the significance of the variables (called covariates here) in the model. For this example, the value is $\chi^2 = 41.7$ with one degree of freedom and p-value ≤ 0.0001. This suggests that calf survival is correlated with the estimated wolf density.

The next table provides parameter estimates, their standard error, and Wald tests for the intercept and variable parameter estimates in the model:

Variable	*df*	Parameter estimate	Standard error	Wald χ^2	p-value	Odds ratio
Intercept	1	5.0993	1.0677	22.8	0.0001	165.9
Wolf	1	–0.1693	0.0352	23.1	0.0001	0.844

We can use the parameter estimates to specify the logistic regression model fitted to the data by first defining the logit that corresponds to the probability of an event; in this case, the event that a caribou calf does survive is estimated to be:

$$\text{logit}(\text{probability of surviving}) = 5.10 - 0.169 \times \text{Wolf}$$

Therefore, the estimated probabilities of surviving or not are given by (see equation 2):

$$\text{probability of surviving, } \hat{p} = \exp(\text{logit})/[1 + \exp(\text{logit})]$$
$$\text{probability of not surviving, } 1 - \hat{p} = 1/[1 + \exp(\text{logit})].$$

To illustrate the use of these equations, let us calculate both probabilities for a wolf density of 10 wolves per 1000 km^2. First, the estimated logit of the survival probability $= 5.10 - 0.169 \times 10 = 3.41$. Therefore, the estimated survival probability is $\exp(3.41)/[1 + \exp(3.41)] = 0.968$, while the estimated mortality probability is $1/[1 + \exp(3.41)] = 0.032$ or $1 - 0.968$. The estimated probabilities for other wolf densities are shown in Table 9.

TABLE 9 *Estimated probabilities of survival and mortality for different wolf densities*

		Probability of	
Wolf density	Logit	Surviving	Not surviving
0	5.10	0.994	0.006
10	3.41	0.968	0.032
20	1.72	0.85	0.15
30	0.03	0.51	0.49
40	–1.66	0.16	0.84

4.1.4 Comparing the regressions with the logistic regression It is interesting to compare the results of the simple regression analysis of the survival proportions with that of the logistic regression. A simple approach is to compare the predicted calf survival proportions for each model with the observed data by looking at the residual sums of squares for the three models:

Model	Uncorrected sums of squares
Linear regression	0.175
Quadratic regression	0.060
Logistic regression	0.103

While the quadratic regression has the smallest residual sums of squares, it is not much smaller than that of the logistic regression, which uses only two parameters to the three of the quadratic model.

The predicted values are plotted in Figure 8. This figure shows that the logistic regression fits the data better than did the linear regression. It should be noted that one of the predicted values from the regression was 1.02 implying that more calves survived than existed! On the other hand, the quadratic regression was a marked improvement on the linear regression and it had a smaller residual sums of squares than the logistic regression. However, note that the fitted values for the quadratic regression at first increase with wolf density and then decrease; a pattern that we might not expect nor believe is reasonable. In summary, the logistic regression is a better model for this data than either the linear and quadratic regressions because:

1. it fits the data adequately with only two parameters compared to the three parameters of the quadratic fit;
2. it has the expected shape; and
3. it produces no predicted probabilities that are greater than one or less than zero.

4.1.5 Examining the fit of the logistic regression model The adequacy of the fit of the logistic regression should be further examined by various plots. The printer plot of predicted values versus observed values (see Figure 9) is reasonably straight with a slope near one (hand drawn on the

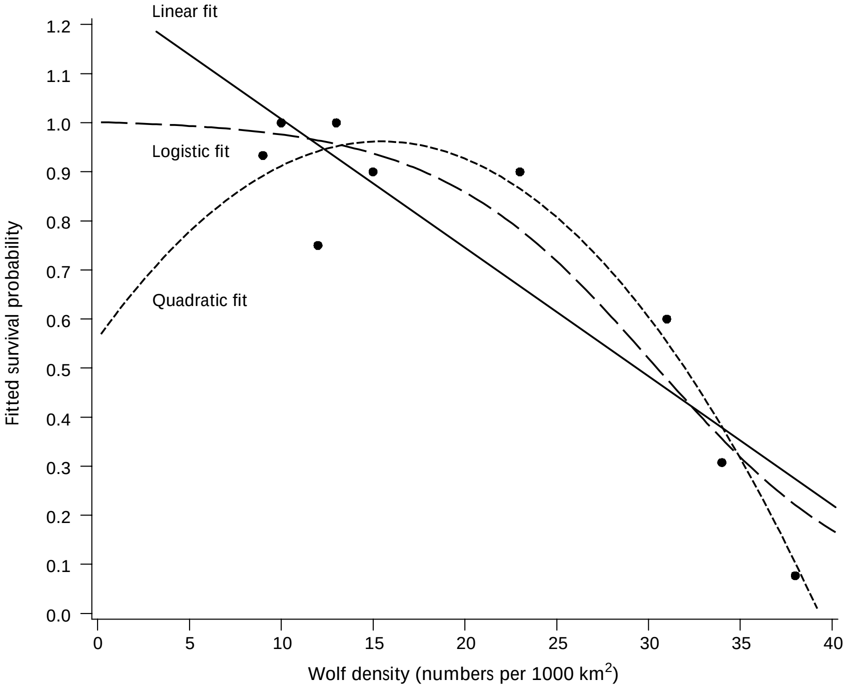


FIGURE 8 *Three predicted curves with observed survival proportions.*

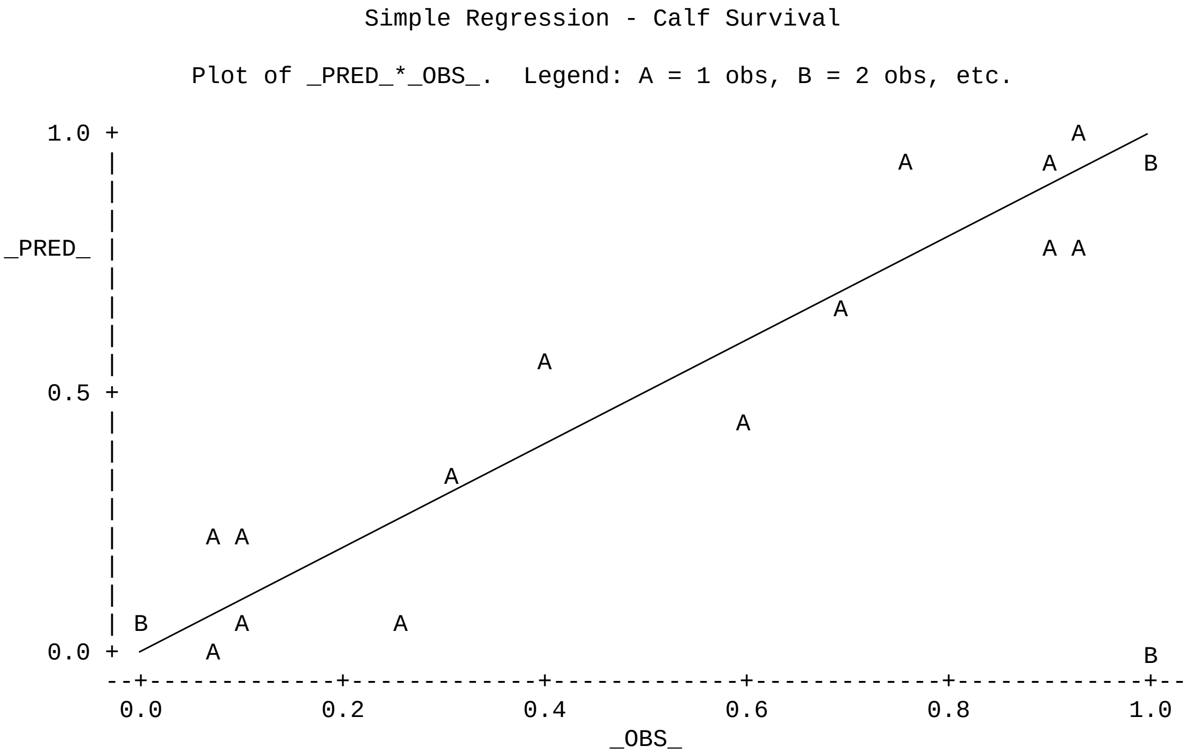


FIGURE 9 *Printer plot of the predicted against the observed proportions for each herd. (Lines are hand-drawn.)*

figure). A slope of one is expected since the predicted values should be about the same as the observed values. The printer plots of the residuals (Figures 10, 11, 12, and 13) show little pattern. Thus, it is reasonable to conclude that the logistic regression fits the data adequately.

4.1.6 Interpreting the odds ratio The odds ratio for survival is exp(parameter estimate for wolf); that is the odds ratio, $\psi = \exp(\beta)$, where β is the parameter estimate for wolf (recall that the logistic model is $\log[\pi/(1 - \pi)] = \alpha + \beta$ wolf). This odds ratio is interpreted as the change in odds of caribou calf survival when the wolf numbers are increased by one. The odds ratio in this example is $\psi = 0.844$. As this is less than one, it implies that the chance of calf survival decreases with increasing wolf numbers. The change in the odds of calf survival for any change in wolf numbers (Δw) is $\psi^{\Delta w}$. Therefore, the change in log-odds is $\Delta w \log(\psi) = \Delta w \beta$. For example, an increase of ten wolves implies that $\psi = 0.844^{10} = 0.184$ (i.e., less chance of survival), while a decrease of ten wolves implies that $\psi = 0.844^{-10} = 5.45$ (i.e., greater chance of survival). Recall that the odds ratio is a ratio of logits, not probabilities. Thus, a constant change in the logits does not mean a constant change in the probabilities. For more discussion of log-odds, refer to section 4.5.2.

4.2 One-way Classification Study: Survival and Age of Stands with Root Rot

The objective of this example is to study the effect of stand age on the survival of trees within stands that are infected with root rot. The researcher examined an inventory database to find suitable even-aged stands that might be considered “similar” based on selected attributes, such as ecosystem classification and infestation level of root rot of different stand ages.

```
          Plot of _RESID_*_PRED_.  Legend: A = 1 obs, B = 2 obs, etc.

_RESID_ |
    0.2 +   A
        |
        |                            A                        AA
        |
        |
        |                                                                 A
        |  A  A                                        A                     A
    0.0 +-----------------------------------------------------------------------
        |  A                     A                                        A  A
        |    A
        |
        |
        |               AA                     A
        |
   -0.2 +                                                                   A
        -+-------------+-------------+-------------+-------------+-------------+-
        0.0           0.2           0.4           0.6           0.8           1.0

                                        _PRED_
```

FIGURE 10 *Printer plot of the simple residuals against the predicted proportions for each herd.*

```
             Plot of _RESID_*HERD.  Legend: A = 1 obs, B = 2 obs, etc.

    0.2 +                  A
        |
        |
        |                                             A        A                 A
_RESID_ |
        |
        |
        |A                          A                                   A
        |         A                          A
    0.0 +-------------------------------------------------------------------------
        |         A                          A
        |A                          A                                   A
        |
        |
        |
        |                                             A        A                 A
        |
        |
   -0.2 +                  A
        -+--------+--------+--------+--------+--------+--------+--------+--------+
         1        2        3        4        5        6        7        8        9
                                        Herd Number


                         Simple Regression - Calf Survival
                                       Plots

             Plot of _RESID_*WOLF.  Legend: A = 1 obs, B = 2 obs, etc.

    0.2 +               A
        |
        |
        |                                    A                A            A
_RESID_ |
        |
        |
        |         A       A                                         A
        |           A         A
    0.0 +-------------------------------------------------------------------------
        |           A         A
        |         A       A                                         A
        |
        |
        |
        |                                    A                A            A
        |
        |
   -0.2 +               A
        --+---------+---------+---------+---------+---------+---------+---------+-
          5        10        15        20        25        30        35        40
                            Wolf Density (Numbers per 1000 km^2)
```

FIGURE 11 *Printer plots of simple residuals to check fit of the logistic regression model.*

```
       Plot of RESDEV*HERD.  Legend: A = 1 obs, B = 2 obs, etc.

Deviance Residuals

  2 +
   |
   |
   |
   |                                          A       A
   |
   |          A               A
   |
   |
  0 +---------------------------------------------------------------------
   |                                  A                       A
   |
   |
   |  A
   |
   |                  A                                               A
   |
   |
 -2 +
   ---+-------+-------+-------+-------+-------+-------+-------+-------+--
      1       2       3       4       5       6       7       8       9

                                 Herd Number
```

```
       Plot of RESDEV*WOLF.  Legend: A = 1 obs, B = 2 obs, etc.

Deviance Residuals

  2 +
   |
   |
   |
   |                                  A              A
   |
   |           A    A
   |
   |
  0 +--------------------------------------------------------------------
   |                    A                                 A
   |
   |
   |         A
   |
   |               A                                             A
   |
   |
 -2 +
   ---+--------+--------+--------+--------+--------+--------+--------+--
      5       10       15       20       25       30       35       40
```

Wolf Density (Numbers per 1000 km^2)

FIGURE 12 *Printer plots of deviance residuals against herd number and wolf density.*

Pearson Residuals

```
       Plot of RESCHI*HERD.  Legend: A = 1 obs, B = 2 obs, etc.

 1 +                                          A       A
   |
   |
   |          A               A
   |
   |
 0 +---------------------------------------------------------------------
   |
   |                                  A                       A
   |
   |
   |
-1 +  A
   |                                                                  A
   |
   |
   |
   |
-2 +                  A
   ---+-------+-------+-------+-------+-------+-------+-------+-------+--
      1       2       3       4       5       6       7       8       9

                                 Herd Number
```

Pearson Residuals

```
       Plot of RESCHI*WOLF.  Legend: A = 1 obs, B = 2 obs, etc.

 1 +                                  A              A
   |
   |
   |           A    A
   |
   |
 0 +--------------------------------------------------------------------
   |
   |                 A                             A
   |
   |
   |
-1 +         A
   |                                                             A
   |
   |
   |
   |
-2 +               A
   ---+--------+--------+--------+--------+--------+--------+--------+--
      5       10       15       20       25       30       35       40
```

Wolf Density (Numbers per 1000 km^2)

FIGURE 13 *Printer plots of Pearson residuals against herd number and wolf density.*

Serious thought is required to determine similar root rot levels; that is, individual trees must have had the same exposure to root rot. One option is to estimate levels of root rot when all stands were young, regardless of current age. Suitable stands should be far enough apart physically that independent response is a reasonable assumption. The stands are surveyed and trees sampled in a simple random manner to estimate the proportion of trees which had died. Since this is an observational study, a correlation between stand age and mortality can be inferred, but any inferences regarding cause and effect must be made on non-statistical grounds.

The data set will have the following variables:

- j = stand number, $j = 1, 2, \ldots, k$ (where there are k stands in all)
- i = stand age, $i = 1, 2, 3$
- y_{ij} = number of surviving trees in stand j at age i
- m_{ij} = number of trees sampled in stand j at age i
- $p_{ij} = y_{ij}/m_{ij}$ = observed proportion of surviving trees in sample from stand j at age i.

An appropriate model for this one-way classification study is:

$$\text{logit}\ (\pi_{ij}) = \mu + \alpha_i \tag{7}$$

where π_{ij} = probability that a tree in stand j of age i is alive, with the number of surviving trees, $y_{ij} \sim \text{binomial}(\pi_{ij}, \mathrm{m}_{ij})$, and $i = 1, 2,$ and 3, for the three stand ages. The parameters α_i indicate the additive effect of stand age i. They are restricted to sum to zero so that μ is the average response of the experiment on the logit scale.

4.2.1 Data for the one-way classification study Trees are sampled from nine stands found suitable for this hypothetical study. The results are shown in Table 10.

Each stand is identified as either young, middle-aged, or old, and values of other variables are "matched" so that they are similar for these stands

TABLE 10 *Data for trees sampled for root rot study*

Stand number	Stand age	Total trees	No. dead trees	No. live trees	Percent dead
1	Young	41	13	28	32
2	Young	35	8	27	23
3	Middle	28	10	18	36
4	Middle	37	15	22	41
5	Middle	16	6	10	38
6	Old	16	7	9	44
7	Old	18	7	11	39
8	Old	41	19	22	46
9	Old	15	7	8	47
Total		247	92	155	37

and need not be included in the modelling process. Three basic models will be fitted to this data:

1. saturated model, where each experimental unit (stand) is assumed to have different survival probabilities;
2. three groups model, where stands are grouped by age into three groups and each group is assumed to have different survival probabilities; and
3. one group (or intercept-only) model, where all the stands are assumed to have the same survival probability.

Comparisons between these increasingly simple models will answer the questions:

- Are the responses between stands within the three groups reasonably homogeneous? If so, this would mean that the three groups model provides a good fit to the data.
- Are the responses between the three groups similar enough that a one group model provides a good fit? In this case, only one survival probability would be needed for all nine stands implying that there is little evidence of a relationship with age.

The analysis examining the three group model is described in section 4.2.2. The differences between the three age groups is examined using contrasts in section 4.2.3. Assuming that a two group model (i.e., no difference between old and middle-aged stands) is appropriate for our objectives, we will study its fit in section 4.2.4.

4.2.2 Logistic regression analysis of three group model The first model fitted to the data is the three groups model. The Wald statistics and likelihood ratio are:

Source	*df*	χ^2	*p*-value
Intercept	1	16.98	0.0000
Age	2	4.96	0.0836
Likelihood ratio	6	1.23	0.9755

The log-likelihoods and –2LogL for various models are summarized in Table 11. The three group model is compared to the saturated model by examining the likelihood ratio (which is calculated by taking the difference in –2LogL's of the two models: 319.8 – 321.0 = 1.2, *df* = 6). When the sample size is adequate, this also provides a goodness-of-fit test for the model to the data. However, in this case, the sample size is not adequate.

The Wald statistic for age (χ^2 = 4.96, *p*-value = 0.084) provides weak evidence against the null hypothesis of no age effect. A more accurate overall test can be calculated by comparing the –2LogL with that of a one group model (for all nine stands). The resulting deviance is 326.2 – 321.0 = 5.1 with two degrees of freedom. This has a *p*-value between 0.05 and

TABLE 11 *The log-likelihoods of four different models*

Model	Number of parameters	Log-likelihood	–2LogL	Difference (*df*)	
1. Saturated	9	–159.9054	319.8108	1.228 (6)	6.356 (8)
2. Three groups	3	–160.5196	321.0392	5.128 (2)	
3. One group	1	–163.0835	326.1670		
4. Two groups	2	–160.8441	321.7092		

0.10, similar to the previous value of 0.084. Thus, the two tests arrive at the same conclusion.

The parameter estimates for the three group model are:

Effect	Parameter	Estimate	Standard error	χ^2	*p*-value
Intercept	1	–0.5547	0.1346	16.98	0.0000
Age	2	–0.4081	0.2001	4.16	0.0414
	3	0.0766	0.1885	0.17	0.6844

The first parameter estimate for age (–0.4081) compares "young" with the overall average effect on the logit scale of "young", "middle," and "old," while the second parameter estimate (0.0766) compares the "middle" with this average. That the first is significant, while the second is not, suggests that young stands may have lower mortality probabilities than the average, while the mortality probability for the middle-aged stands is not much different from the average. The difference between the old and the average effect is obtained by adding the first two parameters and changing the sign, namely –1 × (–0.4081 + 0.0766) = –0.3315. In this case, these parameters do not directly test questions of interest, but such questions can be tested with contrasts.

4.2.3 Contrast analysis of the three stand ages Depending on the objectives of the study, different comparisons might be of interest. Two interesting questions might be:

1. whether the young stands have different mortality compared to both the middle- and old-aged stands, and;
2. whether the middle and old stands are different.

These questions can be tested with contrasts and the results are:

Contrast	*df*	χ^2	*p*-value
Young vs middle and old	1	4.16	0.0414
Middle vs Old	1	0.67	0.4137

The first contrast is significant (*p*-value = 0.041). It matches the earlier test for the first age parameter because the test for the difference between young and the grand mean is the same as the test for the difference between young stands and the other levels (because the parameters are constrained to add to zero). The test for a significant difference between middle and old stands was not significant.

Another approach to determining the age effect might use linear and quadratic contrasts to look for trends in response by age. The results of these contrasts are:

Contrast	*df*	χ^2	*p*-value
Linear	1	4.94	0.0263
Quadratic	1	0.17	0.6844

The tests show that while the overall age effect was not strong, there is some evidence of a simpler response. In this case, a linear relationship of the logits of survival with age appears plausible. A more direct test for the linear effect of age can be obtained by treating age as a continuous variable. Notice that both of these approaches (the contrasts and age treated as a continuous variable) assume that the ages are equally spaced. If we knew the real ages, then we could use them to set up the spacing appropriately. For this example, we will assume that the ages are equally spaced. When this direct analysis is performed, the following Wald test results are obtained:

Source	*df*	χ^2	*p*-value
Intercept	1	11.75	0.0006
Age	1	4.86	0.0274
Likelihood ratio	7	1.39	0.9858

These results are similar to the earlier linear contrast and provide support for the linear trend hypothesis. Because we are fitting the probability for death, the parameter estimate for age (0.36, SE = 0.16) is positive, meaning that the mortality probability increases with age.

There are two ways that we could look at the results:

- probability of survival decreases with increasing age; or
- probability of survival is higher for trees in young stands than for those in middle-aged or old stands.

The conclusion depends, in part, on the objective of the study.

4.2.4 Analysis of a two group model Suppose that the appropriate conclusion for our study objectives is that younger stands have a higher survival rate than either the middle- or old-aged stands. To confirm the

Wald test that the middle- and old-aged stands have similar responses, we can use just two levels in our model: young and older stands. From Table 11, the difference in -2LogL is 321.7092 − 321.0392 = 0.67 (with 1 *df* and *p*-value = 0.30). Thus, the two group model provides as good a fit as the three group model, and we may use it as a final description of the data.

The next step in our analysis is to examine more closely the fit of the two group model to ensure that it is adequate. Printer plots of the simple residuals (for the dead response only) against stand number and age are shown in Figure 14. Because they do not show any unusual patterns, the two group model appears to be an adequate model for this data.

Most of this analysis could also have been done using contingency tables (except for the contrasts and the linear age effect). This is described in section 7.1. In addition, the calculations required to fit the various models and conduct the tests are simple enough to do by hand. This is described in section 7.2.

4.3 One-way Classification Study: Fertilizer Trial

The objective of this example is to study the effectiveness of a fertilizer for decreasing seedling mortality during the first year after outplanting. It is hypothesized that increasing the amount of fertilizer will decrease the seedling mortality.

A simple trial might be designed as follows. Each of four levels of fertilizer (including a control) are applied in a completely randomized manner to six rows of ten seedlings soon after planting. Because each row is treated independently, each row should respond to the treatments independently. Nursery procedures such as watering, shading, and pest control should be conducted in such a way as to maintain that independence of response. A year later, each of the 240 seedlings is assessed for mortality. Since the classification variable (level of fertilizer) is quantitative, and the expected response could be linear (on the logit scale), an appropriate contrast is planned for the analysis.

The data set will have the following variables:

- j = row number, $j = 1, 2, \ldots, 24$ (where there is a total of 24 rows)
- i = fertilizer level, $i = 1, 2, 3, 4$
- x_{ij} = amount of fertilizer applied to row j
- y_{ij} = number of surviving seedlings in row j given level i
- $m_{ij} = 10$ = number of seedlings in row j
- $p_{ij} = y_{ij}/m_{ij}$ = observed proportion of surviving seedlings in row j with treatment i.

An appropriate model for the one-way classification is:

$$\text{logit}\,(\pi_{ij}) = \mu + \alpha_i, \tag{8}$$

where π_{ij} = probability that a seedling receiving fertilizer at level i survives for one year, with the number of surviving seedlings $y_{ij} \sim \text{binomial}(\pi_{ij}, m_{ij} = 10)$, and $i = 1, 2, 3,$ and 4 are the four increasing levels of fertilizer.

```
                 Simple One-Way Classification Example
                          Two Group Analysis

                                 Plots

      Plot of _RESID_*STAND.  Legend: A = 1 obs, B = 2 obs, etc.

      Residual proportion

 0.05 +                                                        A       A
      |A
      |
      |                                        A
      |
      |
 0.00 +------------------------------------------------------------------
      |                        A
      |
      |                                                A
      |
      |                                A
-0.05 +        A
      |                A
      |
      |
      |
      |
-0.10 +
      -+-------+-------+-------+-------+-------+-------+-------+-------+-
       1       2       3       4       5       6       7       8       9
                                  Stand Number

        Plot of _RESID_*AGE.  Legend: A = 1 obs, B = 2 obs, etc.

        Residual proportion

          0.05 +                                              B
               |  A
               |
               |                                              A
               |
               |
          0.00 +--------------------------------------------------
               |                                              A
               |
               |                                              A
               |
               |                                              A
         -0.05 +  A
               |                                              A
               |
               |
               |
               |
         -0.10 +
               ---+-------------------------------------------+---
              1..Young                                      Older
                                       Age
```

FIGURE 14 *Printer plots of the simple residuals for the two group model.*

The parameters α_i indicate the additive effect of each level of fertilizer. They are often restricted to sum to zero, so that $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$. This implies that μ is the average effect for the experiment on the logit scale. Two possible results are plotted in Figure 15 and are labelled Outcome 1 and Outcome 2. Note that the relative distances between points around zero (between, say, –1.4 to 1.4) on the logit scale are little affected by transformation to the probability scale (see outcome 1 in Figure 15), whereas points far from zero are compressed (see outcome 2 in Figure 15). The linear and quadratic contrast coefficients for assessing the significance of trends on the logit scale can be determined similarly as for ANOVA (see section 5.1.1.1).

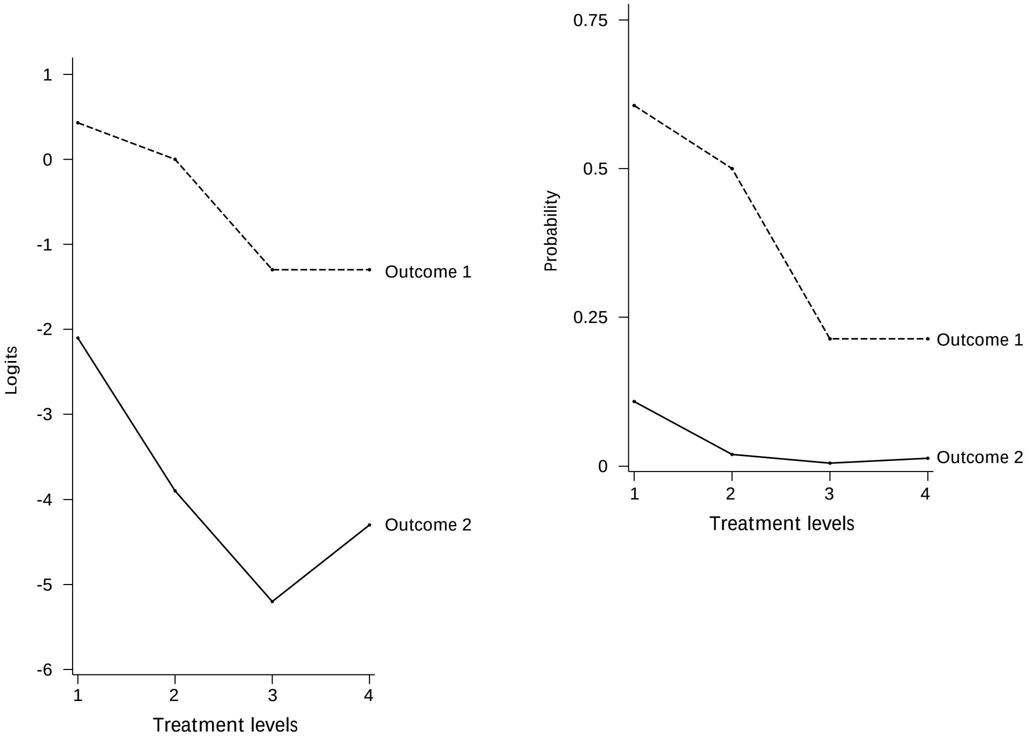


FIGURE 15 *Two possible outcomes of a one-way classification study.*

4.3.1 Data collection and methods In this example, we will look at data from a hypothetical one-way classification. Twenty-four rows, each with ten seedlings, are randomly assigned an amount of fertilizer. Suppose that these amounts are 0, 100, 200, and 300 kg/ha. In all other respects, the seedlings are treated similarly. A growing season later, survival of the seedlings is assessed. The one-way classification model is tested in section 4.3.2, while a simple linear relationship between survival response and fertilizer amount is examined in section 4.3.3.

The initial data set appears below in contingency table form (see Table 12). This was produced using the SAS program code in section 5.4.1.

4.3.2 Contingency table Much of the analysis for this simple example could be done using contingency tables (see section 7.1 for a further

TABLE 12 *Frequency counts for the one-way classification example*

```
---------------------------------------------------------------------------
                           TABLE OF Y BY TREAT

          Y         TREAT

          Frequency|
          Col pct  |       1|       2|       3|       4|  Total
          ---------+--------+--------+--------+--------+
          Alive    |     30 |     42 |     48 |     55 |    175
                   |  50.00 |  70.00 |  80.00 |  91.67 |
          ---------+--------+--------+--------+--------+
          Dead     |     30 |     18 |     12 |      5 |     65
                   |  50.00 |  30.00 |  20.00 |   8.33 |
          ---------+--------+--------+--------+--------+
          Total          60       60       60       60      240

                    STATISTICS FOR TABLE OF Y BY TREAT

          Statistic                     DF     Value        prob
          -------------------------------------------------------
          Chi-Square                     3    28.420       0.001
          Likelihood Ratio Chi-Square    3    29.411       0.001

          Sample Size = 240

---------------------------------------------------------------------------
```

discussion of this topic). The contingency table results shown in Table 12 suggest that there is a treatment effect (likelihood ratio $\chi^2 = 29.41$, 3 *df*, *p*-value ≤ 0.001). It is also useful to test for homogeneity of the row responses within each treatment. The results are shown in Table 13. Three of the four individual contingency tables used to obtain these results received a warning about small sample sizes. This means that we must be cautious about interpreting the results. Nevertheless, the overall test for homogeneity of rows within treatments (obtained by adding the χ^2 values and *df* for tests for each treatment) has a chi-square value of 21.67 with 20 degrees of freedom, which is quite a common value if the null hypothesis is correct.[15] Therefore we might conclude that none of the rows was acting differently from any of the other rows receiving the same treatment. The sample sizes within each row are fairly small. More seedlings would probably be used in a real study. If this homogeneity test was important, then the study should be designed with more trees per row.

[15] See Biometrics Information Pamphlet No. 15 for a discussion of the expected values for the chi-square distribution.

TABLE 13 *Number of surviving trees for each treatment separated by row, including likelihood ratio chi-square from each contingency table analysis*

Treatment	Row within treatment						Likelihood ratio
level (kg/ha)	1st	2nd	3rd	4th	5th	6th	χ^2
0	4	5	6	6	4	5	1.611
100	7	8	6	9	7	5	5.036
200	6	8	6	9	9	10	10.116
300	9	10	9	10	8	9	4.907
						Total with 20 *df*:	21.670

4.3.3 One-way classification analysis The results of the logistic analysis of this data is shown below:

Model	*df*	χ^2	*p*-value
Intercept	1	45.8	0.0001
Treatment	3	24.57	0.0001
Likelihood ratio	20	21.67	0.36

Note the Wald test for treatment effects is χ^2 = 24.57 (with 3 *df*) is similar to the results from Table 12 where the likelihood ratio χ^2 = 29.41 (with 3 *df*). Thus, both tests offer the same conclusion that survival responses differ with treatments. The test for homogeneity of rows within treatment from Table 13 (χ^2 = 21.67) is exactly that shown above. These results show that, with adequate sample sizes, the latter tests the goodness-of-fit of the one-way model. The linear contrast is strongly significant (χ^2 = 21.77, *p*-value ≤ 0.0001) confirming the impression from Table 12 that increasing amounts of fertilizer (over the range tested) are beneficial in improving survival of the seedlings.

4.3.4 Simple linear relationship To obtain an equation for the linear effect of fertilizer on survival, the logistic regression is rerun using treatment level as a continuous variable. A lack-of-fit test for the linear model is obtained by comparing the likelihoods from this model and that from the one-way model: 251.27462 − 250.94999 = 22.00 − 21.67 = 0.325, with two degrees of freedom. This is not significant and implies that the linear model is reasonable. The equation (from Figure 36) is

$$\text{logit} = 0.0143 + 0.00757 \times (\text{treatment level}).$$

The predicted values are given in Table 14 and are quite similar for both models.

The adequacy of the fit of this linear model can be more fully checked by looking at plots of the residuals (see Figure 16). These look fine, remembering that the mirror image effect is due to the two residuals per row (which must add up to zero within each row). The mirror image effect could be eliminated by plotting only the survival response residuals.

TABLE 14 *Predicted survival probabilities for the two different models*

Treatment (kg/ha):	One-way classification	Linear model
0 (control)	0.500	0.504
100	0.700	0.684
200	0.800	0.822
300	0.9167	0.908

```
                                  Fertilizer study
                                   Residual Plots

            Plot of _RESID_*TREAT.  Legend : A = 1 obs, B = 2 obs, etc.
_RESID_ +
        |
    0.4 +
        |
        |
   b0.2 +                    B                 C
        |  B                 A                                   A
        |  B                 A                 B                 B
    0.0 +--D-----------------D-----------------B-----------------F--
        |  B                 A                 B                 B
        |  B                 A                                   A
   -0.2 +                    B                 C
        |
        |
   -0.4 +
        |
        ---+-----------------+-----------------+-----------------+--
           0                100               200               300

                                          TREAT

            Plot of _RESID_*ROW.  Legend: A = 1 obs, B = 2 obs, etc.
_RESID_ |
        |
    0.4 +
        |
        |
    0.2 +                           A     A  A     A        A
        |A           A        A                                            A
        |     A  A              A                    A  A        A     A
    0.0 +---B-----------B--B-----------B--------B--------------B-----B--------B-
        |
        |A           A        A                                            A
   -0.2 +                           A     A  A     A        A
        |
        |
   -0.4 +
        |
        -+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+-
         1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

                                           ROW
```

FIGURE 16 *Printer plots to check the simple residuals for unusual patterns.*

4.4 Multiple Regression: Herbicide Thinning Trial

The objective of this example is to study the effectiveness of a herbicide lance injector for killing aspen trees of different sizes. A tree can receive multiple doses of glyphosate and it is hypothesized that larger trees (as measured by diameter at breast height [dbh]) require more injections to be killed.

A simple trial design is to first select a stand of trees for treatment. Trees to be thinned are chosen according to standard operational procedures. Each tree is then randomly assigned a number of injections. The purpose of any treatment assignment scheme should be to ensure that all treatments are assigned with equal probability to fairly large numbers of trees. Also, it is important that the range of tree sizes assigned to each treatment is large enough so that about half will survive. This will allow good definition of the relationship between treatment, size of tree, and probability of death. Notice that treatments would not be assigned like this on an operational basis because the objective would be to kill all the trees. Here, we are determining the minimum dose required to kill aspen trees.

A year after herbicide injection the treated trees are assessed. The data set has the following variables:

- j = tree number $j = 1, 2, \ldots, 92$
- d_j = dbh of tree j (to the nearest centimetre)
- x_j = number of injections applied to tree j
- y_j = 0 if tree is dead one year later
  = 1 if tree is still alive one year later

A multiple regression model might be appropriate.

$$\text{logit}(\pi_j) = \mu + \beta d_j + \gamma x_j \qquad (9)$$

where: π_j = probability of tree living given d_j and x_j, and whether the tree dies or not is $y_j \sim \text{binomial}(\pi_j, m_j = 1)$.

The number of injections could also be treated as a one-way classification variable with a linear contrast of particular interest. The model shape is presented in Figure 17. The parameter for dbh (β) is assumed to be positive since larger trees are more likely to survive the treatment. On the other hand, the parameter for number of injections (γ) is assumed to be negative since trees of the same size should be less able to survive more injections. Note that the lines on the logit scale are parallel and equally spaced. This shows that the model simply adds the effect of the two independent variables. The curves on the probability scale all have the same shape and are equal distances apart horizontally at any particular probability.

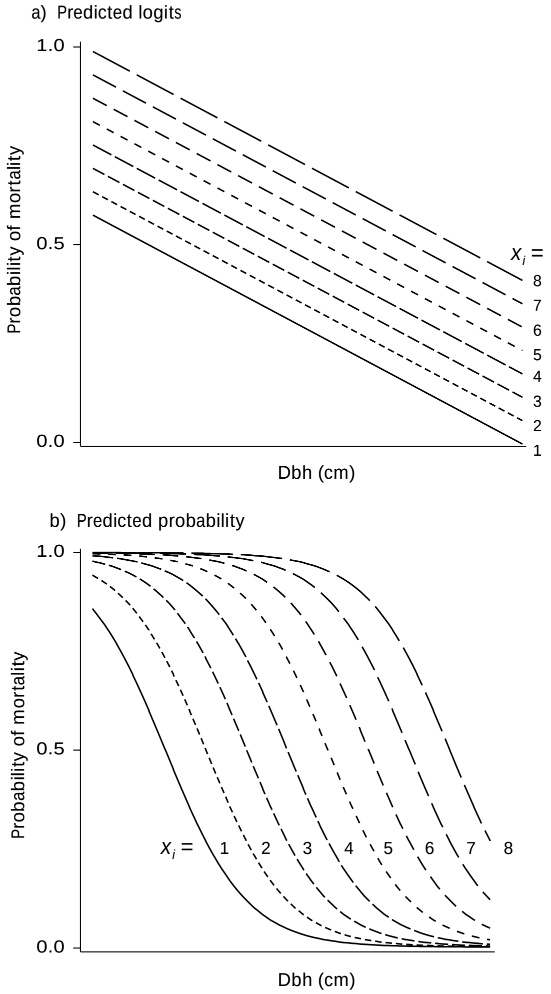


FIGURE 17 *Form of a multiple regression model with one regression parameter negative (dbh) and the other positive (injection number).*

4.4.1 Initial data and methods The data for this study are listed in Appendix 2.[16] Computer output in Table 15 summarizes the tree sizes by dbh class and the number of injections used in the study.

It is clear from this table that the study was designed so that larger trees tended to receive more injections (thus the independent variables are correlated by the design). There is only one tree greater than 35 cm in diameter, and only three trees received just two injections, while another three received eight injections. Otherwise, between 10 and 27 trees received the different injection numbers. If we continue analyzing this data with dbh classes instead of dbh directly, then we may want to pool trees with dbh greater than 30 cm into one class. Table 16 shows the

[16] This data is taken from SX84711Q. The analysis of a similar study with a more complicated design is described in Bergerud (1988).

TABLE 15 *Number of trees in each dbh class with the number of injections given*

```
--------------------------------------------------------------------------
                            TABLE OF DBH BY INJ

   DBH(Aspen dbh (cm))   INJ(Number of Injections)
   Frequency  |     2|     3|     4|     5|     6|     7|     8|  Total
   -----------+------+------+------+------+------+------+------+
    0 - 15 cm |    3 |   10 |    7 |      |      |      |      |     20
   -----------+------+------+------+------+------+------+------+
   15 - 20 cm |      |    5 |    9 |    5 |    1 |      |      |     20
   -----------+------+------+------+------+------+------+------+
   20 - 25 cm |      |    1 |    7 |    6 |    4 |    1 |    1 |     20
   -----------+------+------+------+------+------+------+------+
   25 - 30 cm |      |      |    3 |    7 |    7 |    3 |      |     20
   -----------+------+------+------+------+------+------+------+
   30 - 35 cm |      |      |    1 |    1 |    2 |    6 |    1 |     11
   -----------+------+------+------+------+------+------+------+
   40 - 45 cm |      |      |      |      |      |      |    1 |      1
   -----------+------+------+------+------+------+------+------+
   Total           3     16     27     19     14     10      3       92
--------------------------------------------------------------------------
```

TABLE 16 *Number of dead trees with total trees given a number of injections per dbh class*

```
--------------------------------------------------------------------------
    DBH(Aspen dbh (cm))     INJ(Number of Injections)

  in cell:  number of dead/number of trees:
  Frequency
    Frequency  |     2|      3|     4|     5|     6|     7|     8| Total
    -----------+------+-------+------+------+------+------+------+
     0 - 15 cm |  3/3 | 10/10 |  7/7 |      |      |      |      |    20
    -----------+------+-------+------+------+------+------+------+
    15 - 20 cm |      |   3/5 |  5/9 |  5/5 |  1/1 |      |      |    20
    -----------+------+-------+------+------+------+------+------+
    20 - 25 cm |      |   0/1 |  2/7 |  2/6 |  2/4 |  0/1 |  1/1 |    20
    -----------+------+-------+------+------+------+------+------+
    25 - 30 cm |      |       |  0/3 |  0/7 |  0/7 |  0/3 |      |    20
    -----------+------+-------+------+------+------+------+------+
     > 30 cm   |      |       |  0/1 |  0/1 |  0/2 |  1/6 |  0/2 |    12
    -----------+------+-------+------+------+------+------+------+
    Total:        3/3   13/16  14/27   7/19   3/14   1/10    1/3   42/92
    Percent Dead: 100%   81%    52%    37%    21%    10%     33%    46%
--------------------------------------------------------------------------
```

results when the last two size classes are pooled. We should also consider what to do with the three trees that received two injections: we may either leave them out or pool them with those trees that received three injections.

The response variable is percentage of defoliation, which is measured in 5% increments. For the logistic regression, these values are converted to "dead" or "alive" so that trees with 95% or more defoliation are considered dead.

Overall, about half the trees died (46%) and most of the injection numbers are not completely successful in killing the trees. This is appropriate for an experimental trial because it provides a range of data to which the logistic regression models can be adequately fit. Then the models can be used to estimate the size of tree that each injection number kills. After all, if most or all of the trees either died or survived, it would be hard to assess what minimum dose would have a good chance of killing the trees.

The analysis will proceed as follows: first the multiple regression model is fitted and interpreted, and the residuals of the multiple regression model are examined. Then tests for lack of fit of the multiple regression model are performed. This is accomplished by fitting factorial models using dbh class and/or injection number as categorical variables. Since this is not a complete factorial, some fitting problems will occur. However, we can still compare the results to the multiple regression model.

4.4.2 Multiple logistic regression model We will start with a multiple logistic regression model where both dbh and number of injections are treated as continuous variables (covariates). The results of model fitting are shown in the computer output in Figure 18. The chi-square test for the combined contribution of the two variables is 85.6 with two degrees of freedom, which is strongly significant with a p-value ≤ 0.0001. The Wald tests for `INJ` (p-value = 0.0027) and `DBH` (p-value ≤ 0.0001) suggest that they are both important components in the model.

The Wald tests for the individual effects in the model are approximate and values with marginally significant p-values ($0.05 \leq p$-values ≤ 0.10) can be confirmed by fitting models with and without each term. While this is hardly necessary here, the differences in the $-2\text{Log}L$ statistic between models with and without each variable is quite large and are summarized in Table 17. This value increases from 41.87 to 58.68 (a difference of 16.81 with one degree of freedom) when injection numbers are removed from the model. When only dbh is removed, this value increases to 107.23, a change of 65.36 with one degree of freedom. The model with neither variable fits one probability to all the data assuming that there is no effect of either variable (the intercept only model). In this case, the likelihood ratio increases from 41.87 to 127.50, a change of 85.60 for two

TABLE 17 *Deviance tests for number of injections and dbh*

Model	−2LogL	Test for	Deviance	*df*	*p*-value
Injection and dbh	41.869				
		Dbh	65.364	1	≤0.0001
Injection	107.233				
		Injection	16.809	1	≤0.0001
Dbh	58.678				
		Dbh and Injection	85.627	2	≤0.0001
Intercept only	127.496				

```
---------------------------------------------------------------------------
                        Aspen Injection Trial
                    Logistic Regression Analysis

                       The LOGISTIC Procedure

   Model Fitting Information and Testing Global Null Hypothesis BETA=0

                               Intercept
                  Intercept       and
   Criterion        Only      Covariates  Chi-Square for Covariates

   AIC            129.496      47.869       .
   SC             132.017      55.435       .
   -2 LOG L       127.496      41.869     85.626 with 2 DF (p=0.0001)
   Score             .            .       56.137 with 2 DF (p=0.0001)

                 Analysis of Maximum Likelihood Estimates

              Parameter Standard   Wald       Pr    Standardized  Odds
   Variable DF Estimate   Error Chi-Square Chi-Square  Estimate   Ratio

   INTERCPT 1  -11.0306   2.5545   18.6458     0.0001           .  .
   INJ      1   -2.0743   0.6919    8.9874     0.0027   -1.662060 0.126
   DBH      1    0.9813   0.2292   18.3323     0.0001    3.684249 2.668
---------------------------------------------------------------------------
```

FIGURE 18 *Results of the multiple regression fit.*

degrees of freedom. These likelihood ratio tests provide even stronger evidence for tree size (dbh) and injection number effects than do the Wald tests. Therefore, a reasonable model includes both variables in a linear multiple logistic regression model.

4.4.3 Model interpretation To understand this model, we first look at the parameter estimates that appear in Figure 18. These can be used to determine the equation which describes the model for fitting the probability of survival, namely:

$$\text{logit}(\pi_{ij}) = -11.0306 - 2.0743 \times \texttt{INJ} + 0.9813 \times \texttt{DBH}.$$

It is useful to write this equation out separately for each number of injections and to calculate the size of tree which will have a corresponding 50 and 95% chance of being killed. The results are presented in Table 18.

To show how the predicted size of trees at a given probability of mortality can be calculated, let us use the above equation to make the calculation for four injections and a mortality probability of 95%. The equation for survival is: $\text{logit}(\pi_{ij}) = -19.3278 + 0.9813 \times \texttt{DBH}$. A mortality of 95% (or 5% survival) implies that the above equation should be equal to

TABLE 18 *Survival logit equations for each injection number, with the size of tree (in dbh, cm) expected to have a 50 and 95% chance of survival*

```
--------------------------------------------------------------------------
                            Aspen Injection Trial
                        Logistic Regression analysis
                       Using Multiple Regression Model
                 Listing of equations for each injection number

  Injection                              Predicted Size    Predicted Size
    Number    Intercept  B2 * DBH     at 50% mortality  at 95% mortality

       1     -13.1049    0.98133          13.3542           10.3537
       2     -15.1792    0.98133          15.4679           12.4675
       3     -17.2535    0.98133          17.5817           14.5812
       4     -19.3278    0.98133          19.6955           16.6950
       5     -21.4021    0.98133          21.8092           18.8088
       6     -23.4764    0.98133          23.9230           20.9225
       7     -25.5507    0.98133          26.0368           23.0363
       8     -27.6250    0.98133          28.1505           25.1501
--------------------------------------------------------------------------
```

logit(.05) = log(.05/.95) = –2.94 (see Table 3).[17] Rearranging and solving for dbh, we obtain: DBH = (19.3278 – 2.94) / 0.9813 = 16.70 cm.

It would also be useful to calculate the dbh of trees predicted by the model to die with specific probabilities for each number of injections. These values could be used to prepare field guides about treatment of trees of different sizes. In this case, a graph of the results shows that the constant probability lines are straight and that with each additional injection the size of tree must increase (by about 2.1 cm from Figure 19) to have the same probability of surviving or of dying.

4.4.4 Assessing the adequacy of the multiple regression model As a final check of the multiple regression model, the residuals should be plotted against the independent variables, dbh and inj, and the tree injection number. The plots (Figures 20 and 21) show no obvious patterns and no particular observation appears to stick out, suggesting that the model is adequate.

4.4.5 Factorial models: looking for non-linearity in response It is possible with this example to check for non-linearity in the survival probability responses to both dbh and injection number. We can do this by examining more complicated models. For instance, because many trees (experimental units) are assigned to each injection number, we could treat injection number as a categorical variable and leave dbh as a continuous

[17] We use logit(0.05) instead of logit(0.95) because the equation fits the probability of survival not of mortality. A mortality of 95% corresponds to a 5% survival.

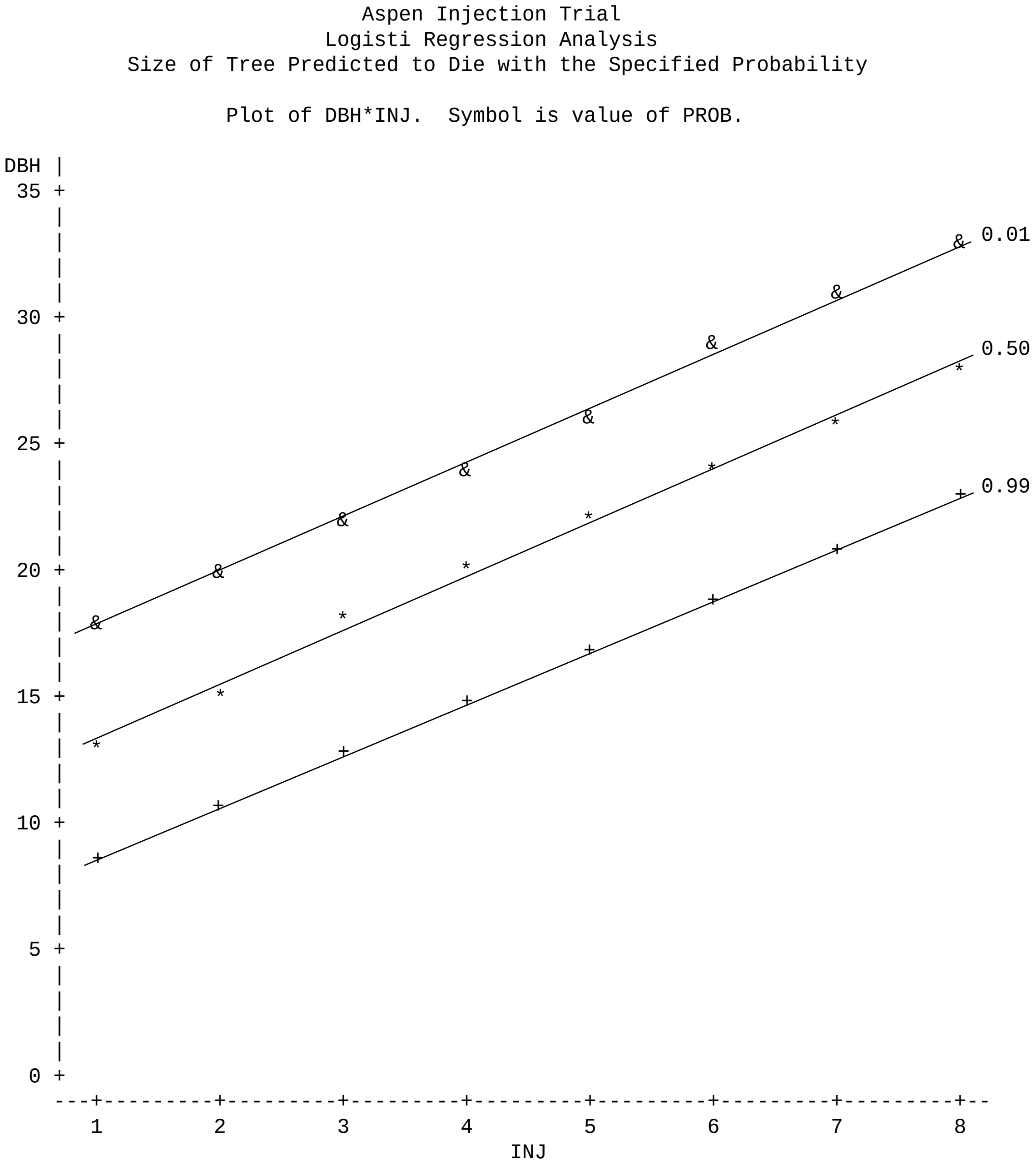


FIGURE 19 *Plot of tree size (dbh, cm) expected to die (with a 1, 50 and 99% probability) as a function of the number of injections received. Notice that these probabilities are not confidence intervals, but are the probability of death. (Lines are hand-drawn.)*

```
                           Aspen Injection Trial
                           Pearson Residual Plots

          Plot of RESCHI*DBH.  Legend: A = 1 obs, B = 2 obs, etc.
   |
 4 +
   |
   |
   |                             A
 2 +             A
   |              A            A
   |                     B
   |                      BA   AB A  A B  A
 0 +--ABAE-CABBB-AAB----A---A--B--A-AB-GAA--BABC-AAC--A--------------A--------
   |           AA B AC B AAA A A
   |                     A
   |                               A
-2 +                       B
   |                                     A
   |
   |
-4 +
   --+---------+---------+---------+---------+---------+---------+---------+--
    10        15        20        25        30        35        40        45

                                  Aspen dbh (cm)
```

Pearson Residual

```
          Plot of RESCHI*INJ.  Legend: A = 1 obs, B = 2 obs, etc.
   |
 4 +
   |
   |
   |                                                            A
 2 +            A
   |            A                                   A
   |                        B
   |                        C           D           C           A
 0 +C-----------J-----------O-----------J-----------G-----------G-----------C-
   |            D           D           E           B
   |                        A
   |                                                A
-2 +                        B
   |                                                            A
   |
   |
-4 +
   -+-----------+-----------+-----------+-----------+-----------+-----------+-
    2           3           4           5           6           7           8

                               Number of Injections
```

Pearson Residual

FIGURE 20 *Printer plots of the Pearson residuals for model diagnostics.*

```
                             Aspen Injection Trial
                            Deviance Residual Plots

            Plot of RESDEV*DBH.  Legend: A = 1 obs, B = 2 obs, etc.

Deviance Residual
     |
   2 +                             A
     |             AA
     |                           A
     |
     |                     B    A
     |                      BA   B
     |                             A  A B  A
     |                        A  B     A C     AA   A
   0 +--A-AD---BA---------------------AA-DAA--B-AC-A-C--A--------------A--------
     |   B A CA AB AAB   BA         A
     |           AA   AC   A   A
     |              B       AA   A
     |
     |                     A
     |
     |                       B       A
  -2 +                                     A
     --+---------+---------+---------+---------+---------+---------+---------+--
      10        15        20        25        30        35        40        45

                                    Aspen dbh (cm)

            Plot of RESDEV*INJ.  Legend: A = 1 obs, B = 2 obs, etc.
Deviance Residual
     |
   2 +                                                            A
     |            B
     |                                                A
     |
     |                        B           A
     |                        C           B
     |                                    A           C           A
     |            A           B           A           C           B           A
   0 +------------D-----------I-----------G-----------C-----------E-----------A-
     |C           E           D           D           A                       A
     |            B           D           A           A
     |            B                       B           A
     |
     |                        A
     |
     |                        B                       A
  -2 +                                                            A
     -+-----------+-----------+-----------+-----------+-----------+-----------+-
      2           3           4           5           6           7           8

                                 Number of Injections
```

FIGURE 21 *Printer plots of the deviance residuals for model diagnostics.*

variable at first. The test for injection number would then check for any differences between the numbers of injections and not just for a linear trend in the response. The difference in −2LogL of this model and the multiple regression would be a test for non-linear trends in the response to number of injections. When this analysis is performed, the resulting difference in −2LogL (comparing the discrete model with the continuous multiple regression model) is 41.87 − 40.56 = 1.31 with four degrees of freedom, which is clearly not significant. The test for a linear response using a contrast is $\chi^2 = 7.52$ with one degree of freedom and a *p*-value of 0.0061. These results suggest that a linear response to numbers of injections is sufficient.

To test for non-linearity for both injection number and dbh, requires that we convert the dbh variable into a categorical variable. We can do this by using the class intervals of Table 16 and fitting a two-way factorial to the data. This treats both dbh class and injection number as categorical variables, which allows us to look for any pattern in the response to these variables. We might expect to have some trouble with this model because many combinations of dbh class and injection number are missing (indicated by the missing cells in Table 16). See section 5.5.5 for more discussion of fitting problems.

An overall test of the lack of fit of the multiple regression model can be calculated by looking at the difference in the −2LogL's of the multiple regression and full factorial models: 41.87 − 40.62 = 1.25, with 89 − 78 = 11 *df*. This is clearly not significant, which implies that the multiple regression model appears to be quite adequate at fitting the data.

4.5 One-way Classification With Covariate: Root Collar Weevil Study

The objective of this example is to study the effectiveness of screefing (removing the duff from around the seedling) to protect recently planted seedlings from the root collar weevil. Since the weevil lives in the duff, it is hypothesized that screefing will reduce its ability to attack the seedling. The problem is complicated because even in small areas the density of root collar weevil is highly variable. Therefore some measure of weevil density is needed to indicate the chances of attack.

A completely randomized design is planned for a cutblock with a high prevalence of the root collar weevil. Twenty-eight square plots, each containing sixteen seedlings, are chosen as experimental units. The screefing treatment is randomly assigned to fourteen of the plots. Individual seedlings within these plots are screefed. The fourteen remaining plots are left as untreated controls. A weevil trap is placed in the middle of each plot. It is expected that the numbers of weevils caught in the traps will not be affected by the treatment assigned to the plots. It is also expected that the trap will accurately indicate weevil density throughout the plot, while not affecting the weevil attack rate for any seedling within the plot. For instance, the seedlings next to the trap are assumed as likely of attack as those farther away. Responses from each plot should be independent.

At the end of the season, presence or absence of attack is assessed for each seedling and the numbers of weevils found in each trap counted. The data set would have the following variables:

- j = plot number, $j = 1, 2, \ldots, 28$ (where there are 28 plots in all)
- $i = 0$ if plot was not screefed
  $= 1$ if plot was screefed
- x_{ij} = number of weevils found in trap of plot j, treatment i
- y_{ij} = number of seedlings attacked in plot j, treatment i
- $m_{ij} = 16$ = number of seedlings in plot j, treatment i
- $p_{ij} = y_{ij}/m_{ij}$ = observed proportion of seedlings attacked in plot j, treatment i

4.5.1 A standard model of parallel lines The probability of weevil attack (π_{ij}) can be hypothesized to increase with weevil numbers, but be reduced by the screefing treatment. Thus the data might be modeled by:

$$\text{logit}(\pi_{ij}) = \mu + \alpha_i + \beta x_{ij}, \tag{10}$$

where: π_{ij} is the proportion of seedlings attacked by the root collar weevil, with $\alpha_1 < \alpha_0$, $\beta > 0$ and the number of attacked seedlings is $y_{ij} \sim \text{binomial}(\pi_{ij}, m_{ij} = 16)$.

Equation (10) can be rewritten as two predictive equations, one for each treatment:

$$\begin{aligned} \text{control: } \text{logit}\ (\pi_{0j}) &= \mu + \alpha_0 + \beta x_{0j} \\ \text{treatment: } \text{logit}\ (\pi_{1j}) &= \mu + \alpha_1 + \beta x_{1j}. \end{aligned} \tag{11}$$

The treatment terms (α_0 and α_1)[18] are usually defined so that $\alpha_0 + \alpha_1 = 0$ or $\alpha_0 = -\alpha_1$.

On the logit scale, this is a model of parallel lines with different intercepts for each treatment (if the response variable were normally distributed, this would be a typical analysis of covariance model: see Biometrics Information Pamphlets Nos. 31 and 46 and Biometrics Information Handbook No. 1). The slope β should be positive because increases in weevil numbers will increase the probability of attack. The intercept α_0 (control) will be larger than α_1 (screefed) if the treatment is effective. The form of the model is shown in Figure 22.

Parallel lines on the logit scale (see Figure 22a) transform into logistic curves of the same shape, but shifted sideways from each other on the probability scale (see Figure 22b). Both the horizontal and vertical differences between parallel logit lines is constant, while only the constant horizontal distance is maintained on the probability scale. The constant vertical difference between two parallel logit lines is $\delta = (\alpha_0 - \alpha_1)$, while the horizontal difference[19] is $(x_{0j} - x_{1j}) = \delta/\beta$. The back-transformation to the probability scale preserves the horizontal distance of δ/β (showing that

[18] Otherwise the model would have three parameters: μ, α_0, and α_1 to describe only two means or intercepts. This is called overparameterization. One way to deal with this is to restrict the parameters so that instead of three, there are really only two (after all, if you know α_0 then you know $\alpha_1 = -\alpha_0$).

[19] This can be found by setting logit (π_{0j}) = logit (π_{1j}) and solving for $(x_{0j} - x_{1j})$.

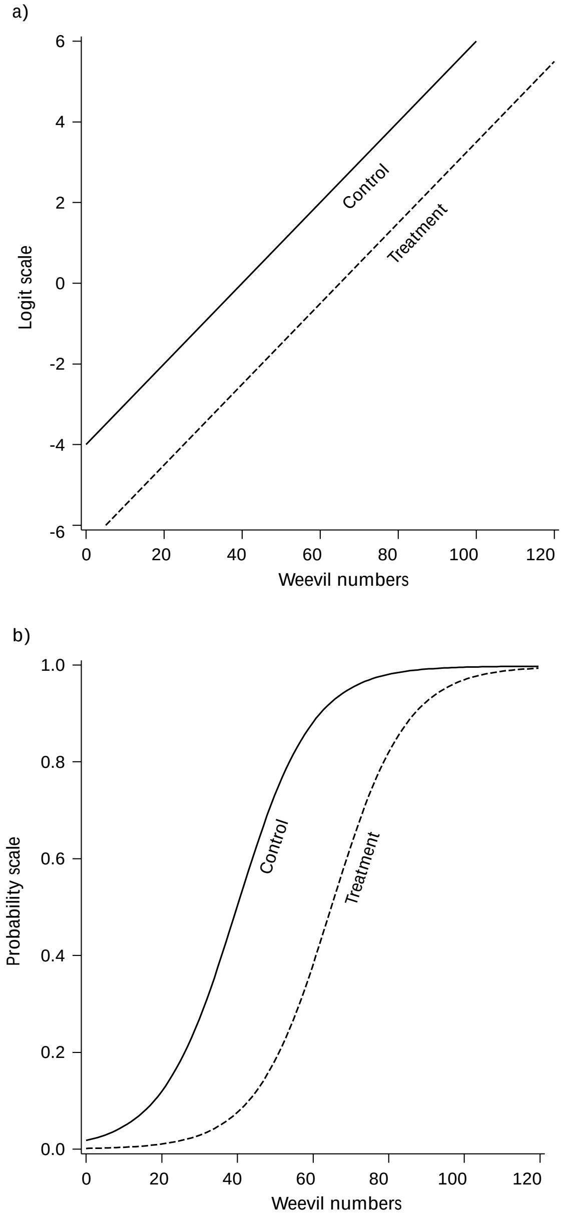


FIGURE 22 *Form of the model where the treatment effect is to shift down or sideways with the attack rate.*

the curves are shifted sideways from each other), but not the vertical distance of δ. This is because the logit lines near low and high values (for values less than –1.4 [π about 0.2] and greater than 1.4 [π about 0.8]) are compressed to make the characteristic S-shaped curves (see Figure 23a and 23c). The lines remain nearly parallel for probability values between 0.2 and 0.8 (see Figure 23b).

Interestingly, the obvious parallelism on the logit scale is apparently lost on the probability scale if the entire logistic curve is not shown on the graph. This would occur, for instance, when the range of the x-variable or covariate is narrow relative to the curve. To see this more clearly let us examine the probability plot in Figure 23 more closely. The whole figure

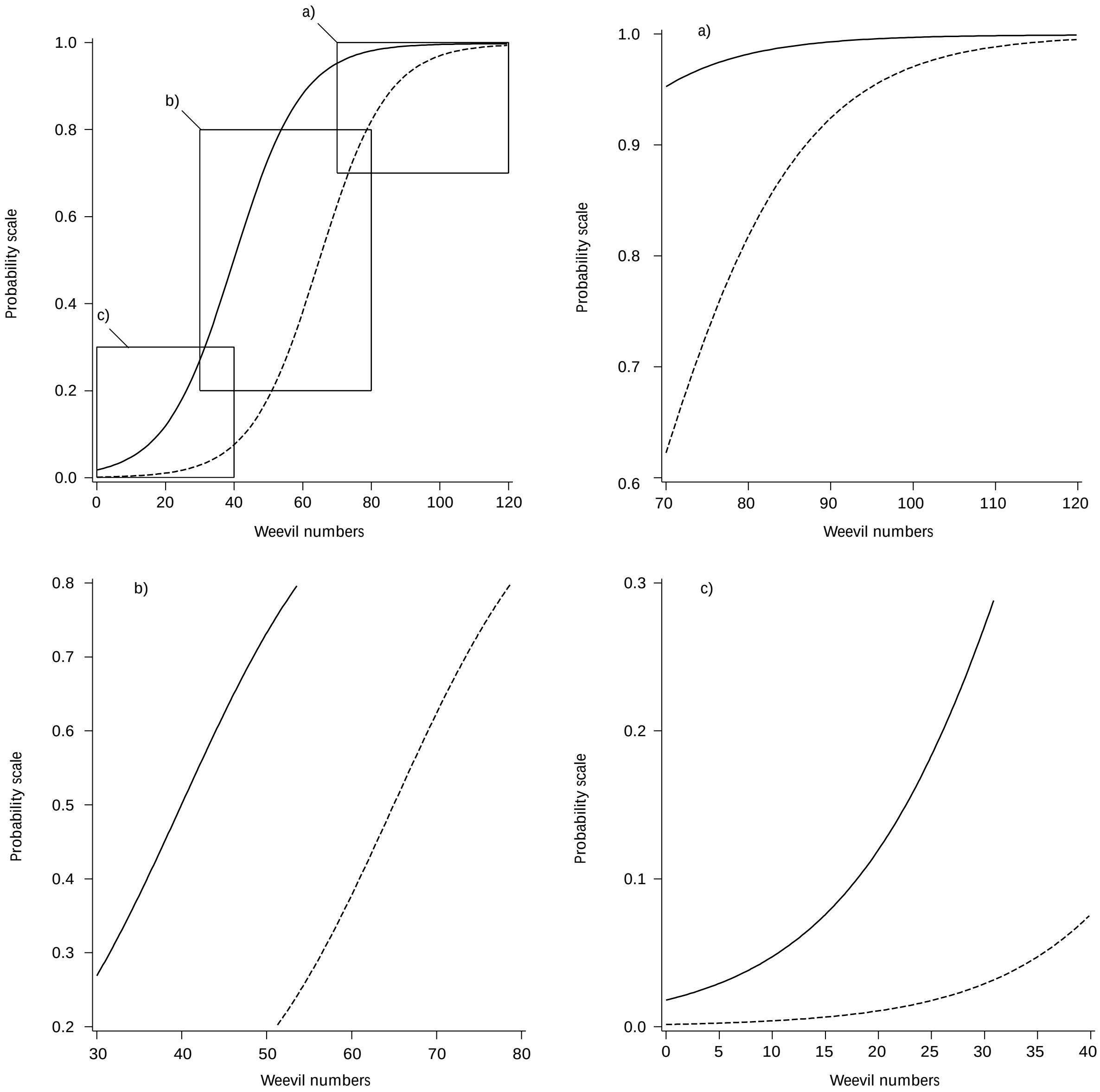


FIGURE 23 *Illustration of how the parallel model can appear non-parallel when the whole curve is not examined.*

has been split into three parts. Only the middle part (Figure 23b) still looks like a parallel-lines model. The other two parts (Figures 23a and 23c) do not appear to be similarly shaped probability models that have been shifted sideways. This can lead to the confusing situation where a parallel logit model will appear to predict non-parallel probability curves. An example of this will be seen in the data analysis (section 4.5.7). When drawing pictures of expected and fitted results, we should remember this possible phenomenon.

4.5.2 The log-odds ratio Another important concept associated with the parallel-lines model is the log-odds: $\log[\pi/(1-\pi)]$, first defined as the logit transformation of the probability (π) in section 2.3. Note that the term, $\pi/(1-\pi)$, is known as the odds in gambling. For instance, if π is the probability of throwing a one on a die, then $1 - \pi$ is the probability of not throwing a one on the die. If these probabilities are 1/6 and 5/6, then the odds of throwing a one are (1/6)/(5/6) = 1/5. That is, the odds of throwing a one are 1:5. On the other hand, the odds of not throwing a one are 5:1. The log-odds is simply the log of this odds. This concept is used extensively in the health sciences, especially in the field of epidemology. See, for example, Breslow and Day (1980). The log-odds ratio is the log of the ratio of two sets of odds. For the root collar weevil study, the log-odds ratio that compares screefed and control plots can be obtained by subtracting the two parts of equation (11):

$$\log[\pi_{0j}/(1 - \pi_{0j})] - \log[\pi_{1j}/(1 - \pi_{1j})] = \mu - \mu + \alpha_0 - \alpha_1 + \beta(x_{0j} - x_{1j})$$

so that,

$$\log \left\{ \frac{\pi_{0j}/(1 - \pi_{0j})}{\pi_{1j}/(1 - \pi_{1j})} \right\} = \delta + \beta(x_{0j} - x_{1j}) \text{ (recall that } \delta = \alpha_0 - \alpha_1) \quad (12)$$

The term on the left is the log-odds ratio, while the term inside the braces is called the odds ratio (ψ). Hence:

$$\log \psi = \delta + \beta(x_{0j} - x_{1j}). \quad (13)$$

This equation shows that if weevil numbers are the same for the control and treated plots (i.e., $x_{0j} = x_{1j}$), then the odds ratio of attack is equal to $\psi = \exp(\delta)$. For example, suppose that $\delta = 0.9$, then the odds ratio is $\exp(0.9)$, which is about 2.5. This means that the odds of attack, regardless of the number of weevils, is about 2.5 times greater for control trees than for treated trees. Note that because of the S-shaped curve, the probabilities of attack do not have a simple relationship which is independent of weevil density. At low levels of weevil density, the attack probabilities may be very low for both groups, while at a medium level of weevil density the attack probabilities may be very different, although the odds ratio is 2.5 in both cases.

4.5.3 Separate-lines model The model just described is very commonly used when the data include one or more categorical variables and one or more continuous-valued covariables. Nevertheless, other models are possible. Considering the possible results of successful treatments can lead to quite a different model. For instance, if the screefing treatment is very successful, then the attack probability should be close to zero, whatever the prevailing weevil population. Expected treatment results are shown in Figure 24. A model that may fit this pattern is:

$$\text{logit}(\pi_{ij}) = \mu + \alpha_i + \beta_i x_{ij} \quad (14)$$

This model implies that:

$$\begin{aligned} \text{control: } \text{logit}(\pi_{0j}) &= \mu + \alpha_0 + \beta_0 x_{0j} \\ \text{treatment: } \text{logit}(\pi_{1j}) &= \mu + \alpha_1 + \beta_1 x_{1j} \end{aligned}$$

In addition to allowing different intercepts (indicated by α_i), this model allows the two treatments to have different slopes (indicated by β_i). For this example, we might expect that when x_{ij} approaches zero, the two treatment probabilities would both be about zero, so that $\text{logit}(\pi_{0j}) = \text{logit}(\pi_{1j}) = -\infty$ (infinity). This implies that $\alpha_0 = \alpha_1 = 0$, with μ having a large negative value. Further, if screefing is very successful, then we may expect that $\beta_1 = 0$. Thus, a final model to consider fitting is:

$$\text{logit}\ (\pi_{ij}) = \mu + \beta_0 x_{ij} \tag{15}$$

This equation implies that:

$$\begin{aligned} \text{control: } \text{logit}(\pi_{0j}) &= \mu + \beta_0\ x_{0j} \\ \text{treatment: } \text{logit}(\pi_{1j}) &= \mu \end{aligned}$$

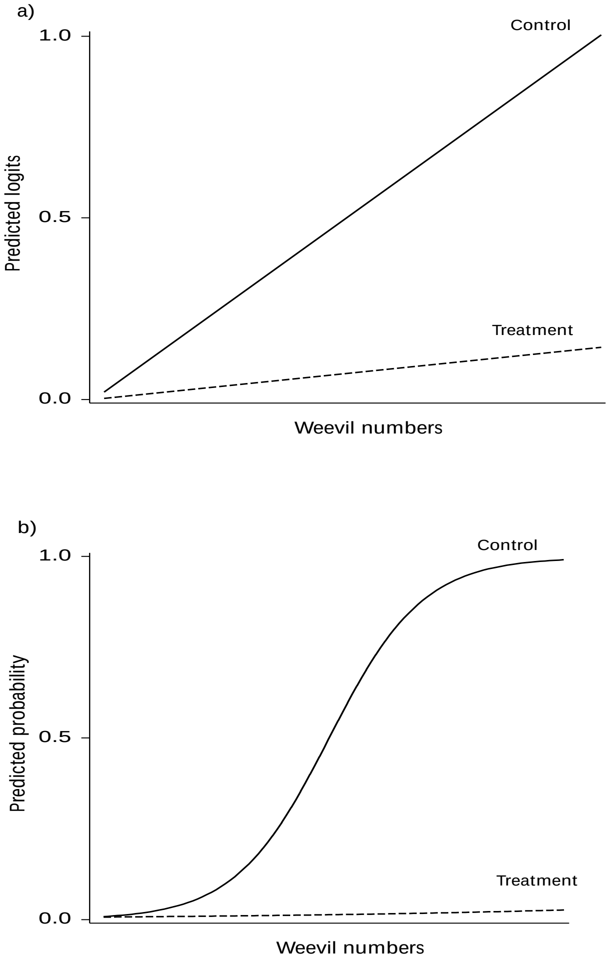


FIGURE 24 *Plots of weevil attack if treatment is very successful. Note that the probability plot (b) might have a corresponding logit plot with parallel lines as shown in Figure 22 and not as shown here.*

The log-odds ratio is log $\psi = \beta_0 x_{0j}$, which is no longer independent of weevil density. As the weevil numbers increase, the odds of attack for the control plots would increase compared to the odds of attack for the treated plots.

4.5.4 Initial data input and methods for root collar weevil study

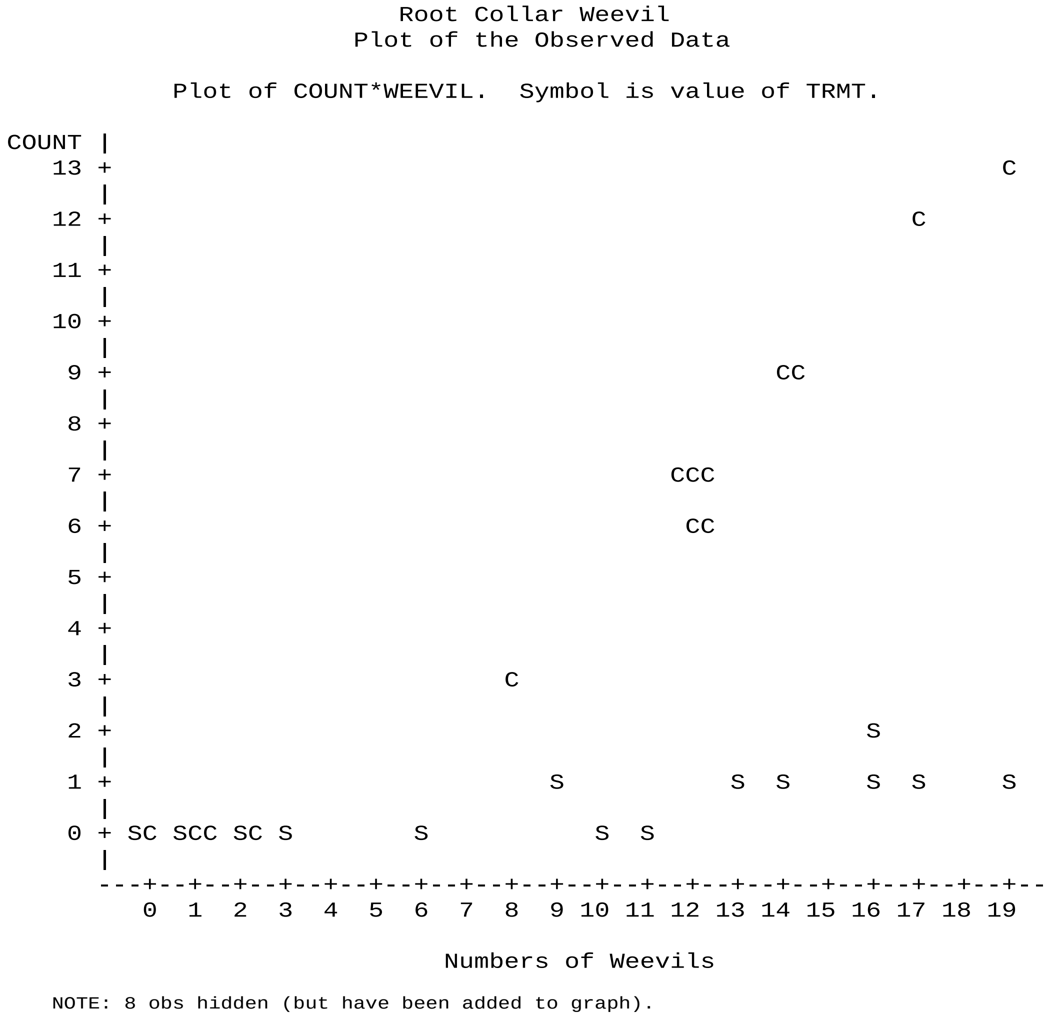
Suppose that this hypothetical study has been designed and conducted as described. The numbers of weevils collected in traps were counted and the number of seedlings that were attacked by the root collar weevil was recorded. These data are listed in Appendix 2 and summarized in Table 19. A printer plot of the data is presented in Figure 25.

The analysis will proceed as follows: first, we fit the standard analysis of covariance type model of two parallel lines (section 4.5.5). We then test for non-parallelism of the lines (heterogeneity of regression). If this is not rejected, then we test if the slope of the lines is zero and if the separation between the lines (the screefing effect for this model) is zero. Secondly, we parameterize the model by using two different variables for weevil numbers (one for each treatment level) instead of just one variable. This reparameterization allows us to easily fit the alternate model (section 4.5.6). Finally, we compare the fit of three of the eight models examined (section 4.5.7).

TABLE 19 *Data collected for root collar weevil study*

	Treatment		Control	
Plot No.	No. of seedlings attacked	No. of weevils	No. of seedlings attacked	No. of weevils
1	0	10	6	12
2	1	13	6	12
3	0	11	12	17
4	2	16	7	12
5	1	16	9	14
6	1	9	7	12
7	0	6	7	12
8	1	14	3	8
9	1	17	9	14
10	1	19	13	19
11	0	0	0	2
12	0	3	0	1
13	0	2	0	0
14	0	1	0	1

4.5.5 Standard analysis of covariance model The explanatory variables in this example include the treatment, screefing or not, as a categorical variable. The second variable, numbers of root collar weevil, is a continuous variable and is required to confirm that root collar weevils were present and a potential threat to the seedlings. This type of design is a one-way classification with a covariable and is typically considered an analysis of covariance. It is in fact a study of two lines as you can see from Figure 25. The sequence of logistic regression models fitted is shown in Table 20. The resulting tests comparing the models are summarized in Table 21.



```
                          Root Collar Weevil
                       Plot of the Observed Data

           Plot of COUNT*WEEVIL.  Symbol is value of TRMT.

COUNT |
   13 +                                                           C
      |
   12 +                                                     C
      |
   11 +
      |
   10 +
      |
    9 +                                            CC
      |
    8 +
      |
    7 +                                     CCC
      |
    6 +                                      CC
      |
    5 +
      |
    4 +
      |
    3 +                          C
      |
    2 +                                                  S
      |
    1 +                             S           S  S     S  S     S
      |
    0 + SC SCC SC S        S           S  S
      |
      ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+---
         0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                             Numbers of Weevils

   NOTE: 8 obs hidden (but have been added to graph).
```

FIGURE 25 *Printer plot of the observed attack counts out of 16 trees per plot.*

TABLE 20 *Models fitted to the root collar weevil data*

Models fit to the data	*df*	Likelihood ratio χ^2
Model 1: Full model with two separate lines	24	8.38
Model 2: With two parallel lines	25	9.85
Model 3: With treatment only (two groups)	26	102.0
Model 4: With weevil only (one line)	26	112.9

TABLE 21 *Test results comparing the four models in Table 20*

Test	Models used	*df*	χ^2-value	*p*-value
H_0: Lines are parallel	1 and 2	1	1.47	0.23
H_0: Parallel: no treatment differences	2 and 4	1	92.15	0.0001
H_0: Parallel: slope of line is zero	2 and 3	1	103.05	0.0001

The first model is the full model which fits two logit lines with separate intercepts and slopes. Since m_i = 16 is small, we ignore the likelihood ratio test as a goodness-of fit test. The Wald tests for treatment, weevils, and their interaction (χ^2 = 0.70, 24.83, and 1.69, respectively with 1 *df* each) suggest a weevil effect, but not much of an interaction or a treatment effect. Nevertheless, these results should be checked by running the reduced (or simpler) models and comparing the change in the likelihood ratio. Model 2 leaves out the interaction between weevil and treatment which forces the two slopes to be the same (i.e., forcing the lines to be parallel).

The difference in χ^2 between Model 1 and 2 is small (9.85 − 8.38 = 1.47 with 1 *df*), so we could decide to consider the slopes as reasonably parallel and that this simpler model fits adequately. The Wald tests for treatment and weevils (χ^2 = 60.4 and 43.2 respectively, with *df* = 1 each) are now much larger with *p*-values ≤ 0.0001. Models 3 and 4 can be used to confirm the Wald tests on whether treatment or weevil should be in the final model (although this is hardly necessary because the probabilities for them are very low; this is more important when the *p*-values are borderline). The likelihood ratio statistics increase dramatically for models 3 and 4 (χ^2 = 102.0 and 112.9, respectively). Hence we conclude that the "ordinary" type ANCOVA model of two parallel lines which are shifted sideways by the treatment fits this data well and that there are both treatment and weevil effects.

4.5.6 Alternative models Recalling the discussion in section 4.5.3 suggests that we should consider some other models. At low numbers of weevils, we could expect the attack probabilities for both treatments to be small and therefore we could fit two separate lines with a common intercept. This is a radiating lines model. Since the probability of attack increases with increasing root collar weevil numbers, the lines should have positive slopes. If the treatment is very effective, then the slope for the treatment could be zero. An easy way to test these questions is to create two new weevil variables from the original weevil variable (x_{ij}), one with the weevil values for the treatment observations (x_{1j} only), but zero values for the control observations, with the other variable arranged in the opposite way (x_{0j} only). The fitting results are summarized in Table 22 and appropriate tests are presented in Table 23. The model which includes both of these variables and treatment (to indicate the intercept) is the same as the full model with non-parallel lines, since separate intercepts and slopes are fit to each line (see models in Table 20 and model 5 in Table 22).

The model that forces the slope to zero for the treated trees, but allows a slope for the control trees (equation 15 and model 8) may be reasonable, but does not fit as well as the two radiating lines model (model 6 which includes both of the new weevil variables). Its fit is worse than that of the full model (χ^2 = 9.32, *p*-value <0.01). On the other hand, the radiating lines model fits almost as well as the full model (χ^2 = 0.80, *p*-value = 0.37), and would be a suitable final model.

TABLE 22 *Alternative models fitted to the root collar weevil data*

Models fit to the data	*df*	Likelihood ratio χ^2
Model 5: Two weevil variables and treatment	24	8.38
Model 6: Two radiating lines model	25	9.18
Model 7: Treatment weevil variable only	26	144.45
Model 8: Control weevil variable only	26	17.70

TABLE 23 *Test results comparing the four models in Table 22*

Test	Models used	*df*	χ^2-value	*p*-value
H_0: Both lines have same intercept	5 and 6	1	0.80	0.37
H_0: Slope for treatment is zero	6 and 7	1	8.52	0.0035
H_0: Slope for control is zero	6 and 8	1	135.27	0.0001
H_0: Same intercept and treated slope is zero	5 and 8	2	9.32	<0.01

4.5.7 Comparing the models To compare these models further, we could look at the sums of squares of the simple residuals (Table 24). They show little difference in fit between the parallel and two radiating lines models. The one line only model shows 10% more variability than the two radiating lines model, while having one less parameter. The equations for the three models are shown in Table 25. Plots of the predicted logits and probabilities are presented in Figures 26, 27, and 28.

We should also look at plots of the residuals to check that the models do adequately fit the data. The Pearson residuals are shown in Figure 29 while the deviance residuals are shown in Figure 30. The plots for the two kinds of residuals are similar and indicate that all three models provide reasonable fits.

The parallel lines model (model 2) and the two radiating lines model (model 6) fit the data similarly, while the one line for control only model (model 8) does not fit quite as well. The final choice of model would depend on the objectives of the study, how the model is to be used, and knowledge of the biology for this specific situation.

It is odd that the parallel lines and the two radiating lines models should fit so similarly. On the logit scale where we are fitting straight lines, they look quite different (see Figure 26). While on the probability scale they look very similar (see Figure 27). We can see from the probability curves that the data covers only the lower part of the S-shaped probability curves for the treatment. This is where the fitted logit values are back-transformed into a narrow range of probability values so that the difference between the lines on the logit scale is not so marked on the probability scale (recall the discussion in section 4.5.1). These plots help us to understand why the parallel lines model fits so well, even though the lines are not parallel on the probability scale.

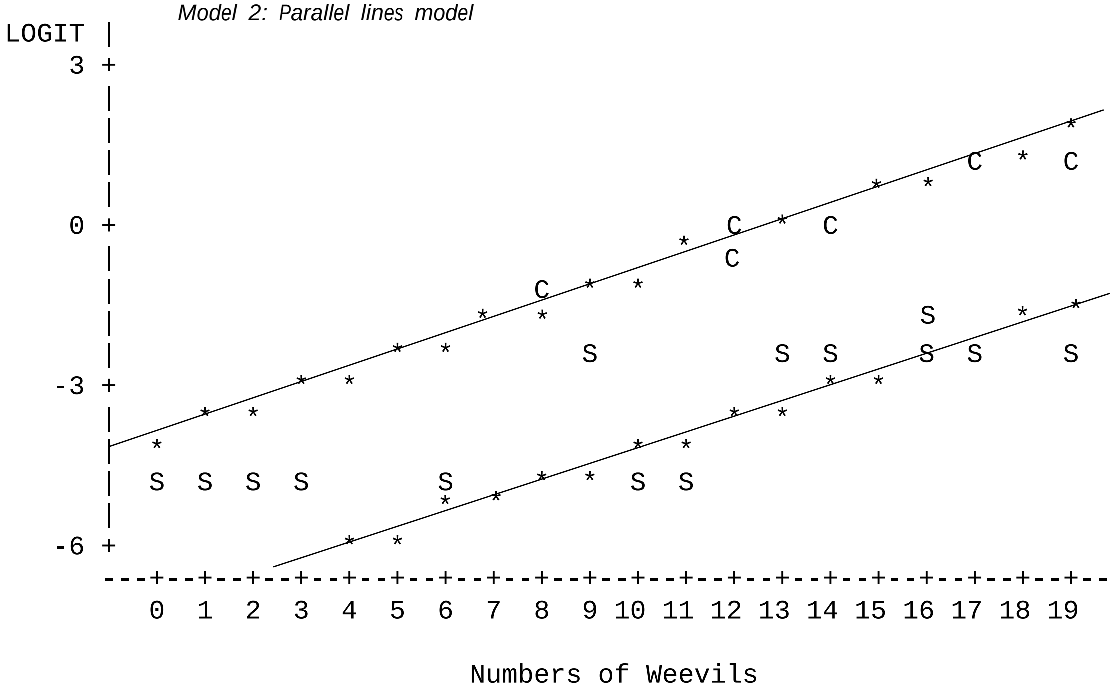
TABLE 24 *Sums of squares of the simple residuals for three models*

Model	Uncorrected sums of squares
Parallel lines	0.02877
Two radiating lines	0.03000
One line for control	0.03297

TABLE 25 *Three final models to fit the root collar weevil data*

Parallel lines model (Model 2)	
Control:	logit = –4.1217 + 0.3086 * Weevil (SE = 0.61) (SE = 0.047)
Screefed:	logit = –7.5321 + 0.3086 * Weevil (SE = 0.61) (SE = 0.047)
Two radiating lines model (Model 6)	
Control:	logit = –4.8370 + 0.3335 * Weevil (SE = 0.64) (SE = 0.049)
Screefed:	logit = –4.8370 + 0.1358 * Weevil (SE = 0.64) (SE = 0.048)
One weevil variable: for control only (Model 8)	
Control:	logit = –3.6086 + 0.2721 * Weevil (SE = 0.34) (SE = 0.028)
Screefed:	logit = –3.6086 + 0.0000 * Weevil (SE = 0.34) (SE = 0.0 since model restricted to this value)

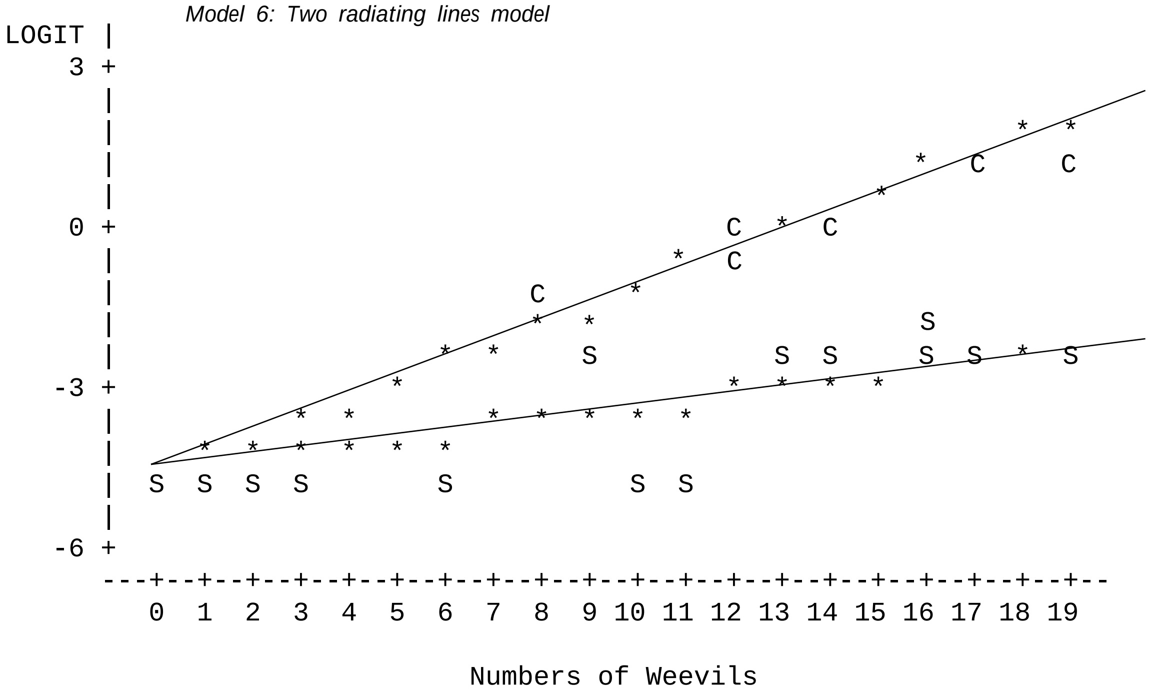
```
Plot of LOGIT*WEEVIL.   Symbol is value of TRMT.
Plot of LOGIT1*WEEVIL.  Symbol used is '*'.
```



```
LOGIT |
    3 +
      |
      |                                                           *
      |                                                     C  *  C
      |                                               *  *
    0 +                                      C  *  C
      |                                   *  C
      |                          C  *  *
      |                       *  *                       S     *  *
      |                 *  *        S           S  S     S  S     S
   -3 +           *  *                             *  *
      |     *  *                             *  *
      |  *                             *  *
      |  S  S  S  S        S     *  *  S  S
      |                    *  *
   -6 +              *  *
      ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
         0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                          Numbers of Weevils
```

NOTE: 19 obs had missing values. 20 obs hidden. 4 obs out of range.



```
LOGIT |
    3 +
      |
      |                                                        *  *
      |                                                  *  C     C
      |                                               *
    0 +                                      C  *  C
      |                                   *  C
      |                          C     *
      |                          *  *                    S
      |                    *  *     S           S  S     S  S  *  S
   -3 +                 *                    *  *  *  *
      |           *  *        *  *  *  *  *
      |     *  *  *  *  *  *
      |  S  S  S  S        S           S  S
      |
   -6 +
      ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
         0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                          Numbers of Weevils
```

NOTE: 19 obs had missing values. 25 obs hidden.

FIGURE 26 *Printer plots of observed (C or S) and fitted (*) values on the logit scale. Zero counts are converted to logits by the empirical logit = log(0.01/1.01) = –4.62. (Lines are hand-drawn.)*

```
     Plot of PROP*WEEVIL.   Symbol is value of TRMT.
     Plot of PRED1*WEEVIL.  Symbol used is '*'.
```

Model 2: Parallel lines model

```
Observed proportion
        |
    1.0 +
        |
        |                                                        *
    0.8 +                                                     *  C
        |                                                  C
        |                                               *
    0.6 +                                            *
        |                                         C
        |                                   C  *
    0.4 +                                   C
        |                                *
        |                             *
    0.2 +                       C  *
        |                    *  *                             S     *  *
        |           *  *  *  *        S           S  S  *  S  S     S
    0.0 +  S  S  S  S  *  *  S  *  *  *  S  S  *  *
        ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
           0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                             Numbers of Weevils

NOTE: 19 obs had missing values.  31 obs hidden.
```

Model 6: Two radiating lines model

```
Observed proportion
        |
    1.0 +
        |
        |                                                     *  *
    0.8 +                                                  *     C
        |                                               *  C
        |                                            *
    0.6 +
        |                                         C
        |                                   C  *
    0.4 +                                   C
        |                                *
        |
    0.2 +                       C  *  *
        |                       *                             S
        |              *  *  *     S     *  *  S  S  *  S  S  *  S
    0.0 +  S  S  S  S  *  *  S  *  *  *  S  S
        ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
           0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                             Numbers of Weevils
NOTE: 19 obs had missing values.  33 obs hidden.
```

FIGURE 27 *Printer plots of data (C or S) and predicted (*) values on the probability scale.*

```
           Plot of PROP*WEEVIL.   Symbol is value of TRMT.
           Plot of PRED3*WEEVIL.  Symbo use is '*'.

          Model 8: One line only for control plots
Observed proportion
      |
  1.0 +
      |
      |
  0.8 +                                                        *  C
      |                                                     C
      |                                                  *
  0.6 +                                               *
      |                                            C
      |                                      C  *
  0.4 +                                      C
      |                                   *
      |                             *  *
  0.2 +                          C
      |                    *  *                          S
      |     *  *  *  *  *           S           S  S     S  S     S
  0.0 +  S  S  S  S  *  *  S  *  *  *  S  S  *  *  *  *  *  *  *  *
      ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
         0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                             Numbers of Weevils
      NOTE: 19 obs had missing values.  28 obs hidden.
```

```
          Plot of LOGIT*WEEVIL.   Symbol is value of TRMT.
          Plot of LOGIT3*WEEVIL.  Symbo use is '*'.

          Model 8: One line only for control plots
LOGIT |
    3 +
      |
      |                                                           *
      |                                                     C  *  C
      |                                               *  *
    0 +                                      C  *  C
      |                                *  *  C
      |                          C  *
      |                    *  *                          S
      |              *  *           S           S  S     S  S     S
   -3 +        *  *
      |  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
      |
      |  S  S  S  S        S           S  S
      |
   -6 +
      ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
         0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                             Numbers of Weevils
      NOTE: 19 obs had missing values.  21 obs hidden.
```

FIGURE 28 *Printer plots of the observed (C or S) and fitted values (*) on both the probability and logit scales for the one line only for control plots model.*

```
           Plot of RESID1*PLOT.  Legend: A = 1 obs, B = 2 obs, etc.

           Model 2: Parallel lines model
  0.05 +                 A         A
       |       A         A         A    A
       |                                     B
       |                      A                   A
  0.00 +------------A-------------------A-------------------A----A----A----A--
       |  A         A         A                                       A
RESID1 |  A    A                                  A         A    A         A
       |                                               A
 -0.05 +
       |
       |
       |
 -0.10 +                                               A
       ---+----+----+----+----+----+----+----+----+----+----+----+----+----+--
          1    2    3    4    5    6    7    8    9   10   11   12   13   14

           Model 6: Two radiating lines model
   0.1 +
       |
       |                 A                   A
RESID2 |                 A         A    A
       |                           A
       |       A                             A
   0.0 +----------------------B-------------------A---------A---------A-------
       |  A    A                        A         A         A    B    A    B
       |  A         A                                  A
       |            A
       |
       |                                               A
  -0.1 +
       ---+----+----+----+----+----+----+----+----+----+----+----+----+----+--
          1    2    3    4    5    6    7    8    9   10   11   12   13   14

           Model 8: One line only for control plots
  0.10 +                 A
       |
       |
       |
  0.05 +
       |       A              A    A         A    A    A
RESID3 |                 A         A    A
       |            A         A                   A
  0.00 +-------------------------------------A--------------------------------
       |                                               A
       |  A         A                   A                   A    A    B    A
       |  A    A                                                 A         A
 -0.05 +                                                    A
       ---+----+----+----+----+----+----+----+----+----+----+----+----+----+--
          1    2    3    4    5    6    7    8    9   10   11   12   13   14

                                     Plot Number
```

FIGURE 29 *Printer plots of simple residuals against the plot number for three models.*

```
            Plot of RESID1*WEEVIL.  Symbol is value of TRMT.

           Model 2: Parallel lines model
  0.05 +                             S                    S
       |                                      C  S
       |                          C                 S
       |                                            C
  0.00 +--S--S--S--S--------S--------------------------------C--------
       |  C                             S  S              S
RESID1 |     C  C                             C              S
       |                                                           C
 -0.05 +
       |
       |
       |
 -0.10 +                                                           S
       ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
          0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

           Model 6: Two radiating lines model
   0.1 +
       |
       |                          C                       S
RESID2 |                                      C
       |                             S
       |                                         S  S
   0.0 +--S-----------------------------------------C-----S-----------
       |     S  S  S        S                 C              S
       |                                S  S                       S
       |                                                     C
       |
       |                                                           C
  -0.1 +
       ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
          0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

           Model 8: One line only for control plots
  0.10 +                                                  S
       |
       |
       |
  0.05 +
       |                             S           S  S     S  S     S
RESID3 |                                      C
       |                                            C        C
  0.00 +--------------------------C-----------------------------------
       |                                                           C
       |  S  S  S  S        S           S  S
       |     C                                C
 -0.05 +        C
       ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
          0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19

                              Numbers of Weevils
```

FIGURE 30 *Printer plots of simple residuals against the weevil count for three models.*

Although several computer packages are available for logistic regression computations, PROC CATMOD and PROC LOGISTIC[20] in SAS are presently the most accessible programs for Ministry staff. This chapter describes and contrasts these two procedures and then goes on to provide a detailed discussion of example programs used to analyze the five studies in Chapter 4. Note that the following discussion is based on Version 6.10 of SAS (SAS Institute 1989). Minor differences exist between the various versions.

5.1 Using **PROC CATMOD** and **PROC LOGISTIC** For Logistic Regression

The differences between PROC LOGISTIC and PROC CATMOD are similar to the differences between PROC GLM and PROC REG. For instance, both PROC GLM and PROC CATMOD are designed to fit models, such as analysis of variance and covariance, that primarily include categorical variables, but may also include covariates or continuous variables. PROC GLM uses a CLASS statement to define the categorical variables, while PROC CATMOD simply assumes that all variables are categorical unless otherwise specified by a DIRECT statement. On the other hand, PROC REG and PROC LOGISTIC are primarily regression procedures and do not directly include categorical variables in the models. To include such variables, they must be converted to indicator or dummy variables in an earlier data step and then included in the model (see section 6 for an advanced discussion of this topic). PROC GLM and PROC CATMOD create these variables in a hidden manner and users may not even realize that the conversion has taken place for model fitting. Note that these two procedures create the indicator variables differently. This is important when examining parameter estimates and defining contrasts. Defining contrast coefficients for PROC CATMOD is discussed in section 5.1.1.1.

The order of the levels for the response variable is presented early in the computer output for both the CATMOD and LOGISTIC procedures and both model the probabilities for the first level that is printed. If the procedure is not fitting the preferred response value, the parameter estimates will simply be different in sign. All the tests and criterion values will be the same. There are various ways to choose the response value to be fitted. See the appropriate SAS manuals for some options.

PROC REG and PROC LOGISTIC have many features that make them valuable for the traditional regression model-building exercise. For instance, both procedures will:

- fit either the model specified in the MODEL statement or select variables from those listed in the MODEL statement to test for inclusion in the model;
- allow forward, backward, or stepwise selection of model variables;
- help to identify if the models provide reasonable fits and to check for unusual points in the data using various regression diagnostics;

[20] PROC GENMOD is another SAS procedure that may be used.

- produce data sets including predicted and residual values (OUTPUT statement) or the estimated parameters (OUTEST option); and
- determine predicted values for observations with missing response values (note that PROC CATMOD will not do this because such observations are deleted from the analysis).

PROC LOGISTIC treats each observation in the data as an experimental unit. On the other hand, PROC CATMOD may lump observations together, so the POPULATION statement should be used to correctly identify the experimental units and keep them properly separated.

For response variables with just two levels, PROC CATMOD has only one way to specify the response, while PROC LOGISTIC has two. In this case, two variables may be used on the left side of the model statement, where one variable contains the count of successes or events and the other contains the total count. This feature avoids the WEIGHT statement that is often necessary with PROC CATMOD (in fact the WEIGHT statement should be avoided with PROC LOGISTIC). See the examples in sections 5.2, 5.3, and 5.4.

The two procedures also produce slightly different test statistics. The likelihood ratio output at the bottom of the analysis of variance table from PROC CATMOD compares the current model with a saturated model (each “population” or experimental unit fit individually). PROC LOGISTIC does not test this value, but generates an overall test for the components of the model (denoted by -2 LOG L under the Chi-Square for Covariates column) by comparing the current model to the intercept-only model.

Both procedures will output residuals, but do so differently. PROC CATMOD will produce simple residuals, two for each experimental unit: one for the success response and one for the failure response. On the other hand, PROC LOGISTIC will produce either or both of the other kinds of residuals: the deviance or Pearson residuals. Only one residual is output for the success response for each experimental unit.

5.1.1 PROC CATMOD Data for this procedure are organized in one of two ways. First, each observation can represent a separate sampling unit with an associated response variable. Groupings of observations into experimental units are specified with a POPULATION statement. The data must be organized this way if $m = 1$ because each experimental unit contains only one sampling unit (as in the herbicide thinning trial example, section 4.4). When $m > 1$ (i.e., each experimental unit contains several sampling units), then the data are either organized in this way or the response for each experimental unit is counted. The second organizational procedure requires two observations in the SAS data set for each experimental unit; one observation specifies the number of failures, while the second specifies the number of successes. PROC CATMOD uses both observations to determine the total number of sampling units in each experimental unit. If a count is zero, then that observation is unnecessary and is ignored by PROC CATMOD, although no problem is created by including the observation. The example programs in this chapter use the variable COUNT to indicate the number. The value of Y indicates whether the count represents success or failure. If Y is zero or one, then the parameter estimates calculated by PROC CATMOD will predict the logits and probability values for Y = 0. Thus the models generated by the programs in this chapter will predict collared calf mortality rate, seedling survival rate, mortality rate of herbicide treated trees, and weevil attack probability.

While PROC CATMOD is well documented in the various editions of the SAS User's Guide, the procedure is so versatile that the discussion relevant to logistic regression is scattered. The following SAS statements are useful for logistic regression analysis.

- POPULATION: defines the experimental units of the study. If this statement is missing, all experimental units with the same variable values in the MODEL statement are grouped together as if they were just one experimental unit. This makes it impossible to compare models when an explanatory variable is completely missing in one of them.
- WEIGHT: denotes the variable that specifies the number of sampling units within an experimental unit which either failed or succeeded.
- DIRECT: denotes those model variables to be treated as continuous variables. Variables not noted here are treated as categorical variables. Continuous variables will be tested with only one degree of freedom, while categorical variables will have degrees of freedom equal to the number of categories minus one.
- MODEL: describes the model to be fitted to the data. Models can be described in exactly the same manner as for PROC's ANOVA and GLM except that the shorthand notation using vertical bars to show interactions is not available in Version 5. If no variables are listed, then a mean or intercept-only model is fit. Some of the useful options for this statement are:

  ML: estimates the model using maximum likelihood methods. This is the default method (for Version 6) and so does not need to be specified.

  PRED=FREQ: prints out the predicted frequencies for each experimental unit (can be a long printout). [You may use only one of the PRED= options.]

  PRED=PROB: prints out the predicted probabilities for each experimental unit (can be a long printout).

  NOPROFILE: suppresses printing of the population and response profiles (may reduce length of printout).

  NOITER: suppresses printing of the iteration history required to converge to an adequate fit.

  COVB: prints out the covariance matrix of the parameter estimates.

  CORRB: prints out the correlation matrix of the parameter estimates.
- CONTRAST: specifies that a specific contrast for a categorical variable be tested. Proper use of this statement requires some understanding of the design matrix that PROC CATMOD uses. This is described in the manuals. See section 5.1.1.1 for more discussion.
- RESPONSE LOGIT/OUT = data name: outputs a data set containing predicted values and simple residuals (two for each experimental unit, one for success and one for failure).

- RESPONSE LOGIT/OUTEST = data name: outputs a data set containing fitted parameter estimates and their variance-covariance matrix.

5.1.1.1 ***Determining Contrast Coefficients for*** **PROC CATMOD** Contrasts of predicted probabilities of logistic regression models can be designed and tested in the same way as contrasts of predicted means in ANOVA. Contrast coefficients designed for ANOVA (and PROC GLM in particular)[21] must be converted to those necessary for PROC CATMOD. This is because PROC's CATMOD and GLM establish indicator variables for categorical variables differently (see Chapter 6 for a discussion of this point). A simple method to convert contrast coefficients is to subtract the last contrast coefficient from all the other coefficients. Thus, the last coefficient becomes zero and is deleted from the list of coefficients in the CONTRAST statement for PROC CATMOD. In this case, the coefficients may not add up to zero. The list of coefficients can be divided or multiplied by a constant.

The following are some common example contrast coefficients:

Treatment level:	1	2	3	4	5
Linear contrast:					
ANOVA:	–2	–1	0	1	2
CATMOD:	–4	–3	–2	–1	[a]
Quadratic contrast:					
ANOVA:	2	–1	–2	–1	2
CATMOD:	0	–3	–4	–3	[a]

[a] No value is used for the last treatment level.

The corresponding CONTRAST statements are:

Linear Contrast:

```
PROC GLM:     Contrast ‘Treatment: Linear’ Treat -2 -1  0  1  2;
PROC CATMOD:  Contrast ‘Treatment: Linear’ Treat -4 -3 -2 -1   ;
```

Quadratic Contrast:

```
PROC GLM:     Contrast ‘Treatment: Quad’ Treat    2 -1 -2 -1  2;
PROC CATMOD:  Contrast ‘Treatment: Quad’ Treat    0 -3 -4 -3   ;
```

The contrast coefficients for the levels of the treatments *must be* in the order assumed by the procedure. This order can be verified by examining the POPULATION PROFILES.

5.1.2 PROC LOGISTIC There are two ways to define the response variable for PROC LOGISTIC. When experimental units contain $m > 1$

[21] Determining contrast coefficients for ANOVAs are described in most textbooks on ANOVA and in Biometrics Information Pamphlets Nos. 12, 14, 16, and 23.

sampling units, it is easiest to use two variables in ratio form as the response variable. For example, if the variable total includes the number of sampling units (m_i), and alive is the variable that represents the number of sampling units still alive at the end of the study, then their ratio alive/total = . . . would be used in the model statement (see example in section 5.2). The response indicated by the variable, alive, is called the EVENT on the printout. On the other hand, if $m = 1$ and Y is the variable that represents 0 or 1 (depending on the unit's response), then a model statement that uses Y = . . . would be easier.

The following basic statements are used with PROC LOGISTIC to perform logistic regression analysis. The WEIGHT statement is used differently with PROC LOGISTIC and should be avoided because many of the statistical tests will be incorrect (SAS Institute 1989:1086).

- MODEL: describes the model to be fitted to the data. Models can be described by listing possible explanatory variables in this statement. If no variables are listed, then an intercept-only model will be fitted. Interactions between variables are not allowed; combination variables must be created in an earlier data step. Only one model statement is allowed for each PROC step. Some useful options for this statement are: INFLUENCE and IPLOTS produce useful influence diagnostics and plots. Discussing their output is beyond the scope of this text. Discussions can be found in Hosmer and Lemeshow (1989) and McCullagh and Nelder (1989), Agresti (1996).
- OUTPUT OUT = data name: produces a data set that includes the original data plus other variables as chosen by keywords. Some of the more useful keywords are:
  RESDEV = var name: outputs one deviance residual for each observation or experimental unit.
  RESCHI = var name: outputs one Pearson residual for each observation or experimental unit.
  PRED = var name: outputs predicted probabilities for each observation.
  LOWER or L = var name: outputs the lower confidence limit for the predicted probability of an event response.
  UPPER or U = var name: outputs the upper confidence limit for the predicted probability of an event response.
  ALPHA = is an option used to select the confidence level for the confidence limits output by LOWER and UPPER. The default level is ALPHA = 0.05. This option must be written after an "/" in the OUTPUT statement.

5.2 Simple Regression: Caribou Calf and Wolf Predation

Recall that the objective of this hypothetical study was to examine the relationship between wolf presence and the survival of caribou calves during their first year. The primary hypothesis of this example was that fewer calves will survive if high numbers of wolves are in their vicinity (see section 4.1 for full details).

5.2.1 Initial data input and examination The following program reads in the data from Appendix 2 (Calf Data) and prints out the data in Table 26. The format `alive` describes the meaning of the zeros and ones for the response variable *y*. To examine the data, the observed proportions are plotted against the wolf density estimate in Figure 7 (page 23). SAS statements in boldface type are essential for the program to do the correct calculations; the other statements are used to improve the appearance of the output and provide comments.

```
title ‘Simple Regression Example- Calf Survival’;
proc format; value alive 0 = ‘0␣␣Died’ 1 = ‘1␣Alive’; run;
data calf;
  infile ‘calf.dat’;
  input herd wolf alive total ;
  count = alive;        y = 1; prop = count/total; output;
  count = total-alive;  y = 0; prop = count/total; output;
label   y = ‘Survived or not’
     wolf = ‘Wolf Density (Numbers per 1000 km2)’
    total = ‘Total Number’    alive = ‘Number Survived’
    count = ‘Count’  herd = ‘Herd Number’   prop = ‘Proportion’;
  format y alive.;
run;
proc print data=calf label;
  id herd;  var wolf total count y prop;
title2 ‘Listing of data’;               * <== Similar to Table 26;
run;
proc plot data = calf;
  where y = 1;     * <== to analyse only the survival proportions;
  plot prop*wolf = herd / haxis = 0 to 40 by 5;*<== Similar to Figure 7;
title2 ‘Plot of the Proportion Surviving against Measure of Wolf Presence’;
run;
```

TABLE 26 *Listing of the data, with two observations per experimental unit (from* PROC CATMOD*)*

```
----------------------------------------------------------------------
                 Simple Regression - Calf Survival
                          Listing of data

         Wolf Density
  Herd   (Numbers per    Total             Survived
Number     1000 km2)    Number    Count     or not     Proportion

   1           9          15        14     1_Alive       0.93333
   1           9          15         1     0__Died       0.06667
   2          10           7         7     1_Alive       1.00000
   2          10           7         0     0__Died       0.00000
   3          12           4         3     1_Alive       0.75000
   3          12           4         1     0__Died       0.25000
   4          13           5         5     1_Alive       1.00000
   4          13           5         0     0__Died       0.00000
   5          15          10         9     1_Alive       0.90000
   5          15          10         1     0__Died       0.10000
   6          23          10         9     1_Alive       0.90000
   6          23          10         1     0__Died       0.10000
   7          31          15         9     1_Alive       0.60000
   7          31          15         6     0__Died       0.40000
   8          34          13         4     1_Alive       0.30769
   8          34          13         9     0__Died       0.69231
   9          38          13         1     1_Alive       0.07692
   9          38          13        12     0__Died       0.92308
----------------------------------------------------------------------
```

5.2.2 Simple and quadratic regression models Based on Figure 7, it seems reasonable that the decreasing trend of the observed proportions could be well fit by a linear or quadratic regression. The following program uses PROC REG to fit both these models to the data. Essential statements appear in boldface type. The computer output is shown in Figure 31.

```
proc reg data=calf;
  where y = 1;     * <== to analyse only the survival proportions;
  model prop = wolf;
title1 ‘Simple Regression Analysis - Linear Fit’;
run;
data calf; set calf; wolf2 = wolf * wolf; run;
proc reg data=calf;
  where y = 1;     * <== to analyse only the survival proportions;
  model prop = wolf wolf2;
title1 ‘Simple Regression Analysis - Quadratic Fit’;
run;
```

```
                   Simple Regression Model - Linear Fit
                           Analysis of Variance

                           Sum of         Mean
  Source          DF      Squares       Square      F Value       Prob>F

  Model            1      0.69095      0.69095       27.640       0.0012
  Error            7      0.17499      0.02500
  C Total          8      0.86594

      Root MSE       0.15811     R-square       0.7979
      Dep Mean       0.71866     Adj R-sq       0.7691
      C.V.          22.00031

                            Parameter Estimates

                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

INTERCEP   1      1.257311    0.11521578        10.913        0.0001
WOLF       1     -0.026205    0.00498432        -5.257        0.0012


                Simple Regression Analysis - Quadratic Fit
                           Analysis of Variance

                           Sum of         Mean
  Source          DF      Squares       Square      F Value       Prob>F

  Model            2      0.80556      0.40278       40.021       0.0003
  Error            6      0.06038      0.01006
  C Total          8      0.86594

      Root MSE       0.10032     R-square       0.9303
      Dep Mean       0.71866     Adj R-sq       0.9070
      C.V.          13.95931

                            Parameter Estimates

                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

INTERCEP   1      0.562394    0.21852403         2.574        0.0421
WOLF       1      0.051475    0.02323602         2.215        0.0686
WOLF2      1     -0.001688    0.00050016        -3.374        0.0150
```

FIGURE 31 *Output from the linear and quadratic regression models.*

5.2.3 Logistic regression analysis using PROC CATMOD and PROC LOGISTIC SAS can fit a simple logistic regression to this data using the maximum likelihood fitting methods available in both PROC CATMOD and LOGISTIC.

5.2.3.1 PROC CATMOD The following SAS code fits the logistic regression using PROC CATMOD. Essential statements appear in boldface type.

```
proc catmod data=calf;
  population herd; * <== identifies herd as the experimental/observational units;
  weight count;    * <== the variable that contains the counts per response;
  direct wolf;     * <== identifies wolf as a continuous and not a class variable;
  model y = wolf;
title1 ‘Simple Logistic Regression Model’;
run;
```

The first part of output is:

```
                        CATMOD PROCEDURE

Response: Y                           Response Levels (R)=     2
Weight Variable: COUNT                Populations     (S)=     9
Data Set: CALF                        Total Frequency (N)=    92
Frequency Missing: 0                  Observations  (Obs)=    16
```

Note that the number of observations is 16 and not 18, because two of the observations had zero counts. PROC CATMOD deletes these observations from the analysis. It would also delete any observations with missing values for the response variable (Y in this case).

```
  POPULATION PROFILES
                Sample
Sample  HERD     Size
----------------------
    1    1         15
    2    2          7
    3    3          4
    4    4          5
    5    5         10
    6    6         10
    7    7         15
    8    8         13
    9    9         13
```

```
RESPONSE PROFILES

Response     Y
-----------------
     1    0__Died
     2    1_Alive
```

Basic information about the data to be analyzed is provided in the preceding output tables. The nine populations (experimental/observational units) are described in the `POPULATION PROFILES` section, while the values of the response variable are displayed in the `RESPONSE PROFILES` section. This allows us to check that the experimental units and response variable are correctly defined. The next part of the output shows the iteration history for the parameter estimates.

```
                    MAXIMUM-LIKELIHOOD ANALYSIS

             Sub        -2 Log     Convergence    Parameter Estimates
Iteration   Iteration   Likelihood    Criterion         1           2
------------------------------------------------------------------------
     0          0       127.53908       1.0000            0           0
     1          0       80.891384       0.3658      -3.0940      0.1070
     2          0       76.384036       0.0557      -4.4174      0.1485
     3          0        75.92853     0.005963      -5.0005      0.1663
     4          0       75.919606     0.000118      -5.0971      0.1692
     5          0       75.919601    6.1244E-8      -5.0993      0.1693
     6          0       75.919601    1.741E-14      -5.0993      0.1693
```

The following output summarizes the fit of the simple logistic regression model to the data. The sample sizes are not large enough to use the `LIKELIHOOD RATIO` as a goodness-of-fit test. Both the `INTERCEPT` and `WOLF` Wald χ^2-statistics (denoted by `Chi-square` below) are significant, which suggests that neither the intercept nor the slope for wolf is zero.

```
           MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

         Source                  DF   Chi-Square      Prob
         --------------------------------------------------
         INTERCEPT                1        22.81    0.0000
         WOLF                     1        23.09    0.0000

         LIKELIHOOD RATIO         7         7.78    0.3524
```

```
              ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                             Standard    Chi-
   Effect            Parameter  Estimate     Error    Square   Prob
   ----------------------------------------------------------------
   INTERCEPT                 1   -5.0993    1.0678     22.81  0.0000
   WOLF                      2    0.1693    0.0352     23.09  0.0000
```

PROC CATMOD can also output predicted and residual values which can be tabulated and plotted. The programming requires the addition of the response statement:

```
proc catmod data=calf;
  population herd; * <== identifies herd as the experimental/observational units;
  weight count;    * <== the variable that contains the counts per response;
  direct wolf;     * <== identifies wolf as a continuous and not a class variable;
  model y = wolf / ml;
  response logit / out = pred;  * <== needed to output predicted values;
title1 ‘Simple Logistic Regression Model’;
run;
proc print;
title2 ‘Listing of output data set’;
run;
```

This statement causes the predicted values to be placed into a data set called pred. It also specifies that the procedure should use the logit link function (identified in the response statement by logit). The logit link function is the default link; therefore the response statement is not really necessary except that we want to use the out = option to create a data set with the predicted values. The printout of this new data set is shown on page 77.

SAS creates this data set by using some of the original variables and then adding special SAS variables. The variables with underscores at the beginning and end of the variable name, such as _SAMPLE_, _NUMBER_, . . . , _RESID_, are the special SAS variables. In this case, the sample or experimental unit number (_SAMPLE_) has the same value as HERD. The variable _TYPE_ describes the type of observation. For logistic regression models there are two types: FUNCTION to indicate the predicted logit of 0__Died (the first response value listed in the RESPONSE PROFILE table) and PROB to indicate the predicted probabilities. For each population/experimental unit or _SAMPLE_ number there are two observations with _TYPE_ = PROB. These are separated by the _NUMBER_ variable whose values are shown in the earlier RESPONSE PROFILES table. In this table, 1 is for 0__Died and 2 is for 1_Alive. The _OBS_ variable contains the observed logit or proportion and _SEOBS_ is its estimated standard error. The variable _PRED_ is the predicted logit or proportion, _SEPRED_ is its estimated standard error, and _RESID_ is the simple difference between _OBS_ and _PRED_. This difference does not have a constant variance and should be interpreted carefully. PROC LOGISTIC calculates residuals that take this into account (the deviance and chi-squared residuals, resdev and reschi, respectively).

To illustrate, let us look at the three lines output for the second herd with a wolf density of 10 wolves per 1000 km^2. This is herd 2 on the output. The predicted logit (for not surviving) is on the first line with _TYPE_ = FUNCTION. The value under the _PRED_ column is –3.40642 as we calculated earlier in section 4.1.3. The two lines with

```
                        Simple Logistic Regression Model
                           Listing of output data set

OBS HERD    Y    _SAMPLE_ _TYPE_   _NUMBER_  _OBS_   _SEOBS_   _PRED_ _SEPRED_  _RESID_

  1   1        .    1    FUNCTION    1   -2.63906 1.03510 -3.57571 0.76691  0.93665
  2   1  0__Died    1    PROB        1    0.06667 0.06441  0.02723 0.02032  0.03943
  3   1  1_Alive    1    PROB        2    0.93333 0.06441  0.97277 0.02032 -0.03943
  4   2        .    2    FUNCTION    1     .       .      -3.40642 0.73433   .
  5   2  0__Died    2    PROB        1    0.00000 0.00000  0.03210 0.02281 -0.03210
  6   2  1_Alive    2    PROB        2    1.00000 0.00000  0.96790 0.02281  0.03210
  7   3        .    3    FUNCTION    1   -1.09861 1.15470 -3.06784 0.66997  1.96923
  8   3  0__Died    3    PROB        1    0.25000 0.21651  0.04445 0.02846  0.20555
  9   3  1_Alive    3    PROB        2    0.75000 0.21651  0.95555 0.02846 -0.20555
 10   4        .    4    FUNCTION    1     .       .      -2.89855 0.63827   .
 11   4  0__Died    4    PROB        1    0.00000 0.00000  0.05223 0.03159 -0.05223
 12   4  1_Alive    4    PROB        2    1.00000 0.00000  0.94777 0.03159  0.05223
 13   5        .    5    FUNCTION    1   -2.19722 1.05409 -2.55997 0.57611  0.36275
 14   5  0__Died    5    PROB        1    0.10000 0.09487  0.07176 0.03837  0.02824
 15   5  1_Alive    5    PROB        2    0.90000 0.09487  0.92824 0.03837 -0.02824
 16   6        .    6    FUNCTION    1   -2.19722 1.05409 -1.20565 0.35953 -0.99157
 17   6  0__Died    6    PROB        1    0.10000 0.09487  0.23047 0.06376 -0.13047
 18   6  1_Alive    6    PROB        2    0.90000 0.09487  0.76953 0.06376  0.13047
 19   7        .    7    FUNCTION    1   -0.40547 0.52705  0.14867 0.29241 -0.55413
```

TYPE = PROB show a fitted probability of 0.03210 for 0__Died and a fitted probability of 0.96790 for 1_Alive which also matches our earlier calculations.

When this data set is merged with the original data set it becomes easier to plot and summarize the results. The following program will do this (essential statements appear in boldface type). The output in Table 27 lists the predicted values.

```
/* Note that version 6.04 will not add herd to the pred data set!  */
proc sort data = pred; by herd y; run;  * <== ensuring that both data sets are;
proc sort data = calf; by herd y; run;  * <== sorted the same way for merging;
data new;
  merge calf pred;
  by herd y;   * <==  matching variables;
  if y ne . ;  * <==  keeping the predicted probabilities only;
  keep y wolf herd count prop _obs_ _pred_ _resid_ total;
proc print data = new;                  *<==  Output is in Table 27;
title2 ‘Listing of final data set with predicted values’;
run;
proc plot data = new;
  plot _pred_*_obs_ ;                   *<== Output is in Figures 9, 10, and 11);
  plot _resid_*(_pred_ herd wolf) / vref = 0;
title2 ‘Plots’;
run;
```

Note that PROC CATMOD produces output for two observations for each experimental unit. This means that each experimental unit has two residuals: one for Alive and one for Died. This leads to a mirror effect in plots of the residuals. This can be eliminated by using a where _number_ = 1; statement, for instance, so that there is only one residual in the graph per observation. PROC LOGISTIC only produces one residual per experimental unit.

TABLE 27 *Listing of predicted values from* PROC CATMOD

```
------------------------------------------------------------------
                  Simple Regression - Calf Survival
             Listing of final data set with predicted values

OBS HERD WOLF TOTAL COUNT    Y      PROP   _OBS_   _PRED_  _RESID_

  1   1    9    15     1 0__Died 0.06667 0.06667 0.02723  0.03943
  2   1    9    15    14 1_Alive 0.93333 0.93333 0.97277 -0.03943
  3   2   10     7     0 0__Died 0.00000 0.00000 0.03210 -0.03210
  4   2   10     7     7 1_Alive 1.00000 1.00000 0.96790  0.03210
  5   3   12     4     1 0__Died 0.25000 0.25000 0.04445  0.20555
  6   3   12     4     3 1_Alive 0.75000 0.75000 0.95555 -0.20555
  7   4   13     5     0 0__Died 0.00000 0.00000 0.05223 -0.05223
  8   4   13     5     5 1_Alive 1.00000 1.00000 0.94777  0.05223
  9   5   15    10     1 0__Died 0.10000 0.10000 0.07176  0.02824
 10   5   15    10     9 1_Alive 0.90000 0.90000 0.92824 -0.02824
 11   6   23    10     1 0__Died 0.10000 0.10000 0.23047 -0.13047
 12   6   23    10     9 1_Alive 0.90000 0.90000 0.76953  0.13047
 13   7   31    15     6 0__Died 0.40000 0.40000 0.53710 -0.13710
 14   7   31    15     9 1_Alive 0.60000 0.60000 0.46290  0.13710
 15   8   34    13     9 0__Died 0.69231 0.69231 0.65848  0.03383
 16   8   34    13     4 1_Alive 0.30769 0.30769 0.34152 -0.03383
 17   9   38    13    12 0__Died 0.92308 0.92308 0.79145  0.13163
 18   9   38    13     1 1_Alive 0.07692 0.07692 0.20855 -0.13163
------------------------------------------------------------------
```

5.2.3.2 PROC LOGISTIC The following SAS code fits the logistic regression model using PROC LOGISTIC for the analysis. The data is read in again because we only need one record per experimental unit and not two as for PROC CATMOD. Essential statements appear in boldface type.

Note that the original input variables are used in the model statement: the variable alive is the count of surviving calves, while total is the total number of calves. PROC LOGISTIC is able to analyze these variables directly when the response variable has only two levels (alive or dead in this case). The output statement creates a data set that, in addition to all the variables in the original data set, will include the predicted values (named pred) and two types of residuals, resdev and reschi, described in section 3.7.

```
title ‘Simple Regression Example - Calf Survival’;
proc format; value alive 0 = ‘0 .. Died’ 1 = ‘1 ..Alive’; run;
data calfr;
  infile ‘calf.dat’;
  input herd wolf alive total ;
label wolf = ‘Wolf Density (Numbers per 1000 km2)’
     total = ‘Total Number’ alive = ‘Number Survived’
      herd = ‘Herd Number’ ;
run;
proc logistic;
  model alive/total = wolf ;
  output out=predr p=pred resdev=resdev reschi=reschi;
title2 ‘Logistic Regression Analysis’;
run;
proc print;
title3 ‘Listing of output data set’; run;
proc plot vpercent = 50;
  plot (resdev reschi)*(herd wolf) / vref = 0;
title2 ‘Plots’; run;
```

The first section of the output below summarizes the variables, the number of observations, and the levels of the response variable. The default link function is the logit, which means that a logistic regression is fitted to the data. The response level `EVENT` is defined by `ALIVE`; that is, for the event that a calf survived.

```
          Simple Regression - Calf Survival
             Logistic Regression Analysis

                The LOGISTIC Procedure

Data Set: WORK.CALF
Response Variable (Events): ALIVE     Number Survived
Response Variable (Trials): TOTAL     Total Number
Number of Observations: 9
Link Function: Logit

                   Response Profile

              Ordered  Binary
                Value  Outcome      Count

                    1  EVENT           61
                    2  NO EVENT        31
```

This next section of output provides various tests of the usefulness of the model variables for fitting the data. The line `–2 LOG L` provides the final values of the –2Log-likelihood from the iterative fitting process. The first value in this line (`Intercept Only`) is for a model with only one mean for all the data, while the second is for the full model (`Intercept and Covariates`). This second value is the same as that produced by the iteration summary from PROC CATMOD (see output in previous section). The difference between the first two values provides an overall test for the model's covariates. This is a more reliable test than the Wald test and is not output by PROC CATMOD. In this case, the test shows a highly significant result (p-value 0.0001). The other criterion measures (except Score) were discussed in section 3.4.

```
                   Testing Global Null Hypothesis: BETA=0

                               Intercept
                 Intercept        and
   Criterion       Only       Covariates    Chi-Square for Covariates

   AIC             119.575        79.920         .
   SC              122.097        84.963         .
   -2 LOG L        117.575        75.920       41.656 with 1 DF (p=0.0001)
   Score              .              .         35.896 with 1 DF (p=0.0001)
```

Wald statistics for the intercept and variables in the model are provided in the next section of output. These tests are identical to those produced by PROC CATMOD. Notice that although the parameter estimates have the same magnitude, they are different in sign. This is because PROC LOGISTIC has fitted the probability of survival, while PROC CATMOD fitted the probability of death.

```
                   Analysis of Maximum Likelihood Estimates

             Parameter Standard    Wald       Pr >    Standardized     Odds
 Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio

 INTERCPT 1     5.0993   1.0678    22.8077     0.0001            .  163.907
 WOLF     1    -0.1693   0.0352    23.0882     0.0001    -1.035538    0.844
```

The final predicted data is given below.

```
                        Listing of output data set

OBS HERD WOLF ALIVE TOTAL COUNT    Y      PROP    PRED   RESCHI   RESDEV

  1   1    9    14    15    14  1_Alive 0.93333 0.97277 -0.93834 -0.79489
  2   2   10     7     7     7  1_Alive 1.00000 0.96790  0.48179  0.67580
  3   3   12     3     4     3  1_Alive 0.75000 0.95555 -1.99463 -1.41449
  4   4   13     5     5     5  1_Alive 1.00000 0.94777  0.52489  0.73238

  5   5   15     9    10     9  1_Alive 0.90000 0.92824 -0.34602 -0.32798
  6   6   23     9    10     9  1_Alive 0.90000 0.76953  0.97970  1.07200
  7   7   31     9    15     9  1_Alive 0.60000 0.46290  1.06489  1.06436
  8   8   34     4    13     4  1_Alive 0.30769 0.34152 -0.25718 -0.25938
  9   9   38     1    13     1  1_Alive 0.07692 0.20855 -1.16815 -1.30289
 10  10   15     .     .     .  1_Alive  .      0.92824   .        .
 11  11   25     .     .     .  1_Alive  .      0.70414   .        .
```

Predicted values for specific wolf densities may be of interest, even if these values do not occur in the data. These values can be obtained by adding them to the original data set with missing response variables. The predicted data set output by PROC LOGISTIC will then contain their predicted values. As an example, two missing observations have been added to the end of the data set as herds 10 and 11. They have missing values for the response variables (alive and total). Note that they have been assigned predicted probabilities based on the fitted logistic regression model. Predicted values are easier to obtain from PROC LOGISTIC than from PROC CATMOD.

5.2.4 Calculating and fitting predicted values to compare the three models A comparison of the results of the linear, quadratic and logistic regression models was presented in section 4.1. The SAS code that was used to calculate the predicted values follows. New and essential SAS code appear in boldface type.

```
** First rerun the regressions and obtain output data sets;
proc reg data=calf;
  where y = 1;  model prop = wolf;
  output out = regpred1 p = regpred1 r = regresd1;
title2 ‘Simple Regression Analysis - Linear Fit’;
run;
proc reg data=calf;
  where y = 1;  model prop = wolf wolf2;
  output out = regpred2 p = regpred2 r = regresd2;
title2 ‘Simple Regression Analysis - Quadratic Fit’;
run;
** Sort the predicted data sets for proper merging;
proc sort data=regpred1; by herd y; run;
proc sort data=regpred2; by herd y; run;
proc sort data=pred    ; by herd y; run;    *<== Data was created previously;
**  Merge the three data sets.  Regpred1 has all the original data plus
    the new predicted values from the regression.     ;
data new;
  merge regpred1 regpred2 pred;
  by herd y;
  if y eq 1 ;  *<==  keeping the predicted probabilities for survival;
  keep y wolf wolf2 herd count prop
       _obs_ _pred_ _resid_ total regpred1 regpred2 regresd1 regresd2;
  label _pred_ =’Logistic Predicted Value’
        regpred1 = ‘Linear Predicted Value’
        regpred2 = ‘Quadratic Predicted Value’
        regresd1 =’Linear Regr Residual’
        regresd2 =’Quadratic Regr Residual’
        _resid_  =’Logistic Residual’;
run;
```

```
proc print data=new label;
title2 ‘Listing of final data set with predicted values’;
  id herd;                     *<== Output is in Table 28;
  var y wolf count _pred_ regpred1 regpred2;  * _resid_ regresd1 regresd2;
title2 ‘Listing of final data set with predicted values’;
run;
** Plot the data with all three models;
proc plot data = new;
  plot _obs_ * wolf = herd
       regpred1 * wolf = ‘*’
       regpred2 * wolf = ‘@’    /*<== Output for Figure 8 in section 4.1.4;*/
       _pred_ * wolf = ‘+’  /overlay;
title2 ‘Comparing the predicted values for all three models’;
run;
** Calculate the sums of squares of the differences between observed and
   predicted values;
proc means n mean uss;  *<== uss = uncorrected sums of squares;
  var regresd1 regresd2 _resid_;   *<== Output is in Table 29;
title2 ‘Residual Sums of Squares’;
run;
```

The predicted values are shown in Table 28 and the sums of squares of the simple residuals (that is, the difference between the observed and predicted values) are shown in Table 29.

TABLE 28 *Predicted values for the three different models fitted. Maximum values are bolded.*

```
 -----------------------------------------------------------------------------
                     Simple Regression - Calf Survival
               Listing of final data set with predicted values

                       Wolf Density            Logistic     Linear    Quadratic
 Herd     Survived    (Numbers per             Predicted   Predicted   Predicted
Number     or not      1000 km2)     Count      Value       Value       Value

   1     1 Alive          9            14      0.97277     1.02147     0.88896
   2     1 Alive         10             7      0.96790     0.99527     0.90837
   3     1 Alive         12             3      0.95555     0.94286     0.93706
   4     1 Alive         13             5      0.94777     0.91665     0.94634
   5     1 Alive         15             9      0.92824     0.86424     0.95477
   6     1 Alive         23             9      0.76953     0.65461     0.85349
   7     1 Alive         31             9      0.46290     0.44497     0.53617
   8     1 Alive         34             4      0.34152     0.36635     0.36148
   9     1 Alive         38             1      0.20855     0.26154     0.08130
 -----------------------------------------------------------------------------
```

TABLE 29 *Residual sums of squares for the three models is tabulated under the USS (for Uncorrected Sums of Squares) column*

```
-------------------------------------------------------------------------
 Variable   Label                   N          Mean            USS
-------------------------------------------------------------------------
 REGRESD1     Linear Regr Residual     9     -2.46716E-17        0.1749861
 REGRESD2     Quadratic Regr Residual  9     -6.74615E-19        0.0603847
 _RESID_      Logistic Residual        9       -0.0096426        0.1026476
-------------------------------------------------------------------------
```

5.2.5 Checking the fit of the logistic regression model The adequacy of the fit of the logistic regression can be examined by various plots. Besides checking the adequacy of the model, the plots help identify unusual data points that might need further investigation. The plot of predicted values versus observed values (see Figure 9, page 26) is reasonably straight with a slope near one (hand drawn on the figure). Since the predicted values should be about the same as the observed values, a slope of one is expected. The plots of the residuals are shown in Figures 10 and 11 (pages 29 and 30), output by `PROC CATMOD`. The four plots of residuals produced using `PROC LOGISTIC` are shown in Figures 12 and 13 (pages 28 and 29). All the plots show little pattern so we might decide that the model fits adequately and that none of the data points are particularly unusual.

5.3 One-way Classification Study: Survival and Age of Stands with Root Rot

In this example we will look at a simple one-way classification. Recall from the study description in section 4.2 that a researcher has searched an inventory database for suitable stands for this study. In this case, nine separate stands with similar levels of root rot infestation are chosen and sampled. This is another observational study and the case for any cause and effect relationships will accordingly be weak.

Section 5.3.1 describes the SAS code for the initial data input. The SAS programs required to accomplish the analysis of the three age group model are described in section 5.3.2. The programs used to test for the linear and quadratic trends in stand age are outlined in section 5.3.3. The model is then refitted with just two age groups (young and old) using the programs in section 5.3.4. Detailed output is included in all sections.

5.3.1 Initial data input The following program reads in the tree sampling data and produces Table 30, which shows frequency counts of dead and live trees. SAS statements appearing in boldface type are essential for the performance of the program. The other statements generally improve the appearance of the output.

```
title ‘Simple One-Way Classification Example’;
**  Reading in the data ;
proc format; value age 1 = ‘1..Young’ 2 = ‘2..Middle’ 3 = ‘3..Old’;
             value alive 0 = ‘0..Dead’ 1 = ‘1..Alive’; run;
data one;
  input dead alive @@;
  stand + 1;                             *<== Creating the stand number;
  age = 1 + (stand ge 3) + (stand ge 6); *<== Creating the age variable;
  total = dead + alive;
  count = dead;  y = 0; prop =  dead/total; output;
  count = alive; y = 1; prop = alive/total; output;
label age = ‘Age’ stand = ‘Stand Number’
      alive = ‘Alive’  dead = ‘Dead’
      count = ‘Count’  total = ‘Total Count’
      prop = ‘Observed Proportion’ y = ‘Survive?’ ;
format age age. y alive.;
cards;                  * The notes below cannot be in the SAS data lines;
13 28  8 27             *<==  Young stands;
10 18 15 22  6 10       *<==  Middle-aged stands;
 7  9  7 11 19 22  7  8 *<==  Old stands;
;
proc print data = one label;
  id age stand; var dead alive count y prop;
title2 ‘Listing of data’;      *<== the output is not shown;
proc freq data = one;
  weight count;                *<== Output is in Table 30;
  table  y * stand / norow nopercent;  * chisq;
title2 ‘Frequency Counts’;
run;
```

TABLE 30 *Counts of dead trees for the root rot and stand age example*

```
--------------------------------------------------------
Y(Survive?)     stand(Stand Number)

Frequency|  ----Young----     -------Middle-Aged-------
Col Pct  |       1|       2|       3|       4|       5|
---------+--------+--------+--------+--------+--------+
0..Dead  |     13 |      8 |     10 |     15 |      6 |
         |  31.71 |  22.86 |  35.71 |  40.54 |  37.50 |
---------+--------+--------+--------+--------+--------+
1..Alive |     28 |     27 |     18 |     22 |     10 |
         |  68.29 |  77.14 |  64.29 |  59.46 |  62.50 |
---------+--------+--------+--------+--------+--------+
Total          41       35       28       37       16

Frequency|  ------------Old Stands-----------
Col Pct  |       6|       7|       8|       9|  Total
---------+--------+--------+--------+--------+
0..Dead  |      7 |      7 |     19 |      7 |     92
         |  43.75 |  38.89 |  46.34 |  46.67 |
---------+--------+--------+--------+--------+
1..Alive |      9 |     11 |     22 |      8 |    155
         |  56.25 |  61.11 |  53.66 |  53.33 |
---------+--------+--------+--------+--------+
Total          16       18       41       15      247
--------------------------------------------------------
```

5.3.2 Logistic regression analysis of the three age group model The following code performs the analysis for the one-way classification for the three age groups. The `population` statement only needs to include the stand variable. However, when both age and stand variables are included they will appear in the `Population Profiles` and in any output data set specified by the `response` statement (for version 6.10 and later). This makes it easier to merge the predicted data set with the original data set when producing a final data set.

```
**  Three Group Model  ;
proc catmod data=one;
  population age stand;
  weight count;
  model y = age / ml;
  contrast ‘Young vs Midl & Old’  age -1 0;
  contrast ‘Middle vs Old’        age  1 2;
title2 ‘Three Group Analysis’;
run;
```

The first part of the output is shown below. It allows us to check that the data have been defined correctly.

```
                       Simple One-Way Classification Example
                                Three Group Analysis

                                  CATMOD PROCEDURE
          Response: Y                           Response Levels (R)=     2
          Weight Variable: COUNT                Populations     (S)=     9
          Data Set: ONE                         Total Frequency (N)=   247
          Frequency Missing: 0                  Observations  (Obs)=    18

                                POPULATION PROFILES
                                                    Sample
                         Sample    Age     Stand     Size
                         ----------------------------------
                             1   1..Young    1          41
                             2   1..Young    2          35
                             3   2..Middle   3          28
                             4   2..Middle   4          37
                             5   2..Middle   5          16
                             6   3..Old      6          16
                             7   3..Old      7          18
                             8   3..Old      8          41
                             9   3..Old      9          15


                                  RESPONSE PROFILES

                                 Response     Y
                                 ------------------
                                      1    0..Dead
                                      2    1..Alive
```

The next part of the output displays the iteration history for the fitting procedure. The last value of the –2 `Log Likelihood` can be used to check hand calculations in section 7.2.

```
                  Simple One-Way Classification Example
                           Three Group Analysis

                        MAXIMUM-LIKELIHOOD ANALYSIS

              Sub       -2 Log     Convergence      Parameter Estimates
Iteration  Iteration  Likelihood    Criterion      1         2         3
--------------------------------------------------------------------------
     0          0      342.41471       1.0000         0         0        0
     1          0      321.11188       0.0622   -0.5287   -0.3660   0.0596
     2          0      321.03923     0.000226   -0.5543   -0.4075   0.0763
     3          0      321.03921    4.5882E-8   -0.5547   -0.4081   0.0766
     4          0      321.03921    2.125E-15   -0.5547   -0.4081   0.0766
```

The rest of the output is shown in Figure 32.

```
          MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

        Source                   DF   Chi-Square      Prob
        --------------------------------------------------
        INTERCEPT                 1        16.98    0.0000
        AGE                       2         4.96    0.0836

        LIKELIHOOD RATIO          6         1.23    0.9755


             ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                        Standard    Chi-
Effect            Parameter  Estimate    Error    Square   Prob
----------------------------------------------------------------
INTERCEPT                 1   -0.5547    0.1346    16.98  0.0000
AGE                       2   -0.4081    0.2001     4.16  0.0414
                          3    0.0766    0.1885     0.17  0.6844


            CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

      Contrast                     DF   Chi-Square      Prob
      ------------------------------------------------------
      Young vs Midl & Old           1         4.16    0.0414
      Middle vs Old                 1         0.67    0.4137


            CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

      Contrast                     DF   Chi-Square      Prob
      ------------------------------------------------------
      Linear                        1         4.94    0.0263
      Quadratic                     1         0.17    0.6844
```

FIGURE 32 *Results of fitting the three groups model to the data, with two sets of contrasts.*

This program fits the one group or intercept only model and the saturated model. Essential statements appear in boldface type.

```
**  One Group Model  ;
proc catmod data=one;
  population age stand;
  weight count;
  model y =  ;
title2 ‘One Group Analysis’;
run;
**  Saturated Model  ;
proc catmod data=one;
  population age stand;
  weight count;
  model y =  stand;
title2 ‘Saturated Model’;
run;
```

The following part of the output is salient to the discussion in sections 4.2.2 and 4.2.3:

```
           Simple One-Way Classification Example
                     One Group Analysis

                      CATMOD PROCEDURE

                 MAXIMUM-LIKELIHOOD ANALYSIS

                                                   Parameter
               Sub        -2 Log     Convergence   Estimates
Iteration   Iteration   Likelihood    Criterion         1
------------------------------------------------------------
     0           0       342.41471       1.0000            0
     1           0       326.17462       0.0474      -0.5101
     2           0       326.16695    0.0000235      -0.5216
     3           0       326.16695     4.91E-11      -0.5216

           Simple One-Way Classification Example
                     One Group Analysis
       MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

     Source                    DF   Chi-Square      Prob
     --------------------------------------------------
     INTERCEPT                  1        15.71    0.0001

     LIKELIHOOD RATIO           8         6.36    0.6074
```

```
                   ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                              Standard    Chi-
       Effect            Parameter  Estimate    Error    Square   Prob
       ----------------------------------------------------------------
       INTERCEPT                 1   -0.5216    0.1316    15.71  0.0001
```

```
                    Simple One-Way Classification Example
                               Saturated Model

                               CATMOD PROCEDURE

                          MAXIMUM-LIKELIHOOD ANALYSIS

              Sub       -2 Log    Convergence     Parameter Estimates
Iteration  Iteration  Likelihood   Criterion      1        2        3
------------------------------------------------------------------------
     0          0      342.41471      1.0000         0        0        0
     1          0      319.93263      0.0657   -0.4713  -0.2604  -0.6145
     2          0        319.811    0.000380   -0.4938  -0.2732  -0.7183
     3          0      319.81088   3.5348E-7   -0.4943  -0.2730  -0.7221
     4          0      319.81088   4.701E-13   -0.4943  -0.2730  -0.7221

Parameter Estimates
Iteration      4          5          6          7          8           9
------------------------------------------------------------------------
     0           0          0          0          0          0         0
     1     -0.1002     0.0929    -0.0287     0.2213     0.0268    0.3249
     2     -0.0940     0.1108    -0.0170     0.2425     0.0418    0.3472
     3     -0.0935     0.1113    -0.0165     0.2430     0.0423    0.3477
     4     -0.0935     0.1113    -0.0165     0.2430     0.0423    0.3477

                MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

              Source                   DF   Chi-Square      Prob
              --------------------------------------------------
              INTERCEPT                 1        11.90    0.0006
              STAND                     8         6.00    0.6471

              LIKELIHOOD RATIO          0          .       .
```

5.3.3 Contrast analysis The objectives of the study will dictate which contrasts are of interest. The program for two of the contrasts was shown in section 5.3.2, and output shown in Figure 32. This section will fit linear and quadratic contrasts to examine a trend response with age. The test results are shown in the bottom half of the output in Figure 32. New statements appear in boldface type.

```
**  Three Group Model  ;
proc catmod data=one;
  population age stand;
  weight count;
  model y = age / ml;
  response logit / out = pred;
  contrast ‘Linear    ‘  age 2  1;
  contrast ‘Quadratic’  age 0 -3;
title2 ‘Three Group Analysis’;
run;
```

A more direct test for the linear effect of age can be obtained by running the above code again, but now specifying that age is a DIRECT effect so that age is treated as a continuous variable. The boldface SAS statement shows where the change in the program is required. Note that the `contrast` statements have been left out since they are only relevant for age when it is a categorical variable.

```
**  Three Group Model - Linear Effect ;
proc catmod data=one;
  population age stand;
  weight count;  direct age;
  model y = age / ml;     *<== Output is in Figure 33;
title2 ‘Three Group Analysis - Linear Effect’;
run;
```

The output is presented in Figure 33. Notice that because the parameter estimate for age is positive and the `CATMOD` procedure fits the probability of the lowest response level (in this case `0.. Dead`), the mortality probability increases with age.

```
           Simple One-Way Classification Example
            Three Group Analysis - Linear Effect

         MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                  DF   Chi-Square      Prob
       --------------------------------------------------
       INTERCEPT                1        11.75    0.0006
       AGE                      1         4.86    0.0274

       LIKELIHOOD RATIO         7         1.39    0.9858

            ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                               Standard    Chi-
Effect            Parameter  Estimate    Error    Square   Prob
----------------------------------------------------------------
INTERCEPT                 1   -1.2788    0.3730    11.75  0.0006
AGE                       2    0.3626    0.1644     4.86  0.0274
```

FIGURE 33 *Output for the model with age as a direct effect (continuous variable).*

5.3.4 Analysis of the two age group model The next step in our analysis is to fit the model with the two groups: younger stands and older stands. The following program includes the required SAS programming code. The middle- and old-aged stands are pooled into one new category by giving them the same formatted value defined in PROC FORMAT. The CATMOD output is presented in Figure 34. The predicted values are placed into a data set called pred and are shown in Table 31, while the residuals are plotted in Figure 14 (page 36). Note that the residuals and predicted values are output only for the dead response observations (by using a where statement when creating the NEW data set). New SAS codes appear in boldface type.

```
proc format; value age 1 = ‘1..Young’ 2, 3 = ‘Older’;
**  Two Group Model  ;
proc catmod data=one;
  population age stand;
  weight count;
  model y = age;
  response logit / out = pred;
  format age age.;
title2 ‘Two Group Analysis’;
run;
proc sort data=one;  by stand y; run;

**  Adding predicted values to original dataset ;
data new;
  merge one pred; by stand y;
  if _type_ ne ‘FUNCTION’ and y = 0; *<== Keeping the Dead Probabilities only;
label age = ‘Age’ stand = ‘Stand Number’
      prop = ‘Observed Proportion’ y = ‘Survive?’
     _obs_ = ‘Observed Proportion’ _pred_ = ‘Predicted Proportion’
   _resid_ = ‘Residual Proportion’;
  keep y age stand count _obs_ _pred_ _resid_ total;
run;
proc print data=new label;
  id age stand;
  var y count _obs_ _pred_ _resid_ total;
title2 ‘Listing of final dataset with predicted values’;
proc plot data=new;
  plot _pred_*_obs_;
  plot _resid_*(stand age) / vref = 0;
title2 ‘Plots’;
run;
```

```
                  Simple One-Way Classification Example
                           Two Group Analysis

                       CATMOD PROCEDURE

    Response: Y                           Response Levels (R)=     2
    Weight Variable: COUNT                Populations     (S)=     9
    Data Set: ONE                         Total Frequency (N)=   247
    Frequency Missing: 0                  Observations  (Obs)=    18

                           POPULATION PROFILES
                                               Sample
                    Sample    AGE     STAND     Size
                    ----------------------------------
                        1   1..Young    1         41
                        2   1..Young    2         35
                        3   Older       3         28
                        4   Older       4         37
                        5   Older       5         16
                        6   Older       6         16
                        7   Older       7         18
                        8   Older       8         41
                        9   Older       9         15

                            RESPONSE PROFILES

                           Response     Y
                           ------------------
                                1    0..Dead
                                2    1..Alive

                       MAXIMUM-LIKELIHOOD ANALYSIS

               Sub        -2 Log     Convergence    Parameter Estimates
Iteration   Iteration   Likelihood    Criterion         1           2
-------------------------------------------------------------------------
     0           0       342.41471       1.0000            0           0
     1           0       321.78082       0.0603      -0.6170     -0.2778
     2           0       321.70925     0.000222      -0.6522     -0.3097
     3           0       321.70923    4.5782E-8      -0.6527     -0.3102
     4           0       321.70923    2.297E-15      -0.6527     -0.3102
```

FIGURE 34 *Fit of the two group model.*

```
              Simple One-Way Classification Example
                       Two Group Analysis

            MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

          Source                  DF   Chi-Square      Prob
          --------------------------------------------------
          INTERCEPT                1        18.96    0.0000
          AGE                      1         4.28    0.0385

          LIKELIHOOD RATIO         7         1.90    0.9653


              ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                                 Standard    Chi-
   Effect           Parameter  Estimate    Error    Square   Prob
   ----------------------------------------------------------------
   INTERCEPT                1   -0.6527    0.1499    18.96  0.0000
   AGE                      2   -0.3102    0.1499     4.28  0.0385
```

FIGURE 34 *(continued)*

TABLE 31 *Listing of the predicted proportions and simple residuals from the two group model*

```
---------------------------------------------------------------------
                Simple One-Way Classification Example
                         Two Group Analysis
           Listing of final dataset with predicted values

          Stand                Observed   Predicted  Residual  Total
Age      Number Survive? Count Proportion Proportion Proportion Count

1..Young    1   0..Dead    13    0.31707    0.27632    0.040757   41
1..Young    2   0..Dead     8    0.22857    0.27632   -0.047744   35
Older       3   0..Dead    10    0.35714    0.41520   -0.058062   28
Older       4   0..Dead    15    0.40541    0.41520   -0.009799   37
Older       5   0..Dead     6    0.37500    0.41520   -0.040205   16
Older       6   0..Dead     7    0.43750    0.41520    0.022295   16
Older       7   0..Dead     7    0.38889    0.41520   -0.026316   18
Older       8   0..Dead    19    0.46341    0.41520    0.048210   41
Older       9   0..Dead     7    0.46667    0.41520    0.051462   15
---------------------------------------------------------------------
```

5.4 One-way Classification Study: Fertilizer Trial

Recall from section 4.3 that this one-way classification study consists of twenty-four rows, each with ten seedlings, which were randomly assigned fertilizer treatments of 0, 100, 200, and 300 kg/ha. In all other respects, the seedlings are treated similarly. The seedling’s survival is assessed after the first growing season. The SAS code for the initial data input and contingency table output appears in the next section. The one-way classification model is programmed in section 5.4.2. Section 5.4.3 presents the program code used to determine a linear equation for the fertilizer response and to test the linear fit.

5.4.1 Contingency table The following SAS program code reads in the fertilizer study data and the first `freq` procedure provides summary counts in the form of a contingency table (see Table 12, page 38). The second `freq` procedure provides the results by rows within the treatments to check for homogeneity of row response. These results were summarized in Table 13 (page 39). Essential SAS statements appear in boldface type.

```
title ‘Fertilizer study’;
proc format; value alive 0=’Alive’ 1=’Dead’;
data fert;  total = 10;
  do treat = 0 to 300 by 100;
    do rw  = 1 to 6;
       row = 6 * (treat/100) + rw;  *To get row numbers from 1 to 24;
       input count @@;
       y = 0; output;
       y = 1; count = total - count; output;
  end; end; format y alive.;
datalines;
 4  5  6  6  4  5
 7  8  6  9  7  5
 6  8  6  9  9 10
 9 10  9 10  8  9
;
proc freq; weight count;
  table y*treat / norow nopercent chisq;
title2 ‘Response Frequencies by Treatments’;
proc freq; by treat notsorted; weight count;
  table y*rw / norow nopercent chisq;
title2 ‘Are rows similar within each treatment?’;
run;
```

5.4.2 One-way classification analysis

The logistic analysis of this data using PROC CATMOD, including a test for linear response to the treatments on the logit scale, is accomplished with the following SAS program code:

```
proc catmod;
  population treat row;
  weight count;  *<== To include the number of trees surviving in each row;
  model y = treat;
  contrast ‘Linear’ treat -3 -2 -1;
title2 ‘Correct Analysis keeping rows as the experimental units’;
run;
```

The output is shown in Figure 35. The initial parts of the output show the populations and response profiles. Note that the likelihood ratio output at the bottom of the MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE is exactly the same as the test results for homogeneity of rows with treatment shown in Table 13 (page 39).

```
                            Fertilizer study
                Correct Analysis keeping rows as the experimental units

                            CATMOD PROCEDURE

        Response: Y                             Response Levels (R)=     2
        Weight Variable: COUNT                  Populations     (S)=    24
        Data Set: FERT                          Total Frequency (N)=   240
        Frequency Missing: 0                    Observations  (Obs)=    45

                          POPULATION PROFILES
                                                 Sample
                          Sample  TREAT  ROW      Size
                          ----------------------------
                              1     0     1         10
                              2     0     2         10
                              3     0     3         10
                              .     .     .          .
                              .     .     .          .
                              .     .     .          .
                             22   300     4         10
                             23   300     5         10
                             24   300     6         10

                          RESPONSE PROFILES

                           Response    Y
                           ---------------
                                1    Alive
                                2    Dead

                            Fertilizer study
                Correct Analysis keeping rows as the experimental units

                         MAXIMUM-LIKELIHOOD ANALYSIS

              Sub       -2 Log    Convergence           Parameter Estimates
Iteration  Iteration  Likelihood   Criterion       1         2          3         4
-------------------------------------------------------------------------------------
     0          0      332.71065      1.0000          0         0          0         0
     1          0      254.33126      0.2356     0.9167   -0.9167    -0.1167    0.2833
     2          0      251.08306      0.0128     1.1139   -1.1139    -0.2671    0.2630
     3          0      250.95052    0.000528     1.1552   -1.1552    -0.3079    0.2311
     4          0      250.94999   2.1132E-6     1.1579   -1.1579    -0.3106    0.2284
     5          0      250.94999   4.183E-11     1.1579   -1.1579    -0.3106    0.2284
```

FIGURE 35 *Logistic regression analysis for the one-way classification.*

```
              MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

            Source                    DF   Chi-Square      prob
            --------------------------------------------------
            INTERCEPT                  1        45.80    0.0000
            TREAT                      3        24.57    0.0000

            LIKELIHOOD RATIO          20        21.67    0.3586

                 ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                                Standard    Chi-
     Effect            parameter  Estimate    Error    Square   prob
     ----------------------------------------------------------------
     INTERCEPT                 1    1.1579    0.1711    45.80  0.0000
     TREAT                     2   -1.1579    0.2502    21.41  0.0000
                               3   -0.3106    0.2626     1.40  0.2369
                               4    0.2284    0.2852     0.64  0.4232

                CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

          Contrast                    DF   Chi-Square      prob
          ------------------------------------------------------
          Linear                       1        21.77    0.0000
```

FIGURE 35 *(continued)*

5.4.3 Simple linear relationship The following program determines a linear equation (on the logit scale) for the treatment responses and examines the adequacy of the linear fit. The `DIRECT` statement is used to designate treatment as a continuous variable. The salient part of the output is shown in Figure 36. The plots of the residuals (Figure 16, page 40) show a good fit. Remember that the mirror image effect is due to the two residuals per row (which must add up to zero within each row). We could plot only the survival residuals if we used a `where = 0;` in the `PLOT` procedure.

```
proc catmod;
  population treat row;  direct treat;
  weight count;  *<== to include the number of trees surviving in each row;
  model y = treat;   *<== Output is in Figure 36;
  response logit / out = pred;
title2 ‘Correct Analysis keeping rows as the experimental units’;
run;
proc print data=pred;
  where _type_ = ‘PROB’;  *<== predicted values in Tables 14;
run;
proc plot data=pred vpercent = 50;
  where _type_ = ‘PROB’;
  plot _resid_ * (treat row) / vref = 0;  *<== plots are in Figure 16;
title2 ‘Residual Plots’;
run;
```

```
                         Fertilizer study
         Correct Analysis keeping rows as the experimental units

                     MAXIMUM-LIKELIHOOD ANALYSIS

              Sub        -2 Log     Convergence    Parameter Estimates
Iteration   Iteration   Likelihood    Criterion         1           2
--------------------------------------------------------------------------
    0           0       332.71065       1.0000            0            0
    1           0       254.53063       0.2350       0.1067     0.005400
    2           0       251.33136       0.0126       0.0270     0.007269
    3           0       251.27465     0.000226       0.0146     0.007562
    4           0       251.27462     1.076E-7       0.0143     0.007568
    5           0       251.27462    2.647E-14       0.0143     0.007568

             MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

           Source                  DF   Chi-Square      Prob
           --------------------------------------------------
           INTERCEPT                1         0.00    0.9495
           TREAT                    1        24.95    0.0000

           LIKELIHOOD RATIO        22        22.00    0.4602

                ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                             Standard    Chi-
    Effect            Parameter  Estimate      Error    Square   Prob
    ----------------------------------------------------------------
    INTERCEPT                 1    0.0143     0.2265      0.00  0.9495
    TREAT                     2   0.00757    0.00152     24.95  0.0000
```

FIGURE 36 *Logistic regression analysis using fertilizer treatment as a continuous variable.*

5.5 Multiple Regression: Herbicide Thinning Trial

In this trial, a herbicide lance that injects glyphosate into trees using a shotgun shell was tested for the number of shells needed to kill aspen trees of different sizes. The size of the trees was measured by their dbh (diameter at breast height). The main hypothesis of this study is that larger trees will require more injections (see section 4.4).

5.5.1 Initial data input and summary The data for this study are listed in Appendix 2.[22] The following program reads in the data and produces a

[22] This data is taken from SX84711Q. The analysis of a similar study with a more complicated design is described in Bergerud (1988).

summary of the data (see Table 15, page 43) by dbh class and number of injections. Essential statements appear in boldface type. Other statements improve the appearance of the output.

```
title ‘Aspen Injection Trial’;
options linesize=90 pagesize=59;
proc format; value dbhclass    0 - 15 = ‘ 0 - 15 cm’
  15.1 - 20 = ‘15 - 20 cm’  20.1 - 25 = ‘20 - 25 cm’
  25.1 - 30 = ‘25 - 30 cm’  30.1 - 35 = ‘30 - 35 cm’
  35.1 - 40 = ‘35 - 40 cm’  40.1 - 45 = ‘40 - 45 cm’  ;
  value dead 0 = ‘0 ..Alive’  1 = ‘1 .. Dead’;
run;
data aspen;
  infile ‘aspen.dat’ ;
  input dbh inj def;
  dead = (def ge 95);** If def ge 95 then dead = 1, otherwise dead = 0;
  tree + 1;  ** assigning each tree a unique identifier;
label dbh = ‘Aspen dbh (cm)’       inj = ‘Number of Injections’
      def = ‘Percent Defoliation’ dead = ‘Dead or Alive’;
format dead dead.;  run;
proc freq data=aspen;
  table dbh*inj / norow nocol nopercent;
format dbh dbhclass.;
run;
```

The `Proc Format` code establishes the 5 cm dbh classes and `Proc Freq` uses these classes to create the table (indicated by the format statement: `format dbh dbhclass.;`). To develop some intuitive feel for the data, Table 16 (page 43) was produced by modifying the standard SAS output to include the number of dead trees in each cell. In addition, the last two size classes are pooled.

5.5.2 Multiple logistic regression model The following program fits the multiple regression model. Essential SAS statements appear in boldface type.

```
proc logistic;
  model dead = dbh inj;
title2 ‘Logistic Regression Analysis’;
run;
```

The first part of the output, which describes the data, is shown on page 100. The model fitting information is presented in Table 17 (page 44).

```
                        Aspen Injection Trial
                     Logistic Regression Analysis

                        The LOGISTIC Procedure

Data Set: WORK.ASPEN
Response Variable: DEAD      Dead or Alive
Response Levels: 2
Number of Observations: 92
Link Function: Logit


                           Response Profile

                      Ordered
                        Value    DEAD     Count

                            1       0        47
                            2       1        45
```

The rest of the output appeared in Figure 18 (page 45). Since `Dead = 0` is the first value listed in the Response Profile above, the program fits the probability of survival.

To confirm the marginally significant p-values, the models are rerun with and without each term. The following program fits these submodels for `dbh` and `inj` to perform deviance tests. The output is presented in Figure 37.

```
proc logistic;
  model dead =     inj;
title2 ‘Logistic Regression Analysis’;
title4 ‘With just Number of Injections’;
run;

proc logistic;
  model dead = dbh    ;
title2 ‘Logistic Regression Analysis’;
title4 ‘With just dbh’;
run;
```

```
                          Aspen Injection Trial
                      Logistic Regression Analysis

                     With just Number of Injections

   Model Fitting Information and Testing Global Null Hypothesis BETA=0

                            Intercept
              Intercept       and
Criterion       Only       Covariates    Chi-Square for Covariates

AIC             129.496      111.233          .
SC              132.017      116.277          .
-2 LOG L        127.496      107.233       20.262 with 1 DF (p=0.0001)
Score              .            .          18.455 with 1 DF (p=0.0001)

                              With just dbh

                          The LOGISTIC Procedure

   Model Fitting Information and Testing Global Null Hypothesis BETA=0

                            Intercept
              Intercept       and
Criterion       Only       Covariates    Chi-Square for Covariates

AIC             129.496       62.678          .
SC              132.017       67.721          .
-2 LOG L        127.496       58.678       68.818 with 1 DF (p=0.0001)
Score              .            .          50.385 with 1 DF (p=0.0001)
```

FIGURE 37 *Results of simple logistic regression fitting procedure.*

5.5.3 Model interpretation The following program calculates the fitted dbh values for 50 and 95% mortality of trees for each number of injections and produces plots of the predicted equations. The program shows both `PROC CATMOD` and `PROC LOGISTIC` approaches, although the output from `PROC LOGISTIC` will be used for subsequent program steps. To do this, we first run the multiple regression model, but now output two data sets, which I have called:

1. `parms` containing the parameter estimates, using the `outest = parms;` option and,
2. `pred` containing the predicted values, using the `out = pred` option.

Note that these option statements appear in different places within the two procedures. The parameter estimates can be used to determine the fitted equations, fitted values, and sizes of trees that would be expected to die if given a specific number of herbicide injections. New SAS statements appear in boldface type.

```
/*  PROC CATMOD Approach:
proc catmod data=aspen;
  direct inj dbh; population tree;
  model dead = inj dbh / noprofile noiter;
  response logit / out = pred outest = parms;
title3 ‘Using Multiple Regression Model’;
*/
/*  PROC LOGISTIC Approach:  */
proc logistic data=aspen outest = parms;
  model dead = inj dbh;
  output out = pred p = pred reschi = reschi resdev = resdev;
title3 ‘Using Multiple Regression Model’;
run;

proc print data=parms; title4 ‘PARMS dataset’;
run;
```

The PROC CATMOD output for the parms data set is:

```
                              PARMS dataset

OBS    _METHOD_    _TYPE_    _NAME_          B1        B2          B3

 1        ML       PARMS               -11.0306    -2.07431     0.98133
 2        ML       COV          B1       6.5255     0.69319    -0.45788
 3        ML       COV          B2       0.6932     0.47875    -0.13877
 4        ML       COV          B3      -0.4579    -0.13877     0.05253
```

Note that the parms data set includes the parameter estimates on the first line—B1 is the intercept, B2 is the slope or coefficient for the number of injections, and B3 is the slope for dbh. The other three lines provide the covariance matrix for these parameter estimates. This matrix can be used to determine the confidence limits around fitted values.

The PROC LOGISTIC output for the parms data set is:

```
                         Aspen Injection Trial
                             Parms Data Set

OBS  _LINK_   _TYPE_    _NAME_      INTERCEP     INJ         DBH     _LNLIKE_

 1    LOGIT    PARMS    ESTIMATE   -11.0306  -2.07431    0.98133   -20.9346
```

The parameter estimates are labelled INTERCEP, INJ, DBH instead of B1, B2, B3, and the log-likelihood for the model is in a variable called _LNLIKE_. The model’s −2LogL can be obtained by multiplying this value by −2. If COVOUT were added to the PROC LOGISTIC statement, the covariance matrix would also be added to the parms data set. Table 18 (page 46), showing the equations and probabilities of mortality for given numbers of injections, was created using the following code (based on the parms data set output by PROC LOGISTIC):

```
data eqns;  set parms;
  if _type_ = ‘PARMS’;      * <== to select only the parameter estimates;
  logit50 = log(0.50/0.50); * <== calculate logit value for mortality of 50%;
  logit95 = log(0.05/0.95); * <== calculate logit value for mortality of 95%;
  do inj = 1 to 8;
   intrcept = b1 + b2 * inj;   * <== calculate intercept for each injection no.;
   dbh50   = (logit50 - intrcept) / b3;  *<==  predicted tree size at 50% mortality;
   dbh95   = (logit95 - intrcept) / b3;  *<==  predicted tree size at 95% mortality;
   output;
  end;
label intrcept = ‘Intercept’ b2 = ‘B2 * DBH’  inj = ‘Injection_number’
      dbh50 = ‘Predicted size_at 50% mortality’
      dbh95 = ‘Predicted size_at 95% mortality’;
run;
proc print split = ‘_’;
  id inj; var intrcept b2 dbh50 dbh95;
title4 ‘Listing of equations for each injection number’; run;
```

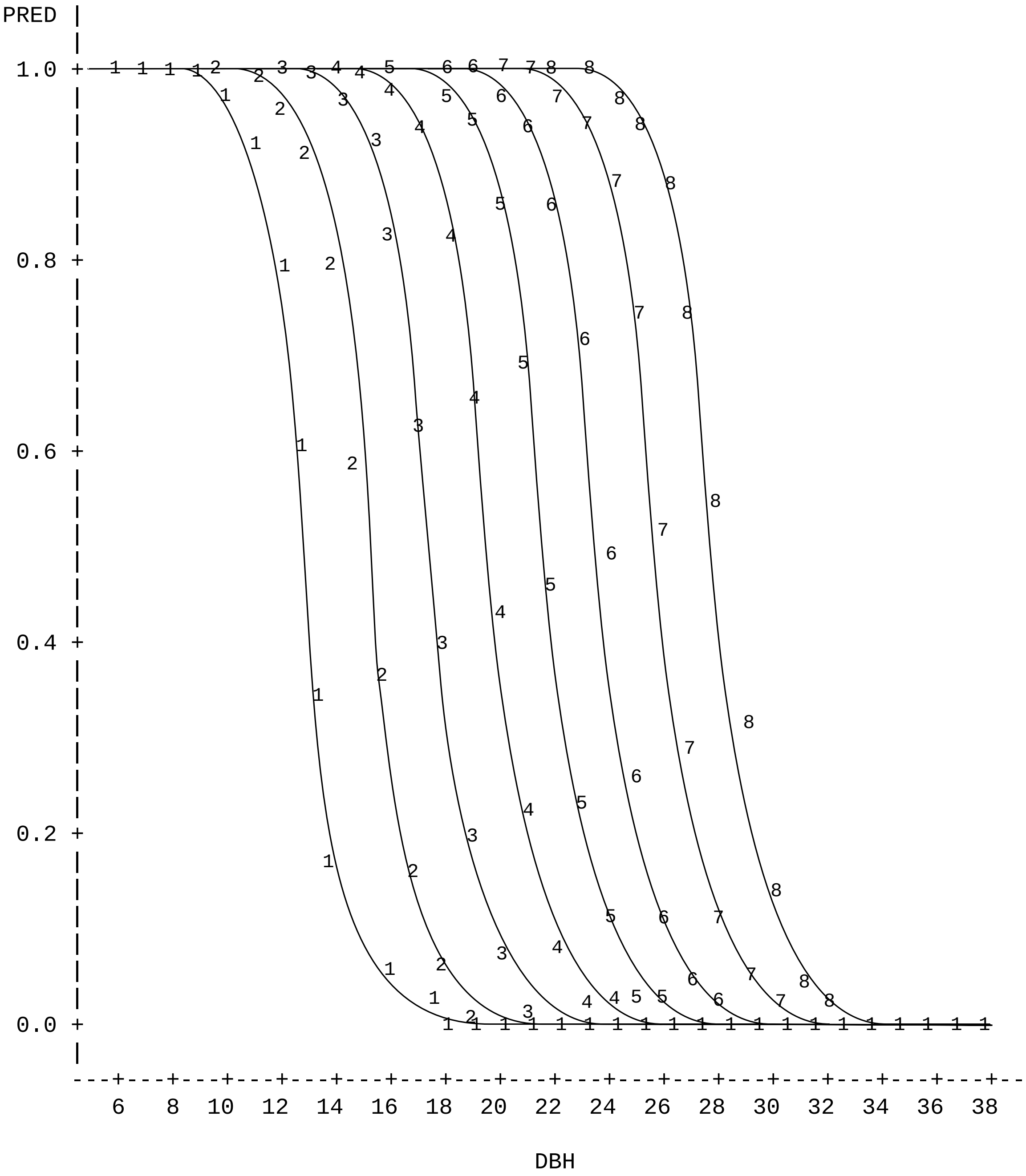
Graphs of the fitted curves will be easier to read if a good range of inj and dbh values is available. These values are calculated using the parms data set above and the following code:

```
data plot;  set parms (rename = (intercep = b1 dbh = b2 inj = b3));
  if _type_ = ‘PARMS’;         * <== to select only the parameter
estimates;
  do inj = 1 to 8;             * <== range of injection numbers;
    do dbh = 6 to 38;          * <== range of tree sizes;
      logit = b1 + b2*inj + b3*dbh;
      pred = 1/(1 + exp(logit));* <== predicted probability of mortality;
      output;  end; end; run;
proc plot data=plot;           *<== Output in Figure 38;
  plot pred*dbh = inj / haxis = 6 to 38 by 2;
title3 ‘Plots of Predicted Probability of Mortality versus Aspen size’;
run;
```

The graph produced (see Figure 38) shows how mortality probability decreases with increasing tree size and that this can be counter-balanced by injecting the tree with more glyphosate capsules.

```
                    Aspen Injection Trial
                 Logistic Regression Analysis
   Plots of Predicted Probability of Mortality versus Aspen Size

           Plot of PRED*DBH.  Symbol is value of INJ.
```



```
NOTE: 156 obs hidden.
```

FIGURE 38 *Printer plot of curves for the multiple regression model showing predicted probability of mortality versus aspen size. (Lines are hand-drawn.)*

The predicted values from the fitting procedure were output into a data set called `pred`. The code and some of that data are shown below:

```
proc print data = pred;
title3 ‘Listing of the PRED dataset’; run;
```

```
                    Aspen Injection Trial
                      Predicted Data Set

OBS   DBH  INJ  DEF    DEAD    TREE  _LEVEL_     PRED    RESCHI    RESDEV

  1  11.0   2   100  1 .. Dead   1  0 ..Alive  0.01232  -0.11166  -0.15743
  2  11.0   2   100  1 .. Dead   2  0 ..Alive  0.01232  -0.11166  -0.15743
  3  12.0   2   100  1 .. Dead   3  0 ..Alive  0.03220  -0.18239  -0.25583
  4  10.5   3   100  1 .. Dead   4  0 ..Alive  0.00096  -0.03097  -0.04379
  5  11.5   3   100  1 .. Dead   5  0 ..Alive  0.00255  -0.05059  -0.07149
  6  12.0   3   100  1 .. Dead   6  0 ..Alive  0.00416  -0.06465  -0.09134
  7  12.0   3   100  1 .. Dead   7  0 ..Alive  0.00416  -0.06465  -0.09134
  8  13.0   3   100  1 .. Dead   8  0 ..Alive  0.01103  -0.10560  -0.14893
  9  13.0   3   100  1 .. Dead   9  0 ..Alive  0.01103  -0.10560  -0.14893
 10  13.0   3   100  1 .. Dead  10  0 ..Alive  0.01103  -0.10560  -0.14893
  .     .   .     .    ..        .    ..          .         .         .
  .     .   .     .    ..        .    ..          .         .         .
  .     .   .     .    ..        .    ..          .         .         .
 82  28.0   7    15  0 ..Alive  82  0 ..Alive  0.87287   0.38163   0.52147
 83  30.0   7     5  0 ..Alive  83  0 ..Alive  0.97995   0.14304   0.20127
 84  30.5   7    25  0 ..Alive  84  0 ..Alive  0.98763   0.11192   0.15779
 85  31.0   7    10  0 ..Alive  85  0 ..Alive  0.99239   0.08757   0.12361
 86  31.0   7    10  0 ..Alive  86  0 ..Alive  0.99239   0.08757   0.12361
 87  33.0   7    20  0 ..Alive  87  0 ..Alive  0.99892   0.03282   0.04641
 88  33.0   7    10  0 ..Alive  88  0 ..Alive  0.99892   0.03282   0.04641
 89  34.5   7     5  0 ..Alive  89  0 ..Alive  0.99975   0.01572   0.02223
 90  24.5   8   100  1 .. Dead  90  0 ..Alive  0.02706  -0.16676  -0.23422
 91  32.5   8    20  0 ..Alive  91  0 ..Alive  0.98619   0.11835   0.16679
 92  42.0   8     0  0 ..Alive  92  0 ..Alive  1.00000   0.00112   0.00158
```

To calculate the dbh of trees predicted by the model to die with specific probabilities (0.01, 0.50, and 0.99) for each number of injections, the following code is used. These values are plotted in Figure 19 (page 47). This graph shows that the constant probability lines are straight. With each additional injection, the size of the tree must increase to have the same probability of surviving.

```
data eqns;  set parms (rename = (intercep = b1 dbh = b2 inj = b3));
  if _type_ = ‘PARMS’;          * <== to select only the parameter
estimates;
  do prob = 0.01, 0.50, 0.99;
    logit = log((1-prob)/prob);
    do inj = 1 to 8;
      intrcept = b1 + b3 * inj;
      dbh = (logit - intrcept) / b2;
      output;
    end; end;
  label intrcept = ‘Intercept’      b2 = ‘Slope for Dbh’
       b3 = ‘Slope for Inj Number’  b1 = ‘Simple Intercept’
       prob = ‘Probability of Survival’
       logit =’Logit of Survival’;
  run;
proc format; value prob 0.01 = ‘&’ 0.50 = ‘*’  0.99 =’+’; run;
proc plot;
  plot dbh * inj = prob/vaxis = 0 to 35 by 5;   *<== Output in Figure 19;
 *plot dbh * inj = logit;
  format prob prob.;
title3 ‘Size of Tree Predicted to Die with the Specified Probability’;
run;
```

5.5.4 Assessing the adequacy of the multiple regression model As a final check of the multiple regression model, the residuals should be plotted against the independent variables, `dbh` and `inj`. The following code will do this.

```
proc plot data = pred;*<== Output in Figures 20 and 21;
  plot (reschi resdev)*(dbh inj) / vref = 0;
title3 ‘Diagnostic Plots’;
run;
```

Recall that `PROC LOGISTIC` produces two types of residuals and these are plotted against `dbh` and `inj` in Figures 20 and 21 (pages 48 and 49). The plots show no obvious patterns and no particular observation appears to stick out, suggesting that the model is adequate.

5.5.5 Factorial models: looking for non-linearity in response More complicated models are necessary to check for non-linearity in survival probability responses. The following program refits the logistic regression model with number of injections as a categorical variable and dbh as a continuous variable (using dbh values directly and not grouping them into classes). First we must define the proper experimental unit in a `Population` statement so that we can compare the various models we fit to each other. In this case, each tree is independently assigned a treatment (number of injections) and is the experimental unit. A further assumption of the analysis is that each tree responds independently to the treatments. Therefore, trees are the correct experimental unit and the variable `tree` identifies each tree uniquely. Essential SAS statements appear in boldface type. The output appears in Figure 39.

```
proc catmod data = aspen;
  population tree;
  direct dbh;      * <== indicates that dbh is a continuous variable;
  format dbh 5.1;  * <== use dbh values and not class (formatted) values;
                   * (only necessary if dbh had been assigned
                      the dbhclass format in the data step);
  model dead = inj dbh / noprofile noiter;
  contrast ‘Injections: linear’ inj 6 5 4 3 2 1;  * 7 levels;
title3 ‘Treating inj as a class variable and dbh as a covariate’;
run;
```

```
   MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

 Source                   DF   Chi-Square      Prob
 --------------------------------------------------
 INTERCEPT                 1        14.71    0.0001
 INJ                       5*         .        .
 DBH                       1        15.92    0.0001

 LIKELIHOOD RATIO         85        40.56    1.0000

 NOTE: Effects marked with ‘*’ contain one or more
       redundant or restricted parameters.

     CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

Contrast                    DF   Chi-Square      Prob
----------------------------------------------------
Injections: linear           1         7.52    0.0061
```

FIGURE 39 *Fit of the model with number of injections as a categorical variable and dbh as a covariate (continuous variable).*

In Figure 39, note that the degrees of freedom for inj are five and not six as expected (since inj has seven levels, it should have six degrees of freedom).

We can expand the non-linearity test to include dbh by using the dbh class intervals. The following program fits a two-way factorial design with dbh class and injection number as the two factors. Essential SAS statements appear in boldface type.

```
proc format; value dbhclass    0 - 15 = ‘0 - 15 cm’
  15.1 - 20 = ‘15 - 20 cm’  20.1 - 25 = ‘20 - 25 cm’
  25.1 - 30 = ‘25 - 30 cm’  30.1 - 45 = ‘ > 30 cm’;
run;
proc catmod;
  population tree;
  model dead = inj dbh inj*dbh / noiter; *<== Output is in Figure 40;
  contrast ‘Injections: linear’ inj 6 5 4 3 2 1;  * 7 levels;
  contrast ‘Dbh: linear’        dbh     4 3 2 1;  * 5 levels;
format dbh dbhclass.; *  <== to use dbh class values defined by the
dbhclass format;
title2 ‘Logistic Regression Analysis’;
run;
```

This program redefines the dbh classes with the proc format so that all the trees greater than 30 cm are pooled together. The population statement identifies the experimental units with the variable tree. The model statement indicates that the response variable is dead, which is fitted with a model including inj, dbh, and inj*dbh as categorical variables (since there is no direct statement). The option in the model statement indicates that information about the number of iterations required to fit the model are not to be output (noiter). The contrast statements are similar to those of PROC GLM and are described in section 5.1.1.1. The format statement specifies that the categorical values (defined by the format dbh class) are used instead of the actual dbh values.

The output is shown in Figure 40. It describes the various “populations” under consideration and the values of the response variable. While the list of population and response profiles can be deleted with the noprofile option in the model statement, it is important to look at this output at least once to ensure that the experimental units are correctly identified by the population statement and to check which level of the response variable is first and being fitted by the procedure. All models run on one data set should have the same defined experimental units so that these models can be correctly compared to each other. This is checked by seeing if the different models have the same total degrees of freedom (calculated by adding up the degrees of freedom column in the MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE produced by each Model statement—92 in this case). This total should be equal to the number of experimental units in the study. The proper definition of the experimental units is an important concern because PROC CATMOD can automatically pool observations in ways that may be inappropriate. This would occur, for instance, when fitting a model with dbh, but not injection number. Such a model could not be compared to the full model without the use of the population statement.

In this example, the main effects and interaction—`dbh`, `inj` and `dbh*inj`—do not have the degrees of freedom expected of them because of the missing dbh and injection number combinations, nor has SAS provided the tests as we would have liked. The output contains a warning message (`NOTE: Effects marked with ‘*’ contain one or more redundant or restricted parameters`.) to let us know that there is a problem. An overall test of the lack of fit of the multiple regression program is calculated by looking at the difference in −2LogL of the multiple regression and factorial models: 41.87 − 40.62 = 1.25 with 89 − 78 = 11 degrees of freedom. This is clearly not significant and implies that the multiple regression model is quite adequate.

```
                           CATMOD PROCEDURE

Response: DEAD                       Response Levels (R)=     2
Weight Variable: None                Populations     (S)=    92
Data Set: ASPEN                      Total Frequency (N)=    92
Frequency Missing: 0                 Observations  (Obs)=    92

                          POPULATION PROFILES
                                    Sample
                        Sample  TREE     Size
                        ----------------------
                           1     1         1
                           2     2         1
                           3     3         1
                           .     .         .
                           .     .         .
                           .     .         .
                          90    90         1
                          91    91         1
                          92    92         1

                           RESPONSE PROFILES

                           Response     DEAD
                          --------------------
                             1     0 ..Alive
                             2     1 .. Dead

                MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                  DF   Chi-Square      Prob
       --------------------------------------------------
       INTERCEPT                1         0.68    0.4101
       INJ                      3*          .       .
       DBH                      2*          .       .
       INJ*DBH                  8*          .       .

       LIKELIHOOD RATIO        78        40.62    0.9999

       NOTE: Effects marked with ‘*’ contain one or more
             redundant or restricted parameters.

                  CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

     Contrast                    DF   Chi-Square      Prob
     ------------------------------------------------------
     Injections: linear           1         1.15    0.2828
     Dbh: linear                  1         2.41    0.1203
```

FIGURE 40 *Fit of the two-way factorial model.*

5.6 One-way Classification With Covariable: Root Collar Weevil Study

The data collected are as in Appendix 2. The purpose of this study was to test if screefing the duff from around seedlings would protect them from the root collar weevil which lives in the duff. The study's main hypothesis was that the probability of weevil attack would increase with weevil numbers, but decrease with the screefing treatment.

5.6.1 Initial data input and plotting The following program reads in the data and plots the observed proportions for an exploratory look. Essential SAS statements appear in boldface type. The output appeared in Figure 25 of Section 4.5.4 (page 57).

```
title ‘Root Collar Weevil Example’;
** To help remind us of the meaning of the levels of y and trmt  ;
proc format; value y    2 =’2: Not Attacked’  1 = ‘1: Attacked’;
             value trmt 0 = ‘Screefed’        1 = ‘Control (1)’;     run;
data weevil;
  infile ‘weevil.dat’;
  input trmt plot number weevil;
** Two observations for each plot are required.  One observation for the
   number of attacked seedlings and another for those not attacked.;
  count = number;      y = 1; prop = count/16;
      logit = log((prop+0.01)/(1-prop+0.01)); output;
  count = 16 - number; y = 2; prop = count/16;
      logit = log((prop+0.01)/(1-prop+0.01)); output;
**  Labels to help remind us of what the variables are.;
label trmt = ‘Treatment’  weevil = ‘Numbers of Weevils’
      prop = ‘Observed Proportion’  logit = ‘Empirical Logit’
     number= ‘Seedlings Attacked’ plot =’Plot Number’;
  format y y. trmt trmt.;
run;
proc sort; by trmt weevil;  run;
proc print;
  id trmt plot;  var weevil number y count prop;
run;
proc plot data=weevil;
  where y = 1;  * <==  plotting the attack counts only;
  plot count * weevil = trmt;  *<== Output in Figure 25;
  *plot logit * weevil = trmt; *<== Output would look much the same for this data;
title2 ‘Plot of the Observed Data’;
run;
```

5.6.2 Standard analysis of covariance This program fits several logistic regression models to the data. Essential SAS statements appear in boldface type. Specific details are discussed with the output after the program.

```
proc catmod;
  population trmt plot;
  direct weevil;  weight count;
  model y = trmt weevil trmt*weevil / noiter;
  title2 ‘Model 1: Full model with two separate lines’;
run;
  model y = trmt weevil             / noiter noprofile;
  title2 ‘Model 2: With two parallel lines’;
run;
**  The next models are included to provide more exact tests of the trmt
    and weevil effects and to check the Wald statistics   ;
  model y = trmt                    / noiter noprofile;
  title2 ‘Model 3: With treatment only - Two groups’;
run;
  model y =      weevil             / noiter noprofile;
  title2 ‘Model 4: With weevil only - One line’;
run;
```

The first part of the output shows the population/experimental unit profiles and describes the response values. This is used to check that the data are properly defined.

```
                          Root Collar Weevil
             Model 1: Full Model with two separate lines

                          CATMOD PROCEDURE

Response: Y                            Response Levels (R)=     2
Weight Variable: COUNT                 Populations     (S)=    28
Data Set: WEEVIL                       Total Frequency (N)=   448
Frequency Missing: 0                   Observations  (Obs)=    45
```

Note that since 11 observations have zero counts, CATMOD recognizes 45 observations instead of 56. Therefore, observations with zero counts need not be kept in the data set.

```
                POPULATION PROFILES
                                      Sample
        Sample     TRMT      PLOT     Size
        -----------------------------------
            1   Screefed       1         16
            2   Screefed       2         16
            3   Screefed       3         16
            4   Screefed       4         16
            .      .           .          .
            .      .           .          .
            .      .           .          .
           25   Control (1)   11         16
           26   Control (1)   12         16
           27   Control (1)   13         16
           28   Control (1)   14         16

                 RESPONSE PROFILES

             Response         Y
             -------------------------
                  1    1: Attacked
                  2    2: Not Attacked
```

The MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE for model 1 is shown in Figure 41, that of model 2 is shown in Figure 42 while those of models 3 and 4 are shown in Figure 43.

```
   Model 1: Full model with two separate lines

  MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source                   DF   Chi-Square      Prob
--------------------------------------------------
INTERCEPT                 1        42.05    0.0000
TRMT                      1         0.70    0.4015
WEEVIL                    1        24.83    0.0000
WEEVIL*TRMT               1         1.69    0.1932

LIKELIHOOD RATIO         24         8.38    0.9987
```

FIGURE 41 *Output from model 1: separate lines for each treatment level.*

```
                  Model 2: With two parallel lines

         MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                  DF   Chi-Square      Prob
       --------------------------------------------------
       INTERCEPT                1        68.66    0.0000
       TRMT                     1        60.42    0.0000
       WEEVIL                   1        43.20    0.0000

       LIKELIHOOD RATIO        25         9.85    0.9970

            ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                                       Standard    Chi-
Effect            Parameter  Estimate    Error    Square   Prob
----------------------------------------------------------------
INTERCEPT                 1   -5.8269    0.7032    68.66  0.0000
TRMT                      2   -1.7052    0.2194    60.42  0.0000
WEEVIL                    3    0.3086    0.0469    43.20  0.0000
```

FIGURE 42 *Output from model 2: parallel lines, one for each treatment level.*

```
          Model 3: With treatment only - Two groups

         MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                  DF   Chi-Square      Prob
       --------------------------------------------------
       INTERCEPT                1       102.12    0.0000
       TRMT                     1        48.45    0.0000

       LIKELIHOOD RATIO        26       102.00    0.0000

              Model 4: With weevil only - One line

         MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                  DF   Chi-Square      Prob
       --------------------------------------------------
       INTERCEPT                1        80.68    0.0000
       WEEVIL                   1        46.41    0.0000

       LIKELIHOOD RATIO        26       112.90    0.0000
```

FIGURE 43 *Partial output for models 3 and 4.*

5.6.3 Comparing the models The following program fits alternative models (models 5 to 8) to the data. Two new variables, `wvltrt` and `wvlcon` are created in the data step and are shown in Table 32. Essential SAS code appears in boldface type.

```
data weevil;
  set weevil;
  wvltrt = (trmt=0) * weevil;
  wvlcon = (trmt=1) * weevil;
run;
proc sort data=weevil; by trmt count y;
proc print data=weevil;  where y = 1;    *<==  Output in Table 32;
  id plot; by trmt; var number weevil wvltrt wvlcon;
title2 'The new variables: wvltrt and wvlcon';
run;
proc catmod;
  population trmt plot;
  weight count;
  direct wvltrt wvlcon;
  model y = trmt wvltrt wvlcon   / noiter noprofile;
  title2 'Model 5: With two weevil variables and an intercept';
run;
  model y = wvltrt wvlcon   / noiter noprofile;
  title2 'Model 6: With two weevil variables';
  title3 'i.e. Two radiating lines from the origin';
run;
  model y = wvltrt         / noiter noprofile;
  title2 'Model 7: With weevil for treatment only';
  title3 'i.e. flat line for control';
run;
  model y =        wvlcon  / noiter noprofile;
  title2 'Model 8: With weevil for control only';
  title3 'i.e. flat line for treatment';
run;
```

The output for model 5 is shown in Figure 44, for model 6 is shown in Figure 45, for model 7 is shown in Figure 46 and for model 8 is shown in Figure 47.

TABLE 32 *The value of the new continuous variables, weevil, wvltrt, and wvlcon*

```
--------------------------------------------------------------------------

                           Root Collar Weevil Example
                      The new variables: wvltrt and wvlcon

--------------------------- Treatment=Screefed ---------------------------

               PLOT    NUMBER    WEEVIL    WVLTRT    WVLCON

                 1        0        10        10         0
                 2        1        13        13         0
                 3        0        11        11         0
                 4        2        16        16         0
                 5        1        16        16         0
                 6        1         9         9         0
                 7        0         6         6         0
                 8        1        14        14         0
                 9        1        17        17         0
                10        1        19        19         0
                11        0         0         0         0
                12        0         3         3         0
                13        0         2         2         0
                14        0         1         1         0

------------------------- Treatment=Control (1) --------------------------

               PLOT    NUMBER    WEEVIL    WVLTRT    WVLCON

                 1        6        12         0        12
                 2        6        12         0        12
                 3       12        17         0        17
                 4        7        12         0        12
                 5        9        14         0        14
                 6        7        12         0        12
                 7        7        12         0        12
                 8        3         8         0         8
                 9        9        14         0        14
                10       13        19         0        19
                11        0         2         0         2
                12        0         1         0         1
                13        0         0         0         0
                14        0         1         0         1
--------------------------------------------------------------------------
```

```
       Model 5: With two weevil variables and an intercept
           i.e. Full model with two separate lines

         MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                  DF   Chi-Square      Prob
       --------------------------------------------------
       INTERCEPT                1        42.05    0.0000
       TRMT                     1         0.70    0.4015
       WVLTRT                   1         4.64    0.0313
       WVLCON                   1        36.70    0.0000

       LIKELIHOOD RATIO        24         8.38    0.9987

            ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES
                                       Standard    Chi-
Effect           Parameter  Estimate    Error    Square   Prob
----------------------------------------------------------------
INTERCEPT                1   -5.1325    0.7915    42.05  0.0000
TRMT                     2   -0.6640    0.7915     0.70  0.4015
WVLTRT                   3    0.1969    0.0914     4.64  0.0313
WVLCON                   4    0.3360    0.0555    36.70  0.0000
```

FIGURE 44 *Output from model 5: full model with two weevil variables and an intercept.*

```
             Model 6: With two weevil variables
                 Two Radiating Lines Model

         MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

       Source                  DF   Chi-Square      Prob
       --------------------------------------------------
       INTERCEPT                1        57.84    0.0000
       WVLTRT                   1         8.09    0.0044
       WVLCON                   1        54.29    0.0000

       LIKELIHOOD RATIO        25         9.18    0.9983

            ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES
                                       Standard    Chi-
Effect           Parameter  Estimate    Error    Square   Prob
----------------------------------------------------------------
INTERCEPT                1   -4.8370    0.6360    57.84  0.0000
WVLTRT                   2    0.1358    0.0477     8.09  0.0044
WVLCON                   3    0.3635    0.0493    54.29  0.0000
```

FIGURE 45 *Output from model 6: two radiating lines model with two weevil variables only.*

```
           Model 7: With weevil for treatment only
                 i.e. flat line for control

        MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

      Source                  DF   Chi-Square      Prob
      --------------------------------------------------
      INTERCEPT                1        49.13    0.0000
      WVLTRT                   1        23.95    0.0000

      LIKELIHOOD RATIO        26       144.45    0.0000
```

FIGURE 46 *Output from model 7: one weevil variable for the treatment and a flat line for the control.*

```
                     Root Collar Weevil
             Model 8: With weevil for control only
                 i.e. flat line for treatment

        MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

      Source                  DF   Chi-Square      Prob
      --------------------------------------------------
      INTERCEPT                1       114.89    0.0000
      WVLCON                   1        94.49    0.0000

      LIKELIHOOD RATIO        26        17.70    0.8865

           ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES
                                          Standard    Chi-
Effect            Parameter  Estimate      Error    Square   Prob
----------------------------------------------------------------
INTERCEPT                 1   -3.6086     0.3367    114.89  0.0000
WVLCON                    2    0.2721     0.0280     94.49  0.0000
```

FIGURE 47 *Output from model 8: one weevil variable for the control and a flat line for the treatment.*

To further compare models 2, 6, and 8, we should look at various plots to confirm that the models do adequately fit the data. The following program reruns these models to obtain residuals and predicted values (see Table 33) and then combines them with the original data set to create the plots in Figures 26–30 in section 4.5.7 (pages 61–65). New SAS code appears in boldface type.

The equations were used to calculate the predicted values plotted in Figures 26–28 (pages 61–63). To show clearly the predicted lines amidst the data in this case required calculating an abundance of predicted values. Often this can be adequately accomplished with the observed independent values, but in this case a separate SAS program was required to generate predicted values for a range of weevil numbers. The SAS program used to do this is not shown.

```
proc catmod;
  population trmt plot;
  direct weevil;  weight count;
  model y = trmt weevil             / noiter;
  response logit / out = pred1;
  title2 ‘Model 2: With parallel lines’;
run;
  direct wvlcon wvltrt;
  model y = wvlcon wvltrt           / noiter noprofile;
  response logit / out = pred2;
  title2 ‘Model 6: Two radiating lines’;
run;
  model y =              wvlcon     / noiter noprofile;
  response logit / out = pred3;
  title2 ‘Model 8: sloping line for control and a flat line for the treatment’;
run;
**  Sorting the data sets prior to merging;
proc sort data=weevil; by trmt plot y; run;
proc sort data=pred1;  by trmt plot y; run;
proc sort data=pred2;  by trmt plot y; run;
proc sort data=pred3;  by trmt plot y; run;
data pred1; set pred1; pred1 = _pred_; resid1 = _resid_;  run;
data pred2; set pred2; pred2 = _pred_; resid2 = _resid_;  run;
data pred3; set pred3; pred3 = _pred_; resid3 = _resid_;  run;
data all;
  merge weevil pred1 pred2 pred3;
  by trmt plot y;
  if y eq 1; * <== keeping the attack probabilities only;
label prop = ‘Observed Proportion’       pred1 = ‘Parallel Lines Pred’
      pred2 = ‘Two Rad. Lines Pred’      pred3 = ‘Control Line Pred’
     resid1 = ‘Parallel Lines Residual’ resid2 = ‘Two Rad. Lines Residual’
     resid3 = ‘Control Line Residual’  ;
run;
proc sort data=all;  by trmt weevil; run;
proc print data = all label;
  by trmt; id plot; format y 2.;         *<== Output in Table 33;
  var weevil number prop pred1 pred2 pred3;
title2 ‘Listing of Predicted Proportions’;
run;
proc plot data = all vpercent = 33;
  plot (pred1  pred2  pred3 )*prop / vaxis = 0 to 1 by 0.2;
  plot (resid1 resid2 resid3)*(trmt)
       (resid1 resid2 resid3)*(plot)
       (resid1 resid2 resid3)*weevil=trmt / vref = 0;
run;
proc means data=all n mean uss;
  var resid1 resid2 resid3;
title2 ‘Listing of Residual Sums of Squares for Three Models’;
run;
```

TABLE 33 *Predicted values for models 2, 6, and 8*

```
-------------------------- Treatment=Screefed --------------------------

          Numbers                                        Two     Control
   Plot      of     Seedlings   Observed    Parallel     Lines     Line
  Number  Weevils   Attacked  Proportion  Lines Pred    Pred      Pred

    11        0        0        0.0000      0.00054   0.007868  0.026375
    14        1        0        0.0000      0.00073   0.009002  0.026375
    13        2        0        0.0000      0.00099   0.010298  0.026375
    12        3        0        0.0000      0.00135   0.011778  0.026375
     7        6        0        0.0000      0.00340   0.017595  0.026375
     6        9        1        0.0625      0.00854   0.026210  0.026375
     1       10        0        0.0000      0.01159   0.029907  0.026375
     3       11        0        0.0000      0.01571   0.034108  0.026375
     2       13        1        0.0625      0.02873   0.044278  0.026375
     8       14        1        0.0625      0.03871   0.050393  0.026375
     4       16        2        0.1250      0.06947   0.065092  0.026375
     5       16        1        0.0625      0.06947   0.065092  0.026375
     9       17        1        0.0625      0.09226   0.073859  0.026375
    10       19        1        0.0625      0.15853   0.094720  0.026375

------------------------- Treatment=Control (1) --------------------------

         Numbers                                           Two     Control
 Plot       of      Seedlings    Observed     Parallel     Lines       Line
Number   Weevils     Attacked   Proportion   Lines Pred     Pred       Pred

  13         0          0        0.0000       0.01596    0.00787   0.02637
  12         1          0        0.0000       0.02160    0.01128   0.03434
  14         1          0        0.0000       0.02160    0.01128   0.03434
  11         2          0        0.0000       0.02918    0.01614   0.04460
   8         8          3        0.1875       0.16069    0.12688   0.19281
   1        12          6        0.3750       0.39680    0.38349   0.41497
   2        12          6        0.3750       0.39680    0.38349   0.41497
   4        12          7        0.4375       0.39680    0.38349   0.41497
   6        12          7        0.4375       0.39680    0.38349   0.41497
   7        12          7        0.4375       0.39680    0.38349   0.41497
   5        14          9        0.5625       0.54943    0.56274   0.55001
   9        14          9        0.5625       0.54943    0.56274   0.55001
   3        17         12        0.7500       0.75474    0.79296   0.73439
  10        19         13        0.8125       0.85084    0.88794   0.82653
--------------------------------------------------------------------------
```

6 USING INDICATOR (DUMMY) VARIABLES WITH LOGISTIC REGRESSION PROGRAMS

This technically oriented chapter discusses the regression approach to modelling a classification factor and will help to explain what is happening "behind the scenes". This information allows you to include categorical variables with a logistic regression package that does not set them up automatically (as does, for instance, `PROC CATMOD`). The one-way classification studies and the `ANCOVA` example will be used to illustrate the differences in calculation methods.

6.1 One-way Classification Study: Root Rot and Stand Age

The regression approach requires the creation of indicator variables (also known as dummy variables). At least as many indicator variables are needed as degrees of freedom for that categorical variable. For the one-way classification with three levels of stand age, two indicator variables are required. Although many ways exist to define the indicator variables, the simplest is to create two new variables such that the first (`AGE1`) has a value of one for the first age class and zero for the other age classes. The second variable (`AGE2`) has a value of one for the second age class and zero otherwise. These two new variables are shown in Table 34. Note that a third variable is unnecessary because the third age class is identified by the zero values in both `AGE1` and `AGE2` (`PROC GLM` uses this method, but also includes an unnecessary `AGE3` variable). Another common way to define the indicator variables is to set the values as for `AGE1` and `AGE2` except that the last classification level is assigned a value of −1 instead of zero (`PROC CATMOD` uses this method). These values are listed as variables `AGE1CAT` and `AGE2CAT` in Table 34. Other approaches for creating indicator variables include choosing values so that one fits a linear trend within the age class, while the second fits a quadratic response. Although these definitions of indicator or dummy variables are independent of the statistical package used to run the analysis, it may be important to determine what approach a particular package is using. This should be described in the package's documentation.

TABLE 34 *Indicator variables created for the root rot study*

Stand Number	Age	AGE1	AGE2	AGE1CAT	AGE2CAT	Dead	Total Count	Observed Proportion
1	1..Young	1	0	1	0	13	41	0.31707
2	1..Young	1	0	1	0	8	35	0.22857
3	2..Middle	0	1	0	1	10	28	0.35714
4	2..Middle	0	1	0	1	15	37	0.40541
5	2..Middle	0	1	0	1	6	16	0.37500
6	3..Old	0	0	-1	-1	7	16	0.43750
7	3..Old	0	0	-1	-1	7	18	0.38889
8	3..Old	0	0	-1	-1	19	41	0.46341
9	3..Old	0	0	-1	-1	7	15	0.46667

The rest of this section explains how to use SAS to conduct an analysis of the root rot and stand age example using `PROC LOGISTIC`, instead of `PROC CATMOD`. First, indicator variables are added to the data set. They are defined using Boolean algebra in variable assignment statements. For instance, SAS assigns the term: (`stand ge 3`) a value of one if true (i.e., the value of stand is greater than or equal to 3), or a value of zero if false (i.e, the value of stand is less than 3). In the program below, for example `AGE1` will have a value of one for stands 1 and 2, and a value of zero for the other stands. The analysis uses the second set of indicator variables. In the following SAS program essential SAS code appears in boldface type.

```
title ‘Simple One-Way Classification Example’;
proc format; value age 1 = ‘1..Young’ 2 = ‘2..Middle’ 3 = ‘3..Old’;
             value alive 0 = ‘0 .. Dead’ 1 = ‘1 ..Alive’;   run;
data one;
  input dead alive @@;
  stand + 1;
*  Creating the indicator variables;
  age  = 1 + (stand ge 3) + (stand ge 6);
  age1 = 1 - (stand ge 3);
  age2 = (stand ge 3 and stand lt 6);
  age1cat = (stand lt 3) - (stand ge 6);
  age2cat = (stand ge 3 and stand lt 6) - (stand ge 6);
  total = dead + alive;
  prop  = dead/(total);
  output;
label age = ‘Age’ stand = ‘Stand Number’
      alive = ‘Alive’  dead = ‘Dead’
      total = ‘Total Count’
      prop = ‘Observed Proportion’ ;
format age age. dead alive. ;
cards;
13 28  8 27
10 18 15 22  6 10
 7  9  7 11 19 22  7  8
;
proc print data = one label;
  id stand age ;  var age1 age2 age1cat age2cat dead total prop;
title2 ‘Listing of data’;
run;
proc logistic data=one;
  model dead/total = age1cat age2cat;
  output out=pred p=pred resdev=resdev reschi=reschi;
title2 ‘Logistic Regression Analysis with indicator variables’;
title3 ‘Using Catmod indicator variables’;
run;
** Output for the following procedures is not shown;
proc print;
```

```
title3 ‘Listing of output data set’;
id stand age;  var age1cat age2cat dead total prop pred reschi resdev;
run;
proc plot vpercent=50;
  plot (resdev reschi)*(age stand) / vref = 0;
title3 ‘Plots’;
run;
```

The output from the first print procedure appears in Table 34. The output from the logistic procedure appears below. This first part identifies the data set and response profile.

```
                          Simple One-Way Classification Example
                    Logisitic Regression Analysis with indicator variables
                              Using Catmod indicator variables

                                 The LOGISTIC Procedure

       Data Set: WORK.ONE
       Response Variable (Events): DEAD      Dead
       Response Variable (Trials): TOTAL     Total Count
       Number of Observations: 9
       Link Function: Logit

                                     Response Profile

                               Ordered  Binary
                                 Value  Outcome      Count

                                     1  EVENT           92
                                     2  NO EVENT       155
```

The following output shows the test results for the overall fit of the model to the data. Note that the −2 LOG L for the model with both intercept and covariates (namely 321.039) is the same as that from the three group analysis in section 4.2.2 (see Table 11, page 33). The test for the covariates (-2 LOG L = 5.128 with 2 DF) is the same for the test of group differences in section 4.2.2.

```
                  Testing Global Null Hypothesis: BETA=0
                             Intercept
               Intercept        and
 Criterion       Only       Covariates    Chi-Square for Covariates

 AIC           328.167       327.039         .
 SC            331.676       337.567         .
 -2 LOG L      326.167       321.039       5.128 with 2 DF (p=0.0770)
 Score            .             .          5.037 with 2 DF (p=0.0806)
```

The Wald statistics are presented in the following output. Note that the parameter estimates for the intercept and both indicator variables are identical to the CATMOD output in Figure 32 (page 87). This occurs because both procedures are modelling the same response level. The odds ratios are the estimated odds of not surviving for each age compared to the average of all three ages.

```
                 Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio

INTERCPT 1    -0.5547   0.1346    16.9808     0.0001            .      .
AGE1CAT  1    -0.4081   0.2001     4.1593     0.0414    -0.184405    0.665
AGE2CAT  1     0.0766   0.1885     0.1652     0.6844     0.035189    1.080
```

Previously, we pooled the last two levels of age into two groups: young and older. To do that here, we rerun the analysis without AGE2CAT. This will confirm the Wald test that showed middle- and old-aged stands to have similar responses.

```
proc logistic data=one;
  model dead/total = age1cat        ;
  output out=pred p=pred resdev=resdev reschi=reschi;
title2 ‘Logistic Regression Analysis with indicator variables’;
title3 ‘Using Catmod indicator variables’;
run;
proc print;
title3 ‘Listing of output data set for two group model’;
id stand age;  var age1cat age2cat dead total prop pred reschi resdev;
run;
proc plot vpercent=50;
  plot (resdev reschi)*(age stand) / vref = 0;
title3 ‘Plots for the two group model’;
run;
```

The output (not shown) provides the same results as that of the previous analysis (see section 4.2.4 and section 5.3.4).

6.2 One-way Classification Study: Fertilizer Trial

The data analysis for this fertilizer trial could also be performed using PROC LOGISTIC instead of PROC CATMOD. In the process, we examine indicator variables more closely and see how the logistic regression program is used to analyze designed studies. Two sets of indicator variables are explored. The first set (dum1, dum2, dum3) matches the indicator variables that PROC CATMOD uses and therefore the output from both procedures will provide the same results. The second set uses polynomials and the output for the linear indicator variable matches that of the linear contrast output by PROC CATMOD because this is a balanced design.

```
data fert;
  set fert;
  if y = 0;  ** keeping only the surviving trees;
*  Creating the PROC CATMOD indicator variables;
  dum1 = (treat=1) - (treat=4);
  dum2 = (treat=2) - (treat=4);
  dum3 = (treat=3) - (treat=4);
*  Creating the polynomial indicator variables;
  linear = -3*(treat=1) -1*(treat=2) +1*(treat=3) +3*(treat=4);
  quad   =  1*(treat=1) -1*(treat=2) -1*(treat=3) +1*(treat=4);
  cubic  = -1*(treat=1) +3*(treat=2) -3*(treat=3) +1*(treat=4);
run;
proc logistic;
  model count/total = dum1 dum2 dum3;
title2 ‘Logistic Regression with indicator variables’;
run;
proc logistic;
  model count/total = linear quad cubic;
title2 ‘Logistic Regression with contrasts for indicator variables’;
run;
proc logistic;
  model count/total = linear;
title2 ‘Logistic Regression with linear contrast only’;
run;
```

The output from the first `PROC LOGISTIC` is:

```
                             Fertilizer study
                 Logistic Regression with indicator variables

                          The LOGISTIC Procedure

Data Set: WORK.FERT
Response Variable (Events): COUNT
Response Variable (Trials): TOTAL
Number of Observations: 24
Link Function: Logit

                             Response Profile

                       Ordered  Binary
                         Value  Outcome      Count

                             1  EVENT          175
                             2  NO EVENT        65

                          The LOGISTIC Procedure

                 Testing Global Null Hypothesis: BETA=0

                            Intercept
              Intercept        and
Criterion       Only       Covariates    Chi-Square for Covariates

AIC             282.361       258.950         .
SC              285.842       272.873         .
-2 LOG L        280.361       250.950       29.411 with 3 DF (p=0.0001)
Score              .             .          28.420 with 3 DF (p=0.0001)

                 Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio

INTERCPT 1     1.1579   0.1711    45.7976     0.0001             .     .
DUM1     1    -1.1579   0.2502    21.4140     0.0001     -0.452338    0.314
DUM2     1    -0.3106   0.2626     1.3988     0.2369     -0.121330    0.733
DUM3     1     0.2284   0.2852     0.6413     0.4232      0.089236    1.257
```

Notice that the parameter estimates, standard errors, and Wald statistics for the three indicator variables (dum1, dum2, dum3) are exactly the same as those calculated by PROC CATMOD (see parameters 2, 3, and 4 in Figure 35, page 96). The chi-square for the covariates is similar to the corresponding Wald statistic from the PROC CATMOD output (χ^2 = 24.57, *df* = 3), but not exactly the same. The test from the PROC LOGISTIC output is generally considered more reliable than the Wald statistic in PROC CATMOD.

The following output lists results from the second PROC LOGISTIC model fit. Notice that the chi-square for the covariates test is exactly the same as in the previous output for dum1, dum2, and dum3. The parameter estimates are different because this model was reparameterized, and the parameters have different meanings. Notice that the test for the linear indicator variable is exactly the same as for the linear contrast used with PROC CATMOD (χ^2 = 21.77, *df* = 1 from Figure 35, page 96). This is generally true with balanced study designs.

```
                             Fertilizer study
           Logistic Regression with contrasts for indicator variables

                          The LOGISTIC Procedure

                   Testing Global Null Hypothesis: BETA=0

                                Intercept
                  Intercept        and
Criterion           Only       Covariates    Chi-Square for Covariates

AIC               282.361       258.950           .
SC                285.842       272.873           .
-2 LOG L          280.361       250.950        29.411 with 3 DF (p=0.0001)
Score                .             .           28.420 with 3 DF (p=0.0001)

                   Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio

INTERCPT 1     1.1579   0.1711    45.7976     0.0001            .      .
LINEAR   1     0.3866   0.0829    21.7658     0.0001     0.477642    1.472
QUAD     1     0.0411   0.1711     0.0576     0.8103     0.022694    1.042
CUBIC    1     0.0390   0.0696     0.3149     0.5747     0.048236    1.040
```

The next output listing was generated from the third logistic model and includes only the linear contrast. The parameter estimate is slightly different and the chi-square for covariates is slightly smaller.

```
                      Fertilizer study
                Logistic Regression with linear contrast only

                      The LOGISTIC Procedure
                 Testing Global Null Hypothesis: BETA=0

                             Intercept
                 Intercept      and
Criterion          Only      Covariates    Chi-Square for Covariates

AIC               282.361       255.275          .
SC                285.842       262.236          .
-2 LOG L          280.361       251.275       29.087 with 1 DF (p=0.0001)
Score                .             .          27.686 with 1 DF (p=0.0001)

                 Analysis of Maximum Likelihood Estimates
              Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio

INTERCPT 1     1.1495   0.1666    47.6331     0.0001            .     .
LINEAR   1     0.3784   0.0758    24.9504     0.0001     0.467472    1.460
```

The second model, with just the linear contrast fits the data just as well as the model with all three contrasts (the difference of χ^2 = 29.411, 3 *df* versus χ^2 = 29.087, 1 *df* yields χ^2 = 0.324, 2 *df*). This is clearly not significant so that a linear model provides an adequate fit. The predicted response equation on the logit scale is:

$$\text{logit}(\text{survival}) = 1.1495 + 0.3784 \times \text{Linear}.$$

The following table shows the predicted logits and probabilities for the four treatments:

Values of “Linear”	Logit equation	Logit	Probability of survival
−3	1.1495 + 0.3784(−3)	0.0143	0.504
−1	1.1495 + 0.3784(−1)	0.7711	0.684
1	1.1495 + 0.3784(1)	1.5279	0.822
3	1.1495 + 0.3784(3)	2.2847	0.908

Note that because the response is linear on the logit scale the differences between successive treatments are constant at 2(0.3786) = 0.76. This is not the case on the probability scale, where differences between successive treatments are 0.180, 0.138, and 0.086.

6.3 One-way Classification with Covariable: Root Collar Weevil Study

Fitting this model using PROC LOGISTIC should now be straightforward to set up. Since the categorical variable for this example has only two values (levels), it can be used as the indicator variable. Nevertheless, to directly compare the parameter estimates between the output from PROC LOGISTIC and PROC CATMOD requires changing the treatment variable so that it has values of 1 and –1. The statistical tests are the same regardless of which values we use for the indicator variables. Many models were fit to this data in section 4.5, but only the parallel lines and two radiating lines models will be fit as examples. A program to do this is:

```
title ‘Root Collar Weevil Example’;
proc format; value y    0 =’0: Not Attacked’  1 = ‘1: Attacked’;
             value trmt 0 = ‘Screefed’        1 = ‘Control (1)’;     run;
data weevil;
  infile ‘weevil.dat’;
  input trmt plot count weevil;
  total = 16;
  trmta = (trmt = 0) - (trmt = 1);  * <=== Creating a new treatment variable;
  wvlcon = (trmt=0)* weevil;  wvltrt = (trmt=1)*weevil;
label trmt = ‘Treatment’  weevil = ‘Numbers of Weevils’
     count= ‘Seedlings Attacked’ plot =’Plot Number’;
  format trmt trmt.;
run;
proc logistic ;
  model count/total = trmt weevil;
  title2 ‘Model with parallel lines - with old treatment variable’;
proc logistic ;
  model count/total = trmta weevil;
  title2 ‘Model with parallel lines’;
proc logistic;
  model count/total = wvlcon wvltrt;
  title2 ‘Two Radiating Lines Model’;
run;
```

Some of the output from this program is shown below.

```
                                 Root Collar Weevil Example
                   Model with parallel lines - with old treatment variable

                                   The LOGISTIC Procedure

         Data Set: WORK.WEEVIL
         Response Variable (Events): COUNT     Seedlings Attacked
         Response Variable (Trials): TOTAL
         Number of Observations: 28
         Link Function: Logit

                                       Response Profile

                                 Ordered  Binary
                                   Value  Outcome      Count

                                       1  EVENT           87
                                       2  NO EVENT       361

                           Testing Global Null Hypothesis: BETA=0

                                      Intercept
                        Intercept        and
          Criterion       Only       Covariates    Chi-Square for Covariates

          AIC             443.057       273.662         .
          SC              447.162       285.976         .
          -2 LOG L        441.057       267.662      173.395 with 2 DF (p=0.0001)
          Score              .             .         131.989 with 2 DF (p=0.0001)

                          Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio

INTERCPT 1    -7.5321   0.8411    80.1897     0.0001           .    0.00054
TRMT     1     3.4104   0.4388    60.4197     0.0001     0.941185   30.278
WEEVIL   1     0.3086   0.0469    43.2031     0.0001     1.042599    1.361
```

```
                                 Root Collar Weevil Example
                                  Model with parallel lines

                           Testing Global Null Hypothesis: BETA=0

                                      Intercept
                        Intercept        and
          Criterion       Only       Covariates    Chi-Square for Covariates

          AIC             443.057       273.662         .
          SC              447.162       285.976         .
          -2 LOG L        441.057       267.662      173.395 with 2 DF (p=0.0001)
          Score              .             .         131.989 with 2 DF (p=0.0001)

                          Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized     Odds
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate      Ratio

INTERCPT 1    -5.8269   0.7032    68.6579     0.0001            .    0.00295
TRMTA    1    -1.7052   0.2194    60.4197     0.0001    -0.941185    0.182
WEEVIL   1     0.3086   0.0469    43.2031     0.0001     1.042599    1.361

                                 Root Collar Weevil Example
                                  Two Radiating Lines Model

                           Testing Global Null Hypothesis: BETA=0

                                      Intercept
                        Intercept        and
          Criterion       Only       Covariates    Chi-Square for Covariates

          AIC             443.057       272.997         .
          SC              447.162       285.311         .
          -2 LOG L        441.057       266.997      174.060 with 2 DF (p=0.0001)
          Score              .             .         171.359 with 2 DF (p=0.0001)

                        Analysis of Maximum Likelihood Estimates

                Parameter   Standard      Wald        Pr >     Standardized    Odds
Variable   DF    Estimate     Error    Chi-Square  Chi-Square    Estimate      Ratio

INTERCPT   1      -4.8370     0.6360      57.8407      0.0001       .          0.0079
WVLCON     1       0.1358     0.0477       8.0946      0.0044      0.492369    1.145
WVLTRT     1       0.3635     0.0493      54.2879      0.0001      1.297800    1.438
```

The results for PROC LOGISTIC and PROC CATMOD (see section 5.6.3) are identical. If the treatment variable (TRMTA) is defined in the same way as PROC CATMOD sets up the indicator variable, then the parameter estimates are also the same (see Table 25, page 60).

This chapter[23] describes alternative methods to analyze the data in the one-way classification study of root rot and stand age (section 4.2.2). The purpose is to show the similarity of logistic regression and contingency tables, and how to do the calculations by hand for a simple design. Section 7.1 uses simple contingency tables to calculate the necessary statistics and perform most of the statistical tests that were done in section 4.2. While contingency tables are easier to understand, logistic regression has the advantage of more easily analyzing complicated models. Section 7.2 will go through the steps necessary to perform the logistic regression by hand. This may help the reader to understand the process behind logistic regression.

Recall that this one-way classification study was designed to study the effect of stand age on the survival of trees in stands infected with similar levels of root rot (see section 4.2). Nine stands were grouped into three age levels (young-, middle-, and older-aged stands). The response variable was Y, where "zero" represents that the tree died and "one" represents that it was still alive when sampled. The number of trees in each category for each stand was stored in a variable called "count".

7.1 Analysis Using Simple Contingency Tables

For this simple design, contingency table methods can arrive at similar conclusions as the logistic regression analysis of section 4.2.2. This is done by fitting various models to the data and then testing hypotheses about parts of the models by comparing models. The hypotheses that we will examine are:

- **Hypothesis 1:** The stands within each age group respond similarly. If the stands within each age group have similar proportions of survival, then this suggests that pooling the stands within each age group is acceptable. This is similar to the homogeneity of variance assumption within ANOVA.
- **Hypothesis 2:** The responses are similar for the three age groups. Combined with an acceptance[24] of Hypothesis 1, this would imply that no differences between the nine stands exist. Rejecting this hypothesis provides evidence that survival and stand age are correlated (but cause and effect statements are not possible because this is an observational study).

If hypothesis 2 is rejected, then we will test for specific differences in stand ages by:

- **Hypothesis 3:** The middle-aged response is similar to the old-aged response.

[23] This chapter is optional and is intended primarily to deepen your understanding of logistic regression.

[24] Strictly speaking, we cannot *accept* a hypothesis, only fail to reject it. Nevertheless, we must make decisions and choose a model for our data. Accordingly, we must accept some hypotheses, although from a purely statistical point of view, we can never do so.

- **Hypothesis 4:** The young-aged response is similar to the average of the middle- and old-aged responses.

These hypotheses can be tested by calculating several tables that pool stand data in various ways. Seven contingency tables must be calculated for the data:

1. one table with all nine stands by *Y*;
2. age by *Y*;
3. age reduced to two levels (young vs older) by *Y*;
4. age by *Y* for the middle and old stands only; and

5. to 7. a separate stand by *Y* table for each age.

The data are presented in Table 30 (page 85) in a *Y* by stand table. All of the contingency tables represent various ways of pooling or removing stands. The results for the seven contingency tables are summarized in Table 35. All the necessary numbers are obtained from the SAS program output shown below. The first contingency table, stand by *Y*, is the saturated model, while the second contingency table is the three group model. Essential SAS statements appear in boldface type.

```
title ‘Simple One-Way Classification Example’;
data one;  set one;
  a2 = 1 + (stand ge 3);  ** Pooling Middle and Old into one level;
run;                      ** Could also have done this with Proc Format;
proc freq data = one;
  weight count;
  tables  stand*y        /*  Result #1 */
          age*y          /*  Result #2 */
          a2*y           /*  Result #3 */
         / nocol nopercent chisq;
title2 ‘Frequency Counts’;
run;
proc freq data=one;
  weight count;  where age ge 2;
  table  age*y           /*  Result #4 */
         / nocol nopercent chisq;
title2 ‘Testing Middle vs Old’;
run;
proc freq; by age;
  weight count;
  table  stand*y         /*  Result #5 */
         / nocol nopercent chisq;
run;
```

TABLE 35 *Results of contingency table analyses*

Result no.	Contingency table	Degrees of freedom	Summary χ^2 statistics			
			Pearson's	p-value	Likelihood	p-value[a]
1	Stand	8	$\chi^2 = 6.17$	0.63	$\chi^2 = 6.356$	0.61
2	Age: young, middle and old	2	$\chi^2 = 5.037$	0.081	$\chi^2 = 5.128$	0.077
3	Age: young vs. middle and old	1	$\chi^2 = 4.342$	0.037	$\chi^2 = 4.458$	0.035
4	Age: middle vs. old	1	$\chi^2 = 0.668$	0.41	$\chi^2 = 0.670$	0.41
5	By age:					
	Age = young	1	$\chi^2 = 0.740$	0.39	$\chi^2 = 0.746$	0.39
	Age = middle	2	$\chi^2 = 0.162$	0.92	$\chi^2 = 0.162$	0.92
	Age = old	3	$\chi^2 = 0.318$	0.96	$\chi^2 = 0.320$	0.96
6	Sum:	6	$\chi^2 = 1.220$	0.98	$\chi^2 = 1.228$	0.98

[a] Probability values not available from printouts have been calculated using Biometrics Information Pamphlet #15.

Table 35 shows the Pearson's `Chi-square` and the `Likelihood Ratio Chi-square` values for each contingency table. Results 1 to 4 correspond to the first four contingency tables, respectively. Result 5 lists the results for tables 5, 6, and 7, while their sum is called Result #6. These results are used to test the hypotheses. Several of the hypotheses can be tested in different ways:

- **Hypothesis 1:** H_0: The stands within each age group respond similarly.

Test: 1. Each test in the By Age section (Result #5) of Table 35 is a test of this hypothesis. This hypothesis is not rejected for any of the ages.

2. The individual tests can be pooled for an overall test. This is presented as Result #6 in Table 35. Since both χ^2-values are less than their respective degrees of freedom, this test also does not reject the null hypothesis.

3. An overall test is calculated as shown in Table 36. This test compares the saturated model with the three group model. Since the three group model constrains the stands within each group to have the same predicted probability of survival, the null hypothesis is that the stands within each group have the same response. The observed χ^2 for the difference between the two models is small (1.133 and 1.22) showing little evidence against the null hypothesis.

Conclusion: It is reasonable to treat the stands within each group as similar so that stands within ages may be pooled.

TABLE 36 *Calculations required to test the similarity of the stands within each age group*

Result no.	Contingency table	Degrees of freedom	Summary statistics			
			Pearson's	p-value	Likelihood	p-value
1	Stand	8	$\chi^2 = 6.17$	0.63	$\chi^2 = 6.356$	0.61
2	Age (3 levels)	2	$\chi^2 = 5.037$	0.081	$\chi^2 = 5.128$	0.077
	Difference	6	$\chi^2 = 1.133$	0.98	$\chi^2 = 1.228$	0.98

- **Hypothesis 2:** H_0: The responses are similar for all ages.

  Test: This is tested by Result #2 in Table 35.

  Conclusion: Both χ^2-values show little evidence against the null hypothesis (p-value = 0.081 and p-value = 0.077). To understand the age effect better, look at the values in the Age by Y table (Table 37).

TABLE 37 *The numbers of trees that are alive or dead for each age and corresponding percentages*

```
------------------------------------------------------
                 TABLE OF AGE BY Y

        age(Age)     Y(Survive?)

        Frequency |
        Row Pct   |0..Dead |1..Alive|  Total
        ----------+--------+--------+
        1..Young  |     21 |     55 |     76
                  |  27.63 |  72.37 |
        ----------+--------+--------+
        2..Middle |     31 |     50 |     81
                  |  38.27 |  61.73 |
        ----------+--------+--------+
        3..Old    |     40 |     50 |     90
                  |  44.44 |  55.56 |
        ----------+--------+--------+
        Total           92      155      247
------------------------------------------------------
```

Evidence from this table and the previous tests suggests that the young stands may differ from the middle and old stands, while the middle and old stands may not. A linear trend in response may also exist, but other methods are required to test that hypothesis (see section 4.2).

- **Hypothesis 3:** H_0: Response to the middle age level is similar to that of the old age.

  Test: 1. This is tested by Result #4 in Table 35.
  2. This can also be tested by comparing Results #2 and #3. Since Result #3 is for a model with only two age levels (young vs middle and old) and Result #2 does not constrain the responses for the three age levels, the difference between them is a test for a similar response between middle-aged stands and old stands as shown in Table 38.

  Conclusion: Both χ^2-values show little evidence for a difference in response between the middle and old age levels (p-value = 0.41).

TABLE 38 *Calculations required to test whether the middle- and old-aged stands are different*

Result	Contingency	Degrees of	Summary statistics			
no.	table	freedom	Pearson's	p-value	Likelihood	p-value
2	Age: young, middle and old	2	$\chi^2 = 5.037$	0.081	$\chi^2 = 5.128$	0.077
3	Age: young vs. middle and old	1	$\chi^2 = 4.342$	0.037	$\chi^2 = 4.458$	0.035
	Difference	1	$\chi^2 = 0.695$	0.40	$\chi^2 = 0.670$	0.41

- **Hypothesis 4**: H_0: The response in young stands is similar to that of both middle- and old-aged stands.

  Test: This is directly tested by Result #3 in Table 35.

  Conclusion: Both χ^2-values show strong evidence for a difference between the response of young stands and that of middle and old stands (p-value = 0.035 and p-value = 0.037). This suggests that most of the age effect occurs in higher survival rates for trees in young stands rather than middle- or old-aged stands.

7.2 Analysis Using Hand Calculations

This section is quite mathematical, but is useful to study if greater understanding of the maximum likelihood method is desired. The saturated, three group, and one group models are compared by determining their log-likelihood equations and calculating the values by hand. The maximization process described here will be familiar to those who remember some first year calculus.

Recall that in this one-way classification study we assume each stand's response to stand age will follow a binomial distribution with parameters π_{ij} and m_{ij}. The parameter π_{ij} is the probability of survival for each tree within an infected stand and may be different for each level of age i (i = 1, 2, and 3) and each stand j (j = 1, 2, . . . , 9). The number of trees within each stand is m_{ij} and is assumed fixed (although this may be unlikely for this example). The probability of observing y_{ij} trees survive is given by the binomial distribution:

$$\text{P(observing } y_{ij} \text{ given } \pi_{\text{ij}} \text{ and } m_{ij}) = \left[\begin{matrix} m_{ij} \\ y_{ij} \end{matrix}\right] \pi_{ij}{}^{y_{ij}} \, (1 - \pi_{ij})^{(m_{ij} - y_{ij})}.$$

The *given* in the above equation indicates that we assume we know all the π_{ij}'s and m_{ij}'s and is often represented by the symbol "|".

If each stand's response is independent of any other stand's response, then the probability for the complete set of responses y_{11}, y_{12}, y_{23}, . . . y_{38}, y_{39} is obtained by multiplying together all the individual stand's probabilities. This can be written as:

$$\begin{aligned} &\text{P}(y_{11}, \ldots, y_{39} | p_{11}, \ldots, p_{39} \text{ and } m_{11}, \ldots, m_{39}) \\ &= \prod_{i=1}^{3} \prod_{j} \left[\begin{matrix} m_{ij} \\ y_{ij} \end{matrix}\right] \pi_{ij}{}^{y_{ij}} \, (1 - \pi_{ij})^{(m_{ij} - y_{ij})}, \end{aligned}$$

where the product symbol means that the following terms are to be multiplied, just as a summation symbol means to add the following terms.

This probability function can predict the values of y_{ij} that would be expected if π_{ij} and m_{ij} were known. However, when faced with a set of data, the y_{ij}, and not π_{ij}, are known! In this situation, we can look at the probability differently. Instead of viewing the above equation as "the probability of observing various y_{ij} values, given π_{ij}'s and m_{ij}'s," we can consider it as a function of the unknown π_{ij} and say: "Given that we know all the y_{ij}'s and m_{ij}'s, what values of the π_{ij}'s would maximize the above function?" The values of π_{ij} that maximize this function also maximize the probability of observing the y_{ij}'s which were observed. These values are called the *maximum likelihood estimates* because they maximize the associated likelihood of the particular sample that has been observed. When the probability function is interpreted in this way, it is called the *likelihood function* and labelled $L(\pi_{ij} | y_{ij}$ and $m_{ij})$:

$$\mathrm{L}(\pi_{ij} | y_{ij},\, m_{ij}) = \prod_{i=1}^{3}\prod_{j} \left[\begin{matrix} m_{ij} \\ y_{ij} \end{matrix}\right] \pi_{ij}^{\,y_{ij}}\,(1 - \pi_{ij})^{(m_{ij} - y_{ij})}.$$

The log-likelihood function:

$$l(\pi_{ij} | y_{ij},\, m_{ij}) = \sum_{i}^{3}\sum_{j} \left\{ (y_{ij}\ \log\ \pi_{ij} + (m_{ij} - y_{ij})\ \log(1 - \pi_{ij}) \right\},$$

is generally used for calculations, instead of the likelihood, because the maximization process uses simpler summation, but yields the same answer. The log-likelihood function given above is missing the term $\sum_{i}\sum_{j} \log \left[\begin{matrix} m_{ij} \\ y_{ij} \end{matrix}\right]$. This term is constant if m_{ij} is fixed and known. It is not necessary in the calculations because it does not depend on π_i and therefore does not affect the result of the maximization process. The notation for the likelihood, $\mathrm{L}(\pi_{ij} | y_{ij},\, m_{ij})$, is often shortened further to $\mathrm{L}(\pi_{ij})$. Similarly, the notation for the log-likelihood, $l(\pi_{ij} | y_{ij},\, m_{ij})$ is shortened to $l(\pi_{ij})$.

Maximum likelihood estimates $(\hat{p}_{ij})$ for the probabilities (π_{ij}) are obtained by differentiating the log-likelihood with respect to the parameters[26] and setting the resulting derivatives to zero. Different estimators will be obtained depending on the restrictions that are placed on the π_{ij}. The three models described earlier restrict the possible values for the parameters in the following way:

Model	Restrictions
1. Saturated	No restrictions
2. Three groups	$\pi_1 = \pi_{11} = \pi_{12}$
	$\pi_2 = \pi_{23} = \pi_{24} = \pi_{25}$
	$\pi_3 = \pi_{36} = \pi_{37} = \pi_{38} = \pi_{39}$
3. One group	$\pi\ = \pi_{11} = \pi_{12} = \pi_{23} = \ldots = \pi_{39}$

[26] The parameters are the π_{ij} only, since the m_{ij} are assumed to be fixed and known.

The corresponding log-likelihoods are:

Model	Log-likelihood formulae
1. Saturated	$l(\pi_{ij}) = \sum_{i=1}^{3}\sum_{j}\left\{y_{ij} \log \pi_{ij} + (m_{ij} - y_{ij}) \log(1 - \pi_{ij})\right\}$
2. Three groups	$l(\pi_i) = \sum_{i=1}^{3}\left\{(\sum_j y_{ij})\log \pi_i + [\sum_j(m_{ij} - y_{ij})] \log(1 - \pi_i)\right\}$
3. One group	$l(\pi) = (\sum_i\sum_j y_{ij})\log \pi + [\sum_i\sum_j(m_{ij} - y_{ij})] \log(1 - \pi)$

The derivatives and estimators are:

Model	Derivatives	Estimators
1. Saturated	$\frac{\partial[1(\pi_{ij})]}{\partial\pi_{ij}} = \frac{y_{ij}}{\pi_{ij}} - \frac{(m_{ij} - y_{ij})}{(1 - \pi_{ij})} = 0$	if $\hat{p}_{ij} = y_{ij}/m_{ij}$
2. Three Groups	$\frac{\partial[1(\pi_i)]}{\partial\pi_i} = \frac{\sum_j y_{ij}}{\pi_{ij}} - \frac{\sum_j(m_{ij} - y_{ij})}{(1 - \pi_{ij})} = 0$	if $\hat{p}_i = \sum_j y_{ij}/\sum_j m_{ij}$
3. One Group	$\frac{\partial[1(\pi)]}{\partial\pi} = \frac{\sum_i\sum_j y_{ij}}{\pi_{ij}} - \frac{\sum_i\sum_j(m_{ij} - y_{ij})}{(1 - \pi_{ij})} = 0$	if $\hat{p} = \sum_i\sum_j y_{ij}/\sum_i\sum_j m_{ij}$

The reader can confirm that these estimates are the maximum and not the minimum by checking that the second derivatives are negative.

These equations are used to calculate the various estimates of the mortality probability ($\hat{p}$, $\hat{p}_i$, and $\hat{p}_{ij}$) by using the data in Table 10 (page 31). For instance, $\hat{p} = \sum_i\sum_j y_{ij}/\sum_i\sum_j m_{ij} = 92/247 = 0.372$, $\hat{p}_1 = \sum_j y_{1j}/\sum_j n_{1j} = (13 + 8)/(41 + 35) = 21/76 = 0.276$, $\hat{p}_{11} = y_{11}/n_{11} = 13/41 = 0.317$, and $\hat{p}_{36} = y_{36}/n_{36} = 7/16 = 0.4375$. The log-likelihood for $\hat{p}$ is $l(\hat{p}) = (\sum_i\sum_j y_{ij})\log \hat{p} + (\sum_i\sum_j(m_{ij} - y_{ij})) \log(1 - \hat{p}) = 92{*}\log(92/247) + 155{*}\log(1 - 92/247) = -90.8592 + -72.2243 = -163.0835$. All the other calculations are done similarly. The results are:

Model		Estimates	Number of parameters	Log-likelihood	–2LogL
1. Saturated	Young:	0.32, 0.23			
	Middle:	0.36, 0.41, 0.38			
	Old:	0.44, 0.39, 0.46, 0.47	9	–159.9054	319.8108
2. Three groups	Young:	0.28			
	Middle:	0.38			
	Old:	0.44	3	–160.5196	321.0392
3. One group		0.37	1	–163.0835	326.1670

The following tests can be conducted:

Test	Models used	Deviance[a]	*df*	*p*-values
Goodness of fit for one group	1 and 3	6.356	8	0.61
Goodness of fit for three groups	1 and 2	1.228	6	0.98
Group differences	2 and 3	5.128	2	0.077

[a] Notice that these deviances have the same value as the L.R. χ^2 statistic in the contingency tables, Results #1, #6, and #2 respectively in Table 35.

What we can conclude from these tests is that both the three group and one group models adequately fit the data, although there is weak evidence (p-value = 0.08) for group differences. Linear and quadratic trends for the groups assume it is possible to estimate numeric values for their ages, but the calculations are more difficult (see section 4.2 for the results and discussion).

7.3 Comparing Hand Calculations with Computer-generated Results

The log-likelihood values and the final parameter estimates determined above can be obtained from the `PROC CATMOD` output by examining the last line of the iteration history. These histories are presented in Table 39 (SAS programs were given in section 5.3). The column titled `–2 Log Likelihood` is the log-likelihood multiplied by –2. This function, –2log-likelihood, has an approximate chi-square distribution.

Note that the last value of the `–2 Log Likelihood` for the three group model is `321.03921` which is –160.5196 when divided by –2. This is the value computed previously. The other calculations for the log-likelihood can be checked by examining the other iteration histories.

The parameter estimates can be checked by first adding together the appropriate parameter estimates. The inverse logit function (equation 2) is then used to obtain the probabilities. The calculations for the three group model are:

Young: logit = –0.5547 – 0.4081 = –0.9628;
prob = exp(–0.9628)/[1 + exp(–0.9628)] = 0.2763.
Middle: logit = –0.5547 + 0.0766 = –0.4781;
prob = exp(–0.4781)/[1 + exp(–0.4781)] = 0.3827.
Old: logit = –0.5547 + 0.4081 – 0.0766 = –0.2232;
prob = exp(–0.2232)/[1 + exp(–0.2232)] = 0.4444.

TABLE 39 *Iteration histories for the three models*

```
--------------------------------------------------------------------------
                    Simple One-Way Classification Example
                               Saturated Model

                          MAXIMUM LIKELIHOOD ANALYSIS

              Sub       -2 Log    Convergence     parameter Estimates
Iteration  Iteration  Likelihood   Criterion      1        2         3
------------------------------------------------------------------------
     0          0      342.41471      1.0000         0        0        0
     1          0      319.93263      0.0657   -0.4713  -0.2604  -0.6145
     2          0        319.811    0.000380   -0.4938  -0.2732  -0.7183
     3          0      319.81088   3.5348E-7   -0.4943  -0.2730  -0.7221
     4          0      319.81088   4.699E-13   -0.4943  -0.2730  -0.7221

                             parameter Estimates
 Iteration      4         5         6         7         8         9
 ----------------------------------------------------------------------
      0             0         0         0         0         0         0
      1       -0.1002    0.0929   -0.0287    0.2213    0.0268    0.3249
      2       -0.0940    0.1108   -0.0170    0.2425    0.0418    0.3472
      3       -0.0935    0.1113   -0.0165    0.2430    0.0423    0.3477
      4       -0.0935    0.1113   -0.0165    0.2430    0.0423    0.3477

                              Three Group Analysis

                           MAXIMUM-LIKELIHOOD ANALYSIS

              Sub      -2 Log   Convergence      parameter Estimates
 Iteration Iteration Likelihood  Criterion       1         2         3
 ------------------------------------------------------------------------
      0         0     342.41471      1.0000          0         0         0
      1         0     321.11188      0.0622    -0.5287   -0.3660    0.0596
      2         0     321.03923    0.000226    -0.5543   -0.4075    0.0763
      3         0     321.03921   4.5882E-8    -0.5547   -0.4081    0.0766
      4         0     321.03921   2.125E-15    -0.5547   -0.4081    0.0766

                   Model with no structure -- one group only

                           MAXIMUM-LIKELIHOOD ANALYSIS

                                                             parameter
                         Sub        -2 Log     Convergence   Estimates
          Iteration   Iteration   Likelihood    Criterion         1
          -------------------------------------------------------------
               0           0       342.41471       1.0000            0
               1           0       326.17462       0.0474      -0.5101
               2           0       326.16695    0.0000235      -0.5216
               3           0       326.16695     4.91E-11      -0.5216
--------------------------------------------------------------------------
```

8 CONCLUSIONS

Logistic regression is a data analysis method that can be of great use in forestry research problems. Many of the concepts associated with the general linear models of ANOVA, regression, and ANCOVA are already familiar and can be readily translated for use with logistic regression. This handbook has used simple design examples to show the differences and similarities between this new method of logistic regression and the better-known traditional data analysis methods. I hope that it will encourage readers to use logistic regression, where appropriate, in their data analysis and to study the technique more deeply by reading such texts as Agresti (1996).

While running a series of logistic regression analyses, a scientist (Les Peterson, B.C. Ministry of Forests, pers. comm., 1988) noted a curious inconsistency with his results. The following two sets of data illustrate the problem:

	Set 1				Set 2			
	Number of successes per e.u.			Overall percent success	Number of successes per e.u.			Overall percent success
Group 1	4	3	3	33	4	3	3	33
Group 2	9	10	10	97	10	10	10	100
Sample size:	10 per e.u.				10 per e.u.			

The results from PROC CATMOD are:

```
                     MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

                               Set  1                          Set  2
                     -------------------------      -------------------------
Source               DF   Chi-Square      Prob      DF   Chi-Square      Prob
-----------------------------------------------------------------------------
INTERCEPT             1         6.04    0.0140      1       188.23    0.0000
GROUP                 1        13.92    0.0002      0*          .       .

LIKELIHOOD RATIO      4         2.56    0.6340      5         0.30    0.9977
```

with the second set obtaining this warning:

```
NOTE: Effects marked with ‘*’ contain one or more
      redundant or restricted parameters.
```

The parameter estimates are:

```
                         ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

                              Set  1                             Set  2
                  -----------------------------    ----------------------------------
                          Standard  Chi-                    Standard    Chi-
Effect  Parameter Estimate  Error  Square   Prob   Estimate    Error    Square   Prob
-------------------------------------------------------------------------------------
INTERCEPT       1   1.3370  0.5441   6.04 0.0140    5.3134    0.3873   188.23  0.0000
GROUP           2  -2.0301  0.5441  13.92 0.0002   -6.0066 #        .      .     .
```

with the second set obtaining this note:

```
NOTE: Parameters marked with ‘#’ are regarded to be
infinite.
```

Significant differences between the two groups were found for data set 1, but no results were obtained for set 2, although the differences between the overall percent success was greater. This occurs because chi-square tests for group differences and the intercept are the approximate Wald’s tests. These are unreliable when complete failure or success occurs in one or more treatment groups. As discussed in section 3.5, these tests should be checked with the more reliable deviance statistics. For this example, the deviance is calculated as twice the difference between the log-likelihood of a model with two groups and that of a model with one group. As described in section 7.2 and 7.3, this can be calculated by hand or the necessary –2LogL values obtained from the iteration history output by `PROC CATMOD`. The hand calculations are as shown below:

Log-likelihood for one group model:†

Set 1: $39*\ln(39/60) + (60 - 39)*\ln[(60 - 39)/60] = -38.8468$
Set 2: $40*\ln(40/60) + (60 - 40)*\ln[(60 - 40)/60] = -38.1909$

Log-likelihood for two group model:

Set 1: $10*\ln(10/30) + (30 - 10)*\ln[(30 - 10)/30]$
$+ 29*\ln(29/30) + (30 - 29)*\ln[(30 - 29)/30]$
$= -19.0954 + -4.3843 = -23.4798$
Set 2: $10*\ln(10/30) + (30 - 10)*\ln[(30 - 10)/30] + 0 = -19.0954$

The deviance for each set is calculated by:

Set 1: Deviance $= 2*(-23.4798 - -38.8468) = 30.7$
Set 2: Deviance $= 2*(-19.0954 - -38.1909) = 38.2$

The degrees of freedom for the χ^2-statistics is one (number of treatments minus one) and the critical value at the 95% confidence level (i.e., $\alpha = 0.05$) is 3.84. Hence, there is very strong evidence for group differences in both sets of data, and it is stronger for the second set.

For this simple example, contingency tables as output by `PROC FREQ` with the `CHISQ` option will produce correct likelihood ratio tests. Another suggestion is to run `PROC LOGISTIC` after all zero values are changed to a small value. This produces the correct deviance tests as standard output. An example program is as follows:

† The methods for these hand calculations are described in Section 7.2.

```
Title ‘Example failure of Wald’’s test’;

data waldlog;
  do set = 1 to 2;
    do group = 1 to 2;
      do rep = 1 to 3;
      input ct @@;
      if ct = 0  then ct =      1E-4;
      if ct = 10 then ct = 10 - 1E-4;
      eu + 1; ten = 10;
  output; end; end; end;
datalines;
4 3 3 9 10 10 4 3 3 10 10 10
;
proc logistic data=waldlog; by set;
  model ct/ten = group;
title2 ‘Logistic Regression Approach’;
run;
```

The relevant parts of the output are:

```
                         Testing Global Null Hypothesis: BETA=0
                            Set 1                                Set 2
           -------------------------------------  -------------------------------------
                       Intercept                             Intercept
           Intercept      and     Chi-Square      Intercept     and       Chi-Square
Criterion    Only     Covariates  for Covariates     Only    Covariates  for Covariates

AIC           79.694      50.961        .           78.382        42.198          .
SC            81.788      55.150        .           80.476        46.387          .
-2 LOG L      77.694      46.961     30.733         76.382        38.198       38.184
Score            .            .      26.446            .             .         29.999
```

Both −2LogL values have one degree of freedom with p-value = 0.0001.

Caribou calf data:

```
1 9 14 15
2 10 7 7
3 12 3 4
4 13 5 5
5 15 9 10
6 23 9 10
7 31 9 15
8 34 4 13
9 38 1 13
```

Weevil data:

```
0  1  0 10
0  2  1 13
0  3  0 11
0  4  2 16
0  5  1 16
0  6  1  9
0  7  0  6
0  8  1 14
0  9  1 17
0 10  1 19
0 11  0  0
0 12  0  3
0 13  0  2
0 14  0  1
1  1  6 12
1  2  6 12
1  3 12 17
1  4  7 12
1  5  9 14
1  6  7 12
1  7  7 12
1  8  3  8
1  9  9 14
1 10 13 19
1 11  0  2
1 12  0  1
1 13  0  0
1 14  0  1
```

Aspen data:

```
11.0  2. 100. 1.
11.0  2. 100. 1.
12.0  2. 100. 1.
10.5  3. 100. 1.
11.5  3. 100. 1.
12.0  3. 100. 1.
12.0  3. 100. 1.
13.0  3. 100. 1.
13.0  3. 100. 1.
13.0  3. 100. 1.
13.5  3. 100. 1.
14.5  3. 100. 1.
15.0  3. 100. 1.
15.5  3. 100. 1.
16.0  3.  70. 0.
16.5  3. 100. 1.
16.5  3. 100. 1.
16.5  3.  65. 0.
21.5  3.  90. 0.
12.0  4. 100. 1.
12.0  4. 100. 1.
14.0  4. 100. 1.
14.0  4. 100. 1.
14.5  4. 100. 1.
15.0  4. 100. 1.
15.0  4. 100. 1.
16.0  4. 100. 1.
16.5  4.  95. 1.
17.5  4. 100. 1.
18.0  4.  95. 1.
18.0  4. 100. 1.
18.0  4. 100. 1.
20.0  4.  90. 0.
20.0  4. 100. 1.
20.0  4.  65. 0.
20.5  4.  10. 0.
20.5  4.  40. 0.
21.0  4.  60. 0.
21.0  4. 100. 1.
21.0  4. 100. 1.
23.0  4.  35. 0.
23.0  4.  90. 0.
25.5  4.  70. 0.
26.0  4.  70. 0.
27.5  4.  50. 0.
33.0  4.   5. 0.
17.0  5. 100. 1.
17.0  5. 100. 1.
19.0  5. 100. 1.
19.0  5. 100. 1.
20.0  5. 100. 1.
20.5  5. 100. 1.
21.0  5. 100. 1.
22.5  5.  90. 0.
23.0  5.  75. 0.
23.0  5.  85. 0.
24.0  5.  80. 0.
26.0  5.  60. 0.
27.0  5.  70. 0.
27.0  5.  10. 0.
27.0  5.  20. 0.
27.0  5.  90. 0.
28.0  5.  10. 0.
29.5  5.   5. 0.
31.0  5.  20. 0.
19.5  6. 100. 1.
22.0  6. 100. 1.
23.0  6.  40. 0.
23.0  6. 100. 1.
25.0  6.  95. 1.
25.5  6.  30. 0.
26.5  6.  15. 0.
26.5  6.  10. 0.
27.0  6.  35. 0.
27.0  6.  35. 0.
27.0  6.   5. 0.
29.5  6.   5. 0.
30.5  6.   5. 0.
32.0  6.   5. 0.
24.0  7.  90. 0.
28.0  7. 100. 1.
28.0  7.  15. 0.
30.0  7.   5. 0.
30.5  7.  25. 0.
31.0  7.  10. 0.
31.0  7.  10. 0.
33.0  7.  20. 0.
33.0  7.  10. 0.
34.5  7.   5. 0.
24.5  8. 100. 1.
32.5  8.  20. 0.
42.0  8.   0. 0.
```

Agresti, Alan. 1996. Introduction to categorical data analysis. John Wiley and Sons, Toronto, Ont.

Bergerud, W.A. 1988. Dose response models for the ezject herbicide lance. B.C. Min. For., Victoria, B.C. Res. Note 102.

———. 1991. Pictures of linear models. B.C. Min. For., For. Res. Branch, Victoria, B.C. Biom. Info. Handb. No. 1.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. Discrete multivariate analysis: theory and practice. The MIT Press, Cambridge, Mass.

Breslow, N.E. and D.G. Clayton. 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88:9–25.

Breslow, N.E. and N.E. Day. 1980. Statistical methods in cancer research. Vol. 1: The analysis of case control studies. International Agency for Research on Cancer, Lyon, Switz. IARC Sci. Publ. No. 32.

Cox, D.R. 1970. The analysis of binary data. Methuen, London, Eng. [A classic].

Dobson, A. 1983. An introduction to statistical modeling. Chapman and Hall, London, Eng.

Fienberg, S.E. 1980. The analysis of cross-classified categorical data. 2nd ed. The MIT Press, Cambridge, Mass.

Follmann, D.A. and D. Lambert. 1989. Generalizing logistic regression by nonparametric mixing. J. Am. Stat. Assoc. 84:295–300.

Freeman, D.H., Jr. 1987. Applied categorical data analysis. Marcel Dekker, New York, N.Y.

Gilchrist, W. 1984. Statistical modeling. John Wiley and Sons, Toronto, Ont.

Hosmer, D.W. and S. Lemeshow. 1989. Applied logistic regression. John Wiley and Sons, Toronto, Ont.

McCullagh, P. and J.A. Nelder. 1983. Generalized linear models. Chapman and Hall, London, Eng.

———. 1989. Generalized linear models. 2nd ed. Chapman and Hall, London, Eng.

SAS Institute Inc. 1989. SAS/STAT user’s guide. Version 6. Vol. 2. 4th ed. SAS Institute Inc., Cary, N.C.

Wetherill, G.B. 1981. Intermediate statistical methods. Chapman and Hall, London, Eng.

Zackin, R., V. De Gruttola and N. Laird. 1996. Nonparametrics mixed-effects models for repeated binary data arising in serial dilution assays: an application to estimating viral burden in AIDS. J. Am. Stat. Assoc. 91:52–61.

Relevant Biometrics Information Pamphlets:

Available from the B.C. Ministry of Forests, Research Branch, Biometrics Section, Victoria, B.C.

No. 5 Understanding replication and pseudo-replication
No. 12 Determining polynomial contrast coefficients
No. 14 ANOVA: factorial designs with a separate control
No. 15 Using SAS to obtain probability values for F-, t- and χ^2-statistics
No. 16 ANOVA: contrasts viewed as t-tests
No. 17 What is the design?
No. 23 ANOVA: Contrasts viewed as correlation coefficients
No. 27 When the t-test and F-test are equivalent
No. 36 Contingency tables and log-linear models
No. 41 Power analysis and sample size determination for contingency table tests
No. 55 Displaying factor relationships in experiments

Lecture 20 - Logistic Regression

Statistics 102

Colin Rundel

April 15, 2013

1. Background
2. GLMs
3. Logistic Regression
4. Additional Example

Regression so far ...

At this point we have covered:

- Simple linear regression
    - Relationship between numerical response and a numerical or categorical predictor

Regression so far ...

At this point we have covered:

- Simple linear regression
    - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
    - Relationship between numerical response and multiple numerical and/or categorical predictors

Regression so far ...

At this point we have covered:

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
  - Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven't seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Recap of what you should know how to do ...

- Model parameter interpretation
- Hypothesis tests for slope and intercept parameters
- Hypothesis tests for all regression parameters
- Confidence intervals for regression parameters
- Confidence and prediction intervals for predicted means and values (SLR only)
- Model diagnostics, residuals plots, outliers
- R^2, Adjusted R^2
- Model selection (MLR only)
- Simple transformations

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

Odds

For some event E,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

1. Background
2. GLMs
3. Logistic Regression
4. Additional Example

Example - Donner Party

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

Example - Donner Party - Data

	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs. Gender:

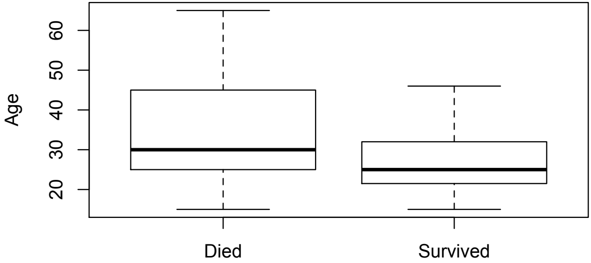
	Male	Female
Died	20	5
Survived	10	10

Example - Donner Party - EDA

Status vs. Gender:

	Male	Female
Died	20	5
Survived	10	10

Status vs. Age:



Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
3. A link function that relates the linear model to the parameter of the outcome distribution
   - $g(p) = \eta$ or $p = g^{-1}(\eta)$

1. Background
2. GLMs
3. Logistic Regression
4. Additional Example

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

Logit function

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \le p \le 1$$

Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Example - Donner Party - Model

In R we fit a GLM in the same was as a linear model except using `glm` instead of `lm` and we must also specify the type of GLM to fit using the `family` argument.

```
summary(glm(Status ~ Age, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
##
## Number of Fisher Scoring iterations: 4
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times \mathit{0}\\ \frac{p}{1-p} &= \exp(1.8185) = 6.16\\ p &= 6.16/7.16 = 0.86\end{aligned}$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times 25\\ \frac{p}{1-p} &= \exp(0.156) = 1.17\\ p &= 1.17/2.17 = 0.539\end{aligned}$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times \mathit{25}\\ \frac{p}{1-p} &= \exp(0.156) = 1.17\\ p &= 1.17/2.17 = 0.539\end{aligned}$$

Odds / Probability of survival for a 50 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

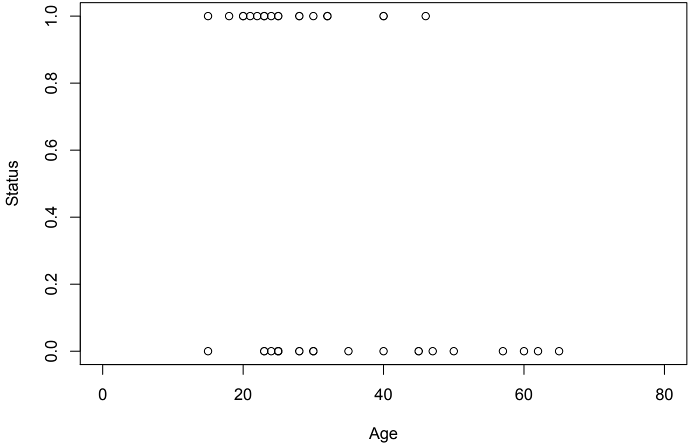
$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times \mathit{25}\\ \frac{p}{1-p} &= \exp(0.156) = 1.17\\ p &= 1.17/2.17 = 0.539\end{aligned}$$

Odds / Probability of survival for a 50 year old:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times \mathit{0}\\ \frac{p}{1-p} &= \exp(-1.5065) = 0.222\\ p &= 0.222/1.222 = 0.181\end{aligned}$$

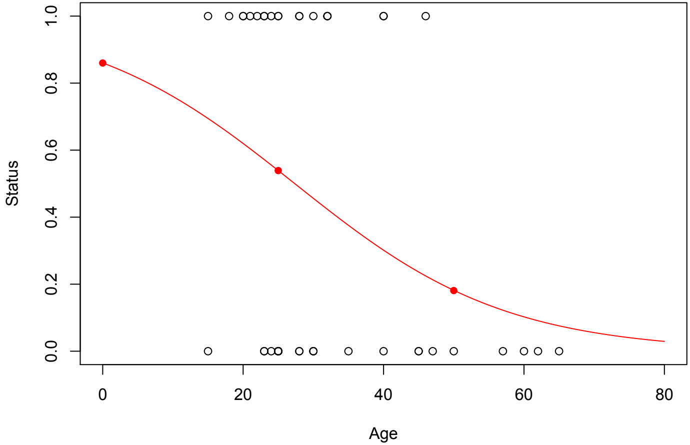
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Simple interpretation is only possible in terms of log odds and log odds ratios for intercept and slope terms.

Intercept: The log odds of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.

Slope: For a unit increase in age (being 1 year older) how much will the log odds ratio change, not particularly intuitive. More often then not we care only about sign and relative magnitude.

Example - Donner Party - Interpretation - Slope

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) &= 1.8185 - 0.0665(x+1) \\
&= 1.8185 - 0.0665x - 0.0665 \\
\log\left(\frac{p_2}{1-p_2}\right) &= 1.8185 - 0.0665x
\end{aligned}$$

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) &= -0.0665 \\
\log\left(\frac{p_1}{1-p_1}\bigg/\frac{p_2}{1-p_2}\right) &= -0.0665 \\
\frac{p_1}{1-p_1}\bigg/\frac{p_2}{1-p_2} &= \exp(-0.0665) = 0.94
\end{aligned}$$

Example - Donner Party - Age and Gender

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Gender slope: When the other predictors are held constant this is the log odds ratio between the given level (Female) and the reference level (Male).

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

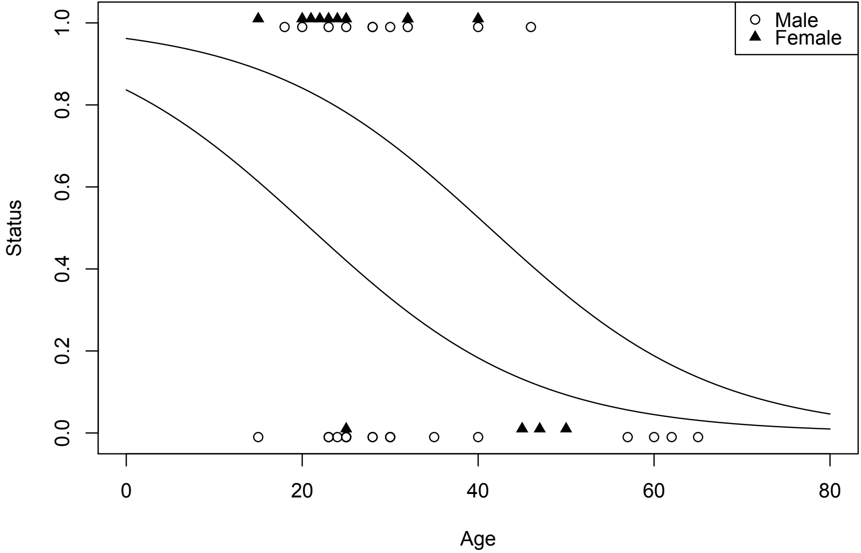
Male model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \mathit{0} \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

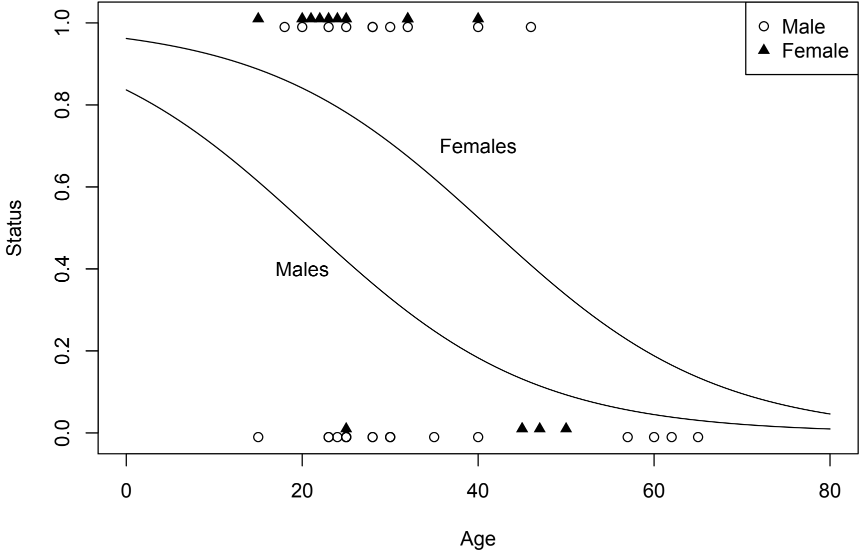
Female model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \mathit{1} \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

Example - Donner Party - Gender Models (cont.)



Example - Donner Party - Gender Models (cont.)



Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Hypothesis test for the whole model

```
summary(glm(Status ˜ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ˜ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Note that the model output does not include any F-statistic, as a general rule there are not single model hypothesis tests for GLM models.

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We are however still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z test.

Note the only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	*-0.0782*	*0.0373*	*-2.10*	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	*-0.0782*	*0.0373*	*-2.10*	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$
$$H_A : \beta_{age} \neq 0$$

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	*-0.0782*	*0.0373*	*-2.10*	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$
$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{\text{-0.0782} - 0}{0.0373} = \text{-2.10}$$

$$\begin{aligned}\text{p-value} &= P(|Z| > 2.10) = P(Z > 2.10) + P(Z < \text{-2.10}) \\ &= 2 \times 0.0178 = 0.0359\end{aligned}$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

$$\exp(CI) = (\exp -0.1513, \exp -0.0051) = (0.85960.9949)$$

1. Background
2. GLMs
3. Logistic Regression
4. Additional Example

Example - Birdkeeping and Lung Cancer

A 1972 - 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

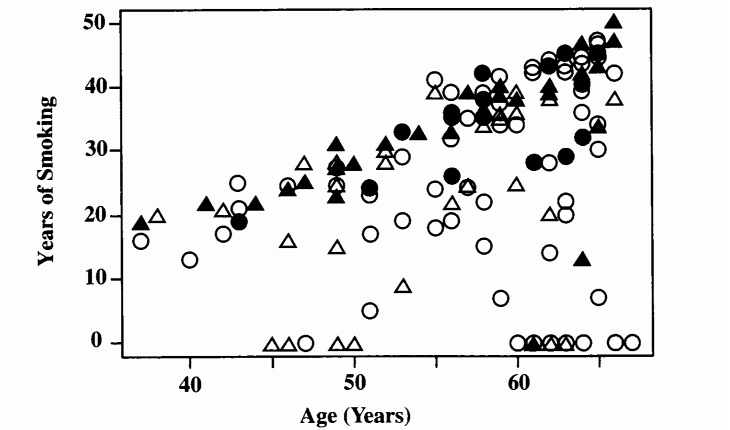
Example - Birdkeeping and Lung Cancer - Data

	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37.00	19.00	12.00
2	LungCancer	Male	Low	Bird	41.00	22.00	15.00
3	LungCancer	Male	High	NoBird	43.00	19.00	15.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	NoCancer	Female	Low	NoBird	65.00	7.00	2.00

LC	Whether subject has lung cancer
FM	Sex of subject
SS	Socioeconomic status
BK	Indicator for birdkeeping
AG	Age of subject (years)
YR	Years of smoking prior to diagnosis or examination
CD	Average rate of smoking (cigarettes per day)

Note - NoCancer is the reference response (0 or failure), LungCancer is the non-reference response (1 or success) - this matters for interpretation.

Example - Birdkeeping and Lung Cancer - EDA



	Bird	No Bird
Lung Cancer	▲	●
No Lung Cancer	△	○

Example - Birdkeeping and Lung Cancer - Model

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))

## Call:
## glm(formula = LC ~ FM + SS + BK + AG + YR + CD, family = binomial,
##     data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736    1.80425  -1.074 0.282924
## FMFemale     0.56127    0.53116   1.057 0.290653
## SSHigh       0.10545    0.46885   0.225 0.822050
## BKBird       1.36259    0.41128   3.313 0.000923 ***
## AG          -0.03976    0.03548  -1.120 0.262503
## YR           0.07287    0.02649   2.751 0.005940 **
## CD           0.02602    0.02552   1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
##
## Number of Fisher Scoring iterations: 5
```

Example - Birdkeeping and Lung Cancer - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Example - Birdkeeping and Lung Cancer - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Keeping all other predictors constant then,

Example - Birdkeeping and Lung Cancer - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.3626) = 3.91$.

Example - Birdkeeping and Lung Cancer - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.3626) = 3.91$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0729) = 1.08$.