Chapter 2
Short History of the Logistic Regression Model

Abstract The logistic regression model, as compared to the probit, Tobit, and complementary log–log models, is worth revisiting based upon the work of Cramer (http://ssrn.com/abstract=360300 or http://dx.doi.org/10.2139/ssrn.360300) and (Logit models from economics and other fields, Cambridge University Press, Cambridge, England, 2003, pp. 149–158). The ability to model the odds has made the logistic regression model a popular method of statistical analysis. The logistic regression model can be used for prospective, retrospective, or cross-sectional data while the probit, Tobit, and the complementary log–log models can only be used with prospective data because they model the probability of the event. This chapter provides a summary (http://ssrn.com/abstract=360300 or http://dx.doi.org/10.2139/ssrn.360300; Logit models from economics and other fields, Cambridge University Press, Cambridge, England, 2003, pp. 149–158).

2.1 Motivating Example

More than 175 years after the advent of the growth curve, we have fully embraced the logistic regression model as a viable tool for binary data. Today, the logistic regression model is one of the most widely used binary models in the analysis of categorical data. The logistic regression model is based on modeling the odds of an outcome, and the idea of odds (as used commonly by the average person) has lots of appeal. Many seem to be familiar with the odds of certain outcomes, whether their discussions are in sports, illness, or almost anything else. Additionally, it is quite interesting from a statistical point of view that whether the data were obtained from prospective, retrospective, or cross-sectional sampling, the covariate's impact on the binary outcome will be the same.

Since this book concentrates on fitting logistic regression models, it is reasonable to spend time elaborating on the history and the origination of those models. The advent of the logistic regression model, as compared to the probit, Tobit, log–log, and complementary log–log models, is worth revisiting (Cramer, 2002, 2003). The ability to model the odds has made it very attractive since the logistic regression relies on the odds, and the odds can always be computed whether the

J.R. Wilson, K.A. Lorenz, *Modeling Binary Correlated Responses using SAS, SPSS and R*, ICSA Book Series in Statistics 9, DOI 10.1007/978-3-319-23805-0_2

data are prospective, retrospective, or cross-sectional. However, since the probit, Tobit, log–log, and complementary log–log models rely on probabilities, they are only applicable to prospective data. Logistic regression models model the probability (nonlinear) or, equivalently, the odds (nonlinear) or logit (linear) of the outcome of an event. Logistic regression models have been used in countless ways, analyzing anything from election data to credit card data to healthcare data. Logistic regression analysis is a useful tool for all of these disciplines because it is ideal for identifying, discriminating, and profiling different types of subpopulations.

2.2 Definition and Notation

2.2.1 Notation

In this discussion, we use the following symbols:

P_t is the probability of the outcome at time t being one.
$1 - P_t$ is the probability of the outcome being zero at time t.
log is the natural logarithm.
logit denotes the log of the odds, i.e., $\left(\log[P_t/(1 - P_t)\right]$.
β_0 represents the value of the logit when the covariate is zero.
β_1 represents the increase in the logit for a unit increase in the covariate (when continuous) or the difference from one category to the next if the covariate is binary.

2.2.2 Definition

A *monotonic function* is a function which is either entirely nonincreasing or nondecreasing. A function is said to be monotonic if its rate of increase or decrease remains the same in direction. So, for $\mathrm{x} > 0$, then $\mathrm{f(x)} = \mathrm{x}^2$ is monotonic increasing since $\mathrm{f(x+2)} > f(\mathrm{x})$ for any $\mathrm{x} > 0$ but is not for all x.

A *probit* model is a type of regression for binary data on a scale that depends on the cumulative distribution function of normal distribution.

A *prospective study* is a study designed to determine the relationship between an outcome and a certain characteristic of the units involved. The researcher follows the population group over a period of time, noting when or how often the event or nonevent (e.g., lung cancer) occurs in the smokers and in the nonsmokers. Prospective studies produce an opportunity to determine probabilities for each group (event or nonevent) and as such provide the relative risk.

A *retrospective study* is a study in which the event or nonevent is unknown, and the information gathered depends on what occurred in the past. One example is

conducting a study of patients with AIDS and whether or not they had used dirty needles or other common practices.

A *case–control study* is a non-experimental research design where researchers collect information on previous cases and compare that information with a control group of persons who have not had those cases (called the control). The two groups (case and control) are matched for age, sex, and other personal data, and are then examined to determine which possible factor (e.g., cigarette smoking, watching television) may account for the increase or decrease in the case group.

A *Tobit model* is also referred to as a censored regression model. The Tobit model is best suited to cases when the response variable is either left- or right-censoring, and we are interested in the linear relationships between variables. For example, in the 1980s there was a time when the law restricted speedometer readings to at most 85 mph. So experiments involving predicting a vehicle's top-speed from a combination of horsepower and engine size, your largest speed value would be 85, regardless of how fast the vehicle was speeding. This is a perfect example of right-censoring (censoring from above) the data. The one thing we are certain about, is that those vehicles recorded as traveling at 85 mph were at least 85 mph. Introduction to SAS. UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/sas/notes2/ (accessed November 24, 2007).

2.3 Exploratory Analyses

The logistic regression model is a tool for presenting the relation between a binary response or a multinomial response and several predictors. Its use is very familiar and common in the fields of health and education, as well as with elections, credit card companies, mortgages, and other cases, where there is a need to profile the sampling unit (Fig. 2.1).

Some example questions to guide a study might be as follow:

1. How do education, ideology, race, and gender predict a vote in favor or not in favor of a US Senator?

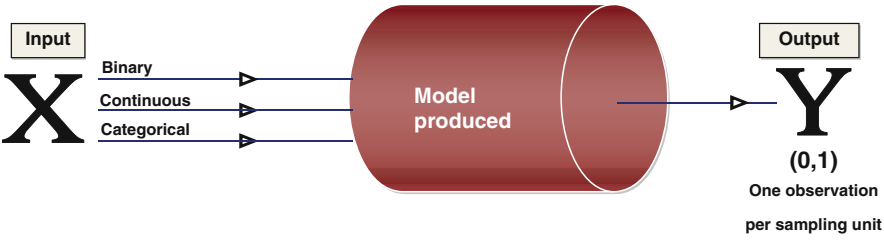

Fig. 2.1 A schematic diagram as X impacts Y

2. What factors predict the type of registered voters who would support the reelection of a President or a Governor?
3. What are the characteristics of the consumer who should be offered a credit card?
4. What are the characteristics of a traveler that will make him or her choose one mode of transportation over another (rail, bus, car, or plane)?

2.4 Statistical Model

The origin of the logistic regression model is in bioassay and some other disciplines. We learned that the logistic function was invented for the purpose of describing the population growth. Also it was given its name by a Belgian mathematician, Verhulst. Figure 2.2 provides a description of the function:

$$P_t = e^{\beta_0+\beta_1 t}/\left[1 + e^{\beta_0+\beta_1 t}\right]$$

This figure shows the relation of proportion P_t as time increases. Let the linear relation be

$$\text{logit}\,[\mathrm{P_t}] = \beta_0 + \beta_1 \mathrm{t},$$

where β_0 denotes the value at time equal to zero, β_1 denotes the rate of change of logit $[\mathrm{P_t}]$ with regard to time and

$$\text{logit}\,[\mathrm{P_t}] = \log\left[\mathrm{P_t}/\left(1 - \mathrm{P_t}\right)\right]$$

The logistic function rises monotonically as t increases. We concur with authors who have noted that for P_t from 0.3 to 0.7, the shape of the logistic curve closely resembles that of the normal probability cumulative distribution function (Fig. 2.3).

One account of the emergence of the logistic function from the growth curve is dated as far back as 1838 when it became a popular formula for certain places in North Africa (Cramer, 2002). In more recent times, Dr. Pearl of the U.S. Food Administration was preoccupied with the food needs of a growing population during World War I and decided to use logistic functions to address it. Additionally, President Dr. Lowell Reed of Johns Hopkins used an application of the logistic curve to catalytic agent formed during a reaction (Reed & Berkson, 1929). The logistic function was also used in chemistry at the same time, but it appears that the basic idea was for logistic growth. Our research support the fact that the function is still used to model population growth as well as the market penetration of new products and technologies.

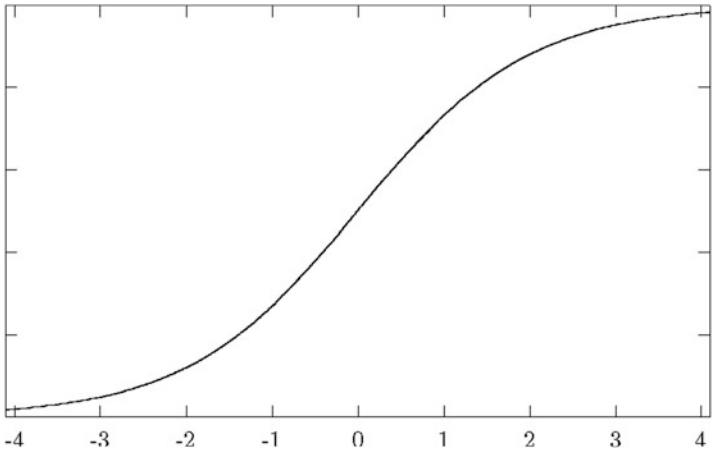

Fig. 2.2 A logistic curve P_t versus time

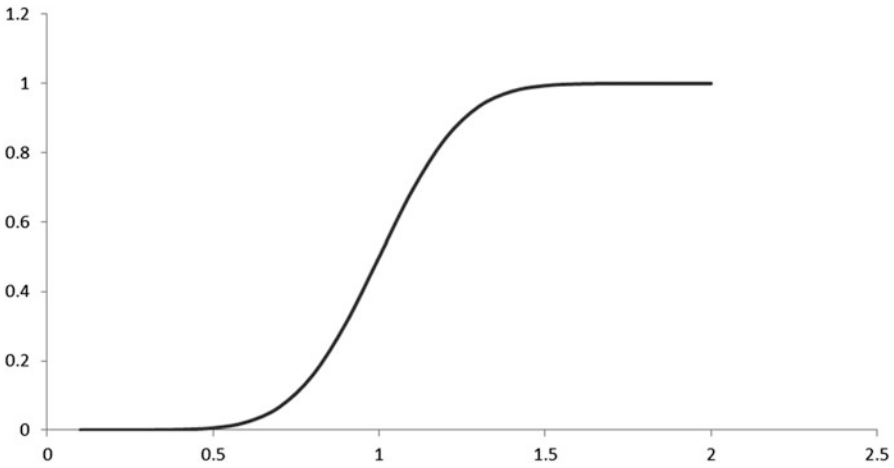

Fig. 2.3 Cumulative distribution function of normal distribution

There is a close resemblance of the logistic to the normal distribution function (Wilson, 1925; Winsor, 1932). As an alternative to the normal probability function, in 1944 Berkson turned his attention to the statistical methodology of bioassay and proposed the use of the logistic instead of the normal probability function of P_t, coining the term "logit" as compared to the "probit" presented by Bliss (1934a, 1934b). The logistic function has presented itself in bioassay in that the logit model of bioassay can easily be generalized to logistic regression, where binary outcomes are related to a number of determinants without a specific theoretical background.

We learned that the earliest developments in statistics and epidemiology took place in the late 1950s and the 1960s. We learned that in the discipline of statistics,

the analytical advantages of the logit transformation as a means of dealing with discrete binary outcomes were put at the forefront of the discussion. This was supported by Dr. Cox as a pioneer in the field by publishing a series of papers in the 1960s about the topic, and then following them up with the outstanding textbook titled *Analysis of Binary Data*, Cox (1969). Later, the close proximity of the logistic model to discriminant analysis was recognized, as well as its unique relationship to log linear models (Bishop, Fienberg, & Holland, 1975). We further learned that epidemiologists were busy developing case–control studies even earlier since the discipline of epidemiology is more directly concerned with odds, odds ratios, log-odds, or logit transformation. It appears that researchers were already clamoring about the theoretical justification, Cornfield (1951, 1956), and we must mention the works of Berkson (1944, 1951).

Our research led us to believe that the first comprehensive textbook with medical applications was published by Hosmer and Lemeshow (1989). I remember using their first edition in my graduate categorical data class in Statistics at Arizona State University shortly after I arrived in Tempe. Until recently, I was unaware that I was touching part of history. I remember back then talking to some researchers from the marketing department and being told that logistic regression was brought to their discipline by certain researchers. The presence of logistic regression models in the behavioral sciences is believed to be due to the works of McKelvey and Zavoina (Cramer, 2003). They adopted the approach based on an ordered probit analysis of the voting behavior of US Congressmen. However, the generalization of logistic regression to the multinomial or polychotomous case is due to Gurland, Lee, and Dahm (1960), Mantel (1966), and Theil (1969).

2.5 Analysis of Data

Our analyses of binary data with logistic regression models will be done mostly with SAS, SPSS, and R. There are several procedures in SAS, SPSS, and R for modeling binary responses under varying conditions and certain assumptions. We attempt to use the most common procedures as we demonstrate the fit of logistic regression models to correlated data with and without time-dependent covariates and with fixed and random effects. There are a few chapters when we were unable to duplicate the fit of the model in all three statistical packages.

2.6 Conclusions

The logistic regression is often preferred as a model for binary responses as it is appropriate for any kind of data: cross-sectional, prospective, and retrospective. Its reliance on the odds makes it an excellent candidate for interpretation as society can

easily relate to such findings. On the contrary, using probit or complementary log–log is only appropriate for modeling prospective data as they rely on probabilities.

References

Berkson, J. (1944). Applications of the logistic function to bioassay. *Journal of the American Statistical Association, 9*, 357–365.

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics, 7*(4), 327–339.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

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

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

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. *Journal of the National Cancer Institute, 11*, 1269–1275.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (pp. 135–148). Berkeley, CA: University of California Press.

Cox, D. R. (1969). *Analysis of binary data*. London: Chapman and Hall.

Cramer, J. S. (2002). *The origins of logistic regression* (Tinbergen Institute Working Paper No. 2002-119/4). Retrieved from SSRN: http://ssrn.com/abstract=360300 or http://dx.doi.org/10.2139/ssrn.360300

Cramer, J. S. (2003). The origins and development of the logit model. In J. S. Cramer (Ed.), *Logit models from economics and other fields* (pp. 149–158). Cambridge, England: Cambridge University Press.

Gurland, J., Lee, I., & Dahm, P. A. (1960). Polychotomous quantal response in biological assay. *Biometrics, 16*, 382–398.

Hosmer, D., & Lemeshow, W. (1989). *Applied logistic regression*. New York: Wiley.

Mantel, N. (1966). Models for complex contingency tables and polychotomous response curves. *Biometrics, 22*, 83–110.

Reed, L. J., & Berkson, J. (1929). The application of the logistic function to experimental data. *Journal of Physical Chemistry, 33*(5), 760–779.

Theil, H. (1969). A multinomial extension of the linear logit model. *International Economic Review, 10*(3), 251–259.

Wilson, E. B. (1925). The logistic or autocatalytic grid. *Proceedings of the National Academy of Science, 11*, 431–456.

Winsor, C. P. (1932). A comparison of certain symmetrical growth curves. *Proceeding of Washington Academy of Sciences, 22*, 73–84.

http://www.springer.com/978-3-319-23804-3

Modeling Binary Correlated Responses using SAS, SPSS and R

Wilson, J.R.; Lorenz, K.A.

2015, XXIII, 264 p. 26 illus., Hardcover
ISBN: 978-3-319-23804-3