

*Volume 3, Number 1 . March 1993*

---

*Journal of*  
***Legal***  
***Economics***

---



*American Academy*  
*of*  
*Economic and Financial Experts*

---

# Applications of Logit Analysis in Civil Litigation

## Introduction

---

**F**orensic economists, acting as expert witnesses, are frequently asked, explicitly or implicitly, to deal with questions of probability. The nature of these questions vary. Sometimes a court merely wants an estimate of a probability for some condition or event. For example, one of the authors was asked to estimate, based on sample survey information, the probability that at least 80% of the households in an allegedly senior community had at least one member aged 55 years or older. Occasionally, experts are asked to compare two sample probabilities. In a product liability case, the court may wish to know if the probability of accidental injury is greater for brand X than for brand Y, and whether the difference is large enough to be statistically meaningful. In this situation, the expert would probably perform a routine statistical hypothesis test on the ob-

served difference in the two sample proportions.

A more challenging question is to identify and quantify various factors which affect the probability of some event. In a discrimination case, for example, a plaintiff might allege that the probability of being fired or passed over for promotion is greater for female than for male employees of a given company. A defendant might counter with the argument that retention and promotion depend on employee education, years of experience, and/or test scores, but not on gender. In such a situation, an economic expert is often asked to identify those factors which tend to increase or reduce the probability of being fired or passed over. If gender, race, or national origin turn out to be among the important explanatory factors, the allegation of discrimination is greatly strengthened.

When faced with this last type of probability question, a statistical technique called logistic or "logit" regression analysis is often an ap-

\*R.R. Trout, Foster Associates, Inc., San Diego, California  
C.B. Foster, Foster Associates, Inc., San Diego, California  
G.M. McCollister, Resource Management International, Sacramento, California

appropriate method for analysis. The aim in this paper is to show how logit analysis can be used and interpreted in the courtroom. In the first sections the fundamentals of the simple logit model are presented, and its use is explained. In the second section the merits of logit analysis are illustrated by presenting a number of applications in actual cases from the authors' own experiences. In the third section an outline is offered of multinomial logit, which is an extension of logit analysis for more complex situations where the possibilities of several outcomes (rather than just two) must be measured simultaneously. The discussion is summarized in the last section.

## The Logit Regression Model

### Introduction to Regression

Most economists are familiar with a technique called *statistical regression analysis*. This is a method for determining how changes in one or more explanatory variables affect the level of some dependent variable of interest.<sup>1</sup> To illustrate, suppose one wants to determine the factors which affect a typical household's consumption spending. An economist might survey a number of households and collect data on how much each family spent on consumption in the last year, the family's total annual income, the number of children and the region of the country where the family lives. A beginning hypothesis might be that consumption depends linearly on income and the number of children, and is different

in the South as compared with the rest of the country.

The above hypothesis is expressed by the following regression equation:

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 N_i + \beta_4 S_i + u_i, \quad (1)$$

$i = 1, \dots, n$

There are  $n$  families in the survey, and for the  $i^{\text{th}}$  family,  $C_i$  is annual consumption spending,  $Y_i$  is annual income,  $N_i$  is the number of children, and  $u_i$  is the *error term* which allows for other influences on consumption which are not directly taken into account by the model. Variable  $S$  is a dummy variable which captures the effect of region:  $S_i = 1$  if the  $i^{\text{th}}$  family lives in the South, and  $= 0$  if not. In this example,  $C$  is the dependent variable [consumption], while  $Y$ ,  $N$  and  $S$  are explanatory variables.

With the data and model specified above, an economist would use a computer-based regression routine to estimate the parameters  $\beta_j$ , where  $j = 1, \dots, 4$ . Suppose the results were as follows<sup>2</sup>:

$$C = 3.0 + 0.7Y + 1.6N - 1.2S$$

(2.4) (3.7) (1.4) (-2.0)

The interpretation goes something like this. Income and consumption were both measured in thousands of dollars. For each \$1,000 increase in income ( $Y$ ), there is a \$700 increase in consumption, holding other factors constant. Each extra child ( $N$ ) causes consumption to rise by \$1,600, holding other factors constant. For given income and family size, consumption is about \$1,200 lower for southern families.

The equation can be used to predict consumption. For a family with \$30,000 of income and 2 chil-

dren living in the West, predicted consumption is  $\hat{C} = 3 + 0.7(30) + 1.6(2) - 1.2(0) = 27.2$ , or \$27,200

The numbers in parentheses below the coefficient estimates are called *t-ratios*. Very roughly, if a coefficient has a t-ratio larger than 2 (in absolute value), then the corresponding variable is thought to be a statistically significant factor in explaining the dependent variable.<sup>3</sup> The example above suggests that income and region are significant in determining consumption behavior, but the number of children is not. The specification, estimation, and interpretation of econometric regression models is a mix of science and common sense and is a good deal more complicated than the preceding discussion would imply. But the general idea is fairly straightforward.<sup>4</sup>

### Regression and Probability

The methods of regression analysis can be extended to situations where the phenomenon to be explained is the probability that some event will occur. In such cases, the dependent variable may be an indicator (i.e., dummy) variable, similar to the explanatory variable  $S_1$  in equation 1, above. An example will explain this idea.

In an age discrimination case, the court is interested in whether the probability of being hired is lower for older job applicants than for younger applicants, other factors considered. From a sample of  $n$  applicants for a job opening, the economist records the applicant's age ( $A_i$ ), prior years of relevant experience ( $X_i$ ), and years of education ( $E_i$ ), where  $i = 1, \dots, n$ . Data on the dependent variable  $P$  is also re-

corded  $P_i = 1$  if the  $i^{\text{th}}$  applicant was hired, and  $P_i = 0$  if not. A simple linear regression model would use the equation below:

$$P_i = \beta_1 + \beta_2 A_i + \beta_3 X_i + \beta_4 E_i + u_i, \quad i = 1, \dots, n \quad (2)$$

Equation 2 is an example of the so-called linear probability model (LPM). After obtaining coefficient estimates  $[\hat{\beta}_j, \text{ for } j=1, \dots, 4]$  the predicted value  $\hat{P}_k = \hat{\beta}_1 + \hat{\beta}_2 A_k + \hat{\beta}_3 X_k + \hat{\beta}_4 E_k$  is interpreted as the estimated probability of being hired for an individual with age, experience and education  $A_k, X_k, E_k$ . If  $\hat{\beta}_2 < 0$ , the probability of being hired is lower if age is higher, other things the same, and age discrimination is a likely possibility. The likelihood of this discrimination can be measured by referring to the t-ratio, as previously discussed. T-ratios with absolute values in excess of 2.0 are deemed to be statistically significant.

### Logit Regression Analysis

There are a number of flaws with the linear probability model shown in equation 2. One of these is that, for given values of  $A, X$  and  $E$ , the estimated probability  $\hat{P}$  may be greater than 1 or less than 0. Since true probabilities must lie between 0 and 1, the linear probability model can provide inaccurate estimates and should not be relied upon. The principal advantage of logit regression is that the final results are guaranteed to yield well-behaved estimates of the specific probability the researcher is interested in:  $0 < \hat{P} < 1$ , always.

The fact that estimated probabilities may not lie between 0 and 1 is the most serious flaw in the

linear probability model, but it is not the only one. Ordinary least squares (OLS) provides good estimates of regression equation parameters only when several assumptions hold true, and the LPM violates two of these assumptions. In the LPM, the variance of the error term is not constant (the problem of *heteroscedasticity*), and the error term has a discrete binary distribution (the problem of *non-normality*). The first problem can be overcome by using weighted least squares, which is cumbersome, the second problem makes conventional tests of statistical significance using t-ratios unreliable.

These statistical problems with the LPM can be avoided by replacing it with one of several models that transform the binary dependent variable so that the predicted probabilities all lie within the zero to one range. The simplest of these transformations is called the logit model. Logit analysis is based on the cumulative logistic probability function.

To illustrate the theory behind logit analysis, consider the age discrimination example again. Logit analysis assumes that the *natural logarithm of the odds* of the  $i^{\text{th}}$  applicant being hired depends linearly on various explanatory factors. In this example, if  $P_i$  is the true probability of being hired, then:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \alpha_1 + \alpha_2 A_i + \alpha_3 X_i + \alpha_4 E_i + \text{error term} \quad (3)$$

Denote  $\ln(P/1-P)$  by  $L$ . Logit regression programs estimate the regression parameters (the  $\alpha$ 's in equation 3), for the dependent variable  $L$ , and therefore estimate  $\hat{L}$ . From any estimate  $\hat{L}_k$ , the estimate

of the underlying probability can be derived<sup>5</sup>:

$$\hat{P}_k = \frac{e^{\hat{L}_k}}{1 + e^{\hat{L}_k}} \quad (4)$$

A little algebra shows that  $0 < \hat{P}_k < 1$ .

The coefficients for the logit model are usually estimated with a maximum likelihood estimation (MLE) routine. The resulting estimates will be unbiased and efficient and will not have the problems associated with the LPM. To perform logit regression analysis, a special computer program is usually required,<sup>6</sup> although the recording and entry of data are the same as for the linear probability or any other regression model.<sup>7</sup>

In the usual case where logit regression analysis is used, the raw data for the dependant variable are recorded as  $P_i=1$  if the event of interest occurred, and  $P_i=0$  if not. In this raw form, the log of the odds appears to be  $\ln(0/1)$  or  $\ln(1/0)$ , both of which are undefined. Regression programs with logit subroutines overcome this difficulty by creating linear indexes of the explanatory variables and working with cumulative probability distribution functions. It is for this reason that it is noted above that logit analysis *usually* requires a special computer program.

There is one situation where logit analysis can be performed directly with an OLS program. In this situation, the original data may have been recorded for individual persons, families, households, etc., but are available to the researcher only as grouped data.<sup>8</sup> For example, one might have the number of male and female applicants for a job who

were in the 20-24 year age group, the 25-29 year group, the 30-34 year group, and so on. And for each group and subgroup, one might know the number of those applicants who were hired. The number hired divided by the total number in subgroup  $i$  is a proportion  $P_j$  where  $0 \leq P_j \leq 1$ . The *logit* of this proportion is easily calculated as  $\text{Ln}(P_j/1-P_j)$ , as long as the groups are chosen so that  $P_j \neq 0$  or  $1$ .

The *logit* for each of the groups is then regressed on the mean value of the explanatory variables in each of the groups to obtain the regression parameters (the  $\alpha$ 's of equation 3). This analysis can be easily performed using an OLS model available in any statistical package. The estimated regression parameters using either the *logit* in an OLS model, or the binary dependent variable in a logistic regression model, will be unbiased estimates of the true values of the regression parameters.<sup>9</sup> Using the OLS routine is quicker, and less costly from a computer cost standpoint but requires that the data be in a grouped format. One could, of course, always assign individual data (e.g., household responses) to groups and use OLS in place of *logit* analysis. However, in doing this the researcher would eliminate a significant amount of the potential explanatory power of the other variables. Given the availability of various MLE alternatives, using the OLS routine on grouped data is a less satisfactory choice when there is an option of using individual observations.

In most courtroom applications, the questions of interest are answered when equation 3 has been estimated. As with the linear prob-

ability model, if  $\hat{\alpha}_2 < 0$ , with a statistically significant *t*-ratio, and with a decent overall model fit, then age discrimination may be inferred.

## Applications

---

In the sections that follow are presented a few interesting applications of *logit* analysis to litigation problems. The cases discussed include a statistical analysis of medical hazards, an employment discrimination case, and the development of a model to determine the loss period in wrongful termination cases. These cases will further illustrate the superiority of the *logit* model over the linear probability model referred to in the first section.

### Case #1 - Medical Hazards

Dr. X was one of four surgeons at a local community hospital. The hospital administration observed that increasingly erratic behavior by Dr. X was affecting the morale of his colleagues and the hospital staff. When a study revealed that Dr. X's patients were much more likely to suffer postoperative complications of a certain type than were the patients of the other three surgeons, the hospital fired him. Dr. X promptly sued, alleging wrongful termination. A major issue in the case was whether the hospital's study of patient complications was valid. The authors were asked to determine if the hospital was justified in concluding from the study that Dr. X was jeopardizing the health of his patients.

*Logit* analysis did not appear necessary at the outset. Of 197 patients whose records were examined in the study, 84 were operated

on by Dr. X, and the remaining 113 patients were operated on by one of the other three surgeons. About 8% of Dr. X's patients suffered a particular complication after surgery, but the proportion was less than 1% for the other patients. A routine statistical hypothesis test was conducted for the difference in population proportions. The null hypothesis was  $H_0: \rho_X \leq \rho_3$ , where  $\rho_X$  is the true (unknown) proportion of Dr. X's patients who develop the complication, and  $\rho_3$  is the same proportion among the patients of the other three surgeons.  $H_0$  was rejected at the 0.005 (0.5%) level. The conclusion was that Dr. X's patients faced significantly greater hazards after surgery than the patients of the other surgeons.<sup>10</sup>

There was a potential complication, however. The 197 patient cases in the study were not a random sample of patients, but a census of everyone operated on by any of the four surgeons in recent years. If there were consistent differences in the patients assigned to, or selecting, Dr. X, then there might be reasons for different postoperative complication rates that had nothing to do with the intrinsic qualifications of Dr. X *per se*. For example, if Dr. X operated on sicker patients, or dealt with conditions requiring more delicate surgical procedures, his case load might be expected to exhibit higher postoperative complication rates.

In order to examine these possibilities, it was surmised that the probability of postoperative complication could be related to the health of the patient, and the type of surgery, as well as the identity or competence of the surgeon. The

hospital study data base included one measure that could be interpreted as an index of patient health at the time of admission to the hospital. The hospital staff identified a particular procedure – type I – that might be particularly hazardous. Including these data allowed the construction of the following logit model:

$$\ln\left(\frac{P}{1-P}\right) = \alpha_0 + \alpha_1 HEALTH + \alpha_2 TYPE + \alpha_3 X + \text{error term} \quad (5)$$

In this model,  $P$  is the probability that patient suffers a particular complication after surgery:  $P=1$  if the complication develops, 0 if not. Among the explanatory variables,  $HEALTH$  is an index of patient health (severity of condition) prior to surgery, where  $0\% < HEALTH < 100\%$ ;  $TYPE = 1$  if a type I surgical procedure was employed, and 0 if not;  $X = 1$  if Dr. X performed the surgery, and 0 if not.

The following regression results were obtained:<sup>11</sup>

$$\ln\left(\frac{\hat{P}}{1-\hat{P}}\right) = -4.217 - 0.008HEALTH + 0.2696TYPE + 2.108X$$

P-Values (0.001) (0.253) (0.376) (0.027)  
Latent  $R^2$ : 0.283

Like the t-ratios described in part I, the p-values are used to test if the coefficient estimates are significantly different from zero and that the corresponding variable is a significant factor in determining the probability of complication.<sup>12</sup> A statistically significant p-value is 0.05 or less, which is roughly consistent with a t-ratio of 2.0 or greater.

The results above suggest that neither  $HEALTH$  nor  $TYPE$  has much influence on the probability of postoperative complication. The

high p-values for  $\hat{\alpha}_1$  and  $\hat{\alpha}_2$  imply that the true coefficients may well be zero. However, the statistically significant positive sign for  $\hat{\alpha}_3$  indicates that even after the possible effects of patient health and surgical procedure are taken into account, the odds of postoperative complications are higher if Dr. X was the surgeon.

Using the conversion shown in equation 4, the probability of a postoperative complication can be computed for any value of HEALTH.

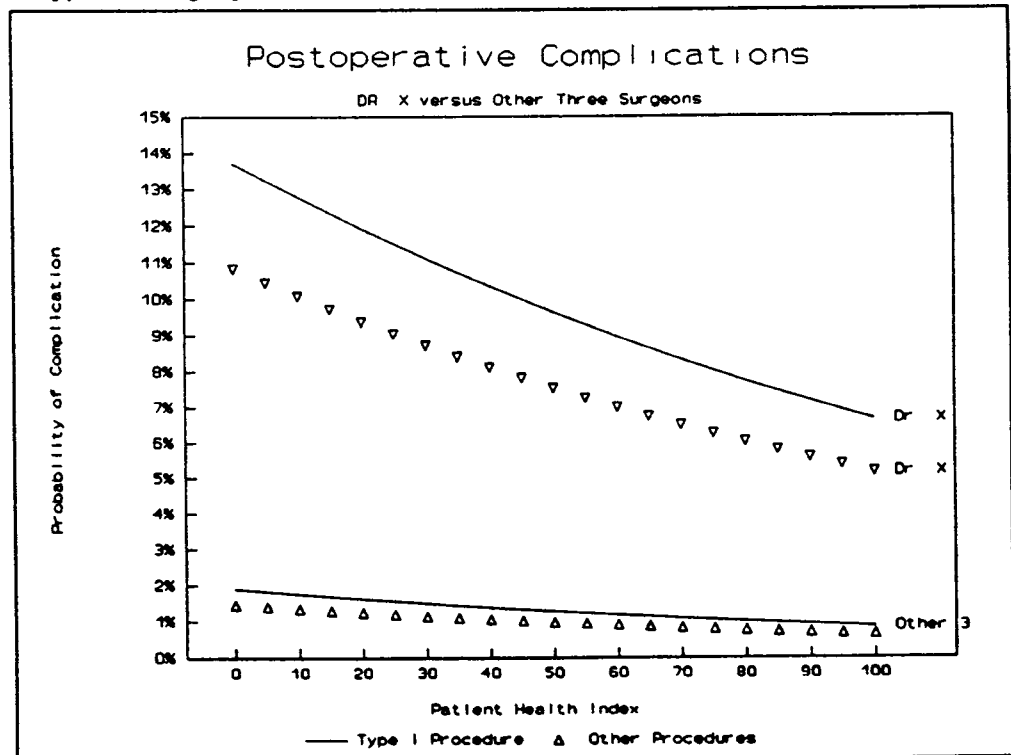
These probabilities are graphed in figure 1. Regardless of the type of surgery and level of the patient health index, the chance of a complication is notably higher if Dr. X is the operating physician. The different probabilities are calculated in table I for patients of average health.

The overall conclusions were that (1) Dr. X's patients were exposed to greater health hazards than those of the other surgeons with the same specialty in the hospital, and (2) the difference in hazard was not due to differences in the health of the patients or the types of surgery required. The inference was that the hazard differential *was* due to some intrinsic characteristic of Dr. X's skill or behavior. In other words, his termination by the hospital was *not* wrongful.

	Dr. X	Other 3
Type I	8.69%	1.14%
Other Type	6.78%	0.88%
(For HEALTH = 63.52%)		

It is interesting to compare the LPM with the logistic regression

Figure 1 - Probability of postoperative complication by surgeon and type of surgery



model using the results of this case. The coefficients in the LPM indicate the change in the probability of postoperative complications for a one unit change in the corresponding explanatory variable. When the linear probability model was run on the Dr. X data, the coefficient for physician identity was 0.072, as shown below.

$$P = 0.032 - 0.0004\text{HEALTH} + 0.0084\text{TYPE} + 0.072X$$

P-Values. (0.48) (0.80) (0.026)

$$R^2 = 0.021$$

This indicates that the probability of a complication was 7 percentage points higher if Dr. X performed the operation, regardless of the type of surgery or level of patient health.

The coefficients for the logistic model indicate the change in the logit [i.e., log of the odds] associated with a one unit change in an explanatory variable. The rate of change of probability with respect to the  $j$ th variable is  $\hat{\alpha}_j [\hat{P} (1-\hat{P})]$ , where  $\hat{P}$  is evaluated for selected values of all the explanatory variables. Thus, the incremental probability of a postoperative complication when surgery is performed by Dr. X depends on both patient health and the type of surgery. This can readily be seen by comparing the vertical distance between a *Dr. X* probability curve and a corresponding *other 3* curve in figure 1. The distance shrinks as the patient health index increases and is different depending on whether a type I surgical procedure was employed or not.

When the data for an explanatory variable are in binary form (such as with the physician indicator variable), the coefficient estimated by logit regression has an

interesting interpretation. For coefficient  $\alpha$ ,  $e^\alpha$  is the odds ratio.<sup>13</sup> For the physician indicator variable, the logit coefficient estimate was 2.108, and the odds ratio is therefore 8.23 [i.e.,  $e^{2.108} = 8.23$ ]. This means that when  $x=1$ , the odds of postoperative complications are *eight times higher* than when  $x=0$  [i.e., someone other than Dr. X performed the operation].<sup>14</sup> This can be verified by using the probabilities displayed in table I. Assuming a type I procedure, the odds of complication are found as follows:

$$\text{odds} = \frac{\hat{P}}{1-\hat{P}}$$

For Dr. X

$$\text{odds} = \frac{0.0869}{(1-0.0869)} = 0.0952$$

$$\text{reciprocal odds} = \frac{1}{0.0952} = 10.5:1$$

For other physicians

$$\text{odds} = \frac{0.0114}{(1-0.0114)} = 0.0115$$

$$\text{reciprocal odds} = \frac{1}{0.0115} = 86.7:1$$

$$\text{The odds ratio} = \frac{86.7}{10.5} = 8.257^{15}$$

That is, the odds against developing a complication are almost 87 to 1 if another surgeon does the procedure but only a little better than 10 to 1 if Dr. X does it.

## Case #2 - Employment Discrimination

Regression techniques have been used in discrimination cases for some time, usually where the dependent variable is wage level, and the issue is whether the sex or race of an employee affects wage rates.<sup>16</sup> A natural extension of this type of analysis using logistic re-

gression arises when the dependent variable represents whether an individual was retained or laid off, and one wishes to determine if an employee's sex, race or age affected the employer's decision to terminate any specific individual. The same type of model can be used to examine a pool of workers who were either promoted or not promoted over some specific time period, and whether age, race or sex affected the promotions.

One of the authors had a particular case where a manufacturing firm terminated ten employees, eight of whom were over the age of 40. The two-by-two results shown in table II below indicate that the termination rates for the over and under 40 age groups were 22% and 5%, respectively. The difference in the termination rates is 0.17, with a t-test value for the difference in means of 2.12. A Fisher exact test probability was 0.04, while a Mann-Whitney test was significant at the 0.02 level. Both of these simple statistical tests indicated a likelihood of discrimination.<sup>17</sup>

	Age Group		
	Under 40	Over 40	Total
Retained	35	28	63
Terminated	2	8	10
Total	37	36	73
Termination Rate	0.05	0.22	

The two-by-two table offers an interesting way to compare the linear probability model and the logit model. The results of a LPM regression using the data summarized in table II are as follows:

$$\text{TERM} = 0.54 + 0.168\text{AGE}$$

P-Value (2.12)

The dependent variable, TERM, equals one if an employee was terminated, zero if not. The coefficient for the AGE variable represents the difference in termination rates. TERM increases in value (probability) as AGE increases. The t-statistic is the same for the LPM as it is for the test on the difference in group means. The LPM measures the difference in observed rates for the two groups, and also the likelihood that the observed coefficient (i.e., difference) is statistically significant (i.e., not equal to zero).

The two-by-two table can also be analyzed in a variety of other ways.<sup>18</sup> For example, the odds that an individual will be terminated, given his or her group membership, can be computed directly from the data in table II. Remember that the logit transformation is simply the log of the odds ratio. For workers under 40 years of age, the odds of termination are 17.5:1 against being terminated.<sup>19</sup> For those above the age of 40, the odds of termination are 3.5:1.

Finally, a useful comparison of the two ratios is called the odds ratio, which is simply 17.5/3.5, or 5.0. This indicates that the odds of being terminated if one is above the age of 40 are five times greater than if one is below the age of 40.<sup>20</sup> While the odds ratio has no particular distribution, the log of the odds ratio is normally distributed for large samples. This leads back to the logit model, which asserts that the log of the odds ratio is linearly related to a group of explanatory variables. The logistic regression equation using the data in table II is

$$\text{Logit (TERM)} = -2.86 + 1.609\text{AGE}$$

(1.94)

P-Value 0.05

Note that  $e$  raised to the power of 1.609 equals 5.0, which is the odds ratio, and obviously therefore the log of the odds ratio is 1.609. The logit model measures both the size and the statistical significance of the relative odds of some event, given one or more factors that affect the event. In contrast, the LPM measures the size and statistical significance of the difference in probabilities of some event, given one or more factors that affect the event.<sup>21</sup>

The analyses presented thus far indicate a statistical likelihood of discrimination. However, it was possible that the rather simple two-way analysis could omit important factors that would account for the higher termination rate among those over 40. To examine this potential problem, logistic multiple regression was used to determine if variables other than age could account for the disparity in termination rates.

The results of the logit multiple regression are shown below, where AGE is the age of the employee at the time of the layoffs (rather than an indicator variable as used above), SERVICE is the employee's years of service with the company at that point, and DEGREE measured the highest level of college education reached.

$$\text{Ln} \left[ \frac{P}{1-P} \right] = -9.0 + 0.22\text{AGE} - 0.27\text{SERVICE} - 1.4\text{DEGREE}$$

(2.4)    (-2.3)    (-1.2)

Even with prior service and education accounted for, the probability of termination was significantly higher for older employees, since

the coefficient of AGE, 0.22, was positive and its t-ratio was 2.4 with the other variables in the model. Since AGE is no longer a binary variable, one cannot interpret  $e^\beta$  as a measure of the odds ratio.

Figure 2 shows that, for average values of SERVICE and DEGREE, the probability of being terminated was rising sharply for employees past their mid-forties, thus confirming the earlier analysis using two-by-two contingency table results.

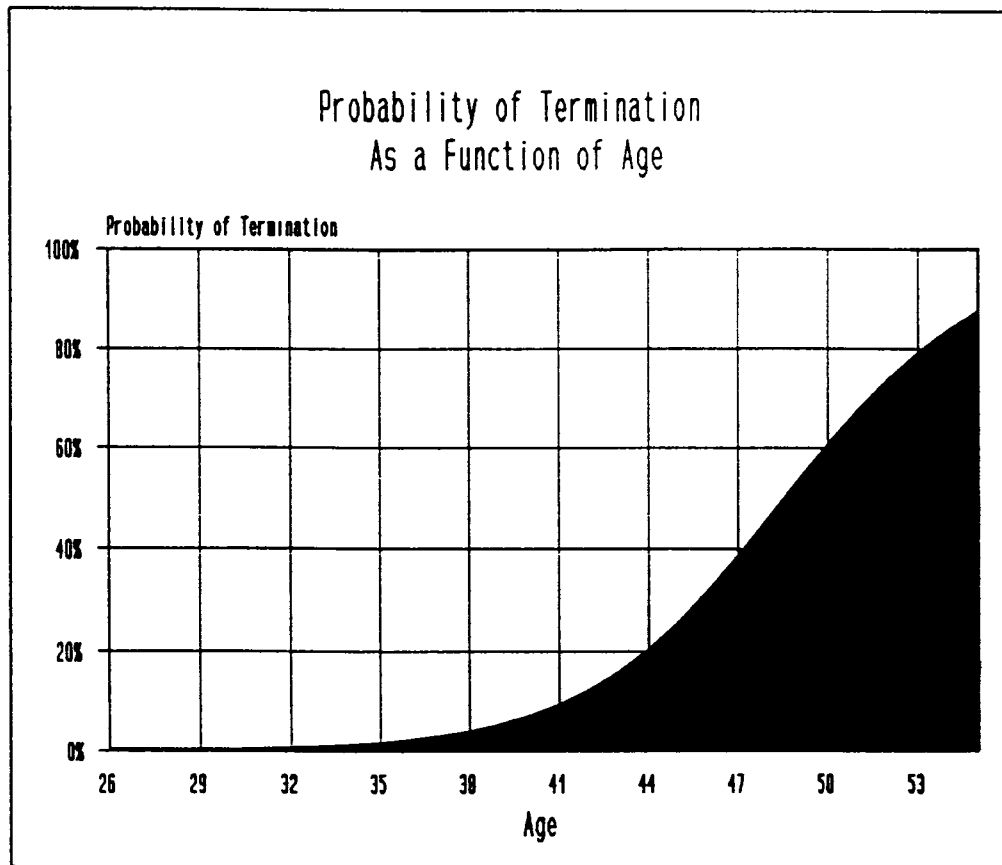
### Case #3 - Conditional Probability of Employment

In personal injury cases, the economist must determine the economic losses to an individual resulting from an accident or injury. The claim is against some defendant that allegedly caused a temporary or permanent interruption to the earnings capacity of the injured party.<sup>22</sup> In permanent injury cases, losses extend throughout the plaintiff's expected work life. The injured party usually had a job at the time of the injury. There is no need for the plaintiff to assert or show that he or she would have remained with that same employer for any given period of time if the injuries had not been suffered.

In employment cases such as wrongful terminations, the relationship between the plaintiff and the employer is quite different. For one thing, the employer is usually the defendant. For another, it is now materially relevant to determine how long the plaintiff employee would have remained with the defendant employer if the wrongful firing had never happened.

Basically, annual economic loss to the individual is the compen-

Figure 2



sation from the original employer, including job-related benefits, less any actual earnings and benefits. This annual loss is ordinarily divided into two components: (1) past economic losses, and (2) present value of future economic losses. Important unknown factors that must be considered in computing economic losses are the duration of unemployment if the terminated employee has not yet found employment, and the likely continuation of employment by the defendant company, had there been no termination. In some cases there may be a contract between employee and employer that specifies some length of time for the employment relationship, but this is rare. In the absence of such a contract, the plaintiff may claim he or she would

have worked for the defendant employer until retirement. While it would be nice (for the plaintiff) to be able to make this assertion with some belief, evidence suggests that, on average, it is not very likely that any particular individual will remain working for the same employer very long. This is in sharp contrast to the way forensic economists often present economic loss estimates on behalf of wrongfully terminated plaintiffs. Often the assumption is made that the plaintiff would have worked for the defendant employer until retirement. Alternatively, economists sometimes select an arbitrary number of years into the future, say five years, to use in determining future damages.

The reality of employment is that people change jobs in this

country at an alarming rate. Current Population Survey (CPS) data indicate historical annual turnover rates of about 10%. About 88% of the turnovers are voluntary, with the remaining 12% being involuntary. The most common reason given in the survey for a change of employer is better pay or different working conditions.

The Bureau of Labor Statistics (BLS) has published studies of occupational tenure, and regularly publishes information about unemployment duration. However, neither the BLS nor any other labor researchers have examined the likelihood that an individual will remain in a specific job in the future. Knowledge of this likelihood is exactly what is needed to estimate accurately the economic losses in an employment litigation case.

Since the BLS does not directly publish data on job retention probabilities, obtaining the probabilities proved to be difficult. First, the CPS data were obtained from surveys, conducted monthly by the U. S. Department of Commerce, Bureau of the Census. In January 1987 the CPS included in its survey a supplement that contained a series of questions about employment in the prior year, employment changes during the year, and current status of employment. From the series of responses to these questions it was possible to create a data set consisting of more than 100,000 individual or household observations on income, education, age, and job related information. This sample population was reduced by eliminating those individuals not in the labor force and those unemployed at the beginning of the period. The

final population consisted of individuals employed in January 1986 and either employed or unemployed [but in the labor force] in January 1987.

This final data set was then used to analyze the probability of an individual being employed by the same employer one year in the future. A logistic regression model was used to determine the probability that an individual with specific economic and demographic characteristics would be working for the same employer one year in the future. The dependent variable for our analyses is a binary indicator variable which takes on a value of one if the individual was employed both in 1986 and 1987, and a value of zero if the individual was employed in 1986, but not in 1987. The explanatory variables examined included the following:

- Age of individual
- Education of individual (in years)
- Family income class (13 classes)
- Duration of employment with current employer (in years)
- Three indicator variables representing sex and race
  - White female
  - Non-white male
  - Non-white female
- Interaction variables
  - Education X Income
  - Age X Income
  - White male X Income
  - Duration X age
- Non-linear variables
  - Age squared
  - Income squared

Variable	Coefficient	t-Statistic	Rate of Change at Mean
Constant	-2.030000	17.5	
Age	0.142300	28.2	0.011770
Age Sqd	-0.001600	26.6	-0.000130
Income	0.047950	24.2	0.003972
Income Sqd	-0.000350	17.4	-0.000029
Education	0.047950	8.6	0.000464
Duration X Age	-0.000064	1.7	-0.000005

The results of the logistic regression analysis, including the rates of change associated with each variable, evaluated at the mean value of the dependent variable, are presented in table III. Both the race and sex indicator variables turned out to be statistically insignificant and were dropped from the equation. This was also true of all the interaction variables, with the exception of duration x age, which was found to be statistically significant.

All of the variables in the final logistic model were significant at the 5% level, using a one-tailed test, with the correct signs. The coefficient of determination for the model was 0.1, which is reasonably good for a binary dependent variable having a mean value of 0.9. The rates of change for each variable are also shown. The rate of change indicates the change in the annual probability that results from a one unit change in the independent variable, evaluated for an individual with a probability of 0.9 [i.e., the mean value for the dependent variable].

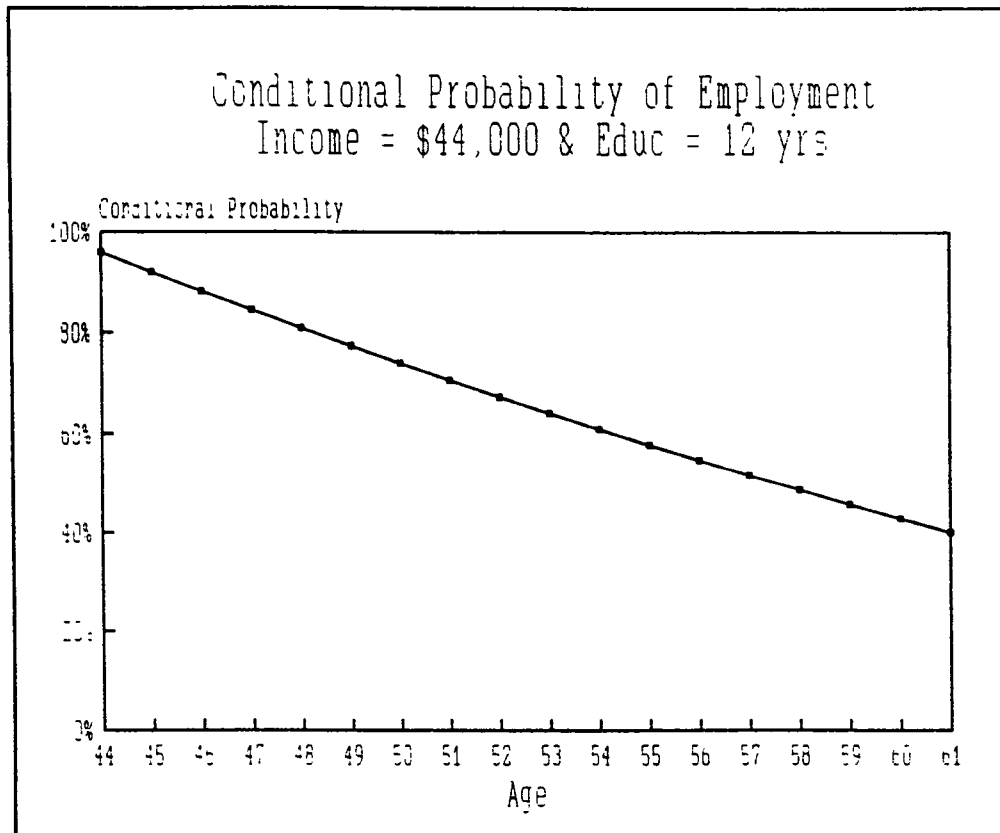
The model was then employed in a wrongful termination case in the computation of past and future economic loss. In this particular case, the plaintiff was 44 years of

age at the time of the termination, had 9 years of employment with the defendant, and earned \$46,000 per year in income, including job-related benefits. The probability of being employed at the end of any particular year is a function of the individual's age, income, education, and duration of prior employment as of that year. The conditional probability of being employed each year in the future is the probability of being employed at the end of the current year, given employment at the beginning of the year, multiplied by the probability of being employed at the beginning of the year. For the individual in this case, both the annual probability of employment and the conditional probability of employment each year is shown in figure 3.<sup>23</sup>

### **Multinomial Logit**

Up to this point, the discussion has focused on probabilities of a single event: a job applicant gets hired; a patient develops a given complication; an employee is still with the same employer a year later. But it is easy to imagine situations where probabilities of a set of events may be of interest. Suppose one wishes to determine whether

**Figure 3**



members of a minority group receive equal treatment in some firm, industry, or occupation. The relevant treatment *events* might be (1) employee is promoted, (2) employee is retained at same level, (3) employee is fired. One would like to see whether the probabilities of these three events, taken jointly, are affected by the employee's minority status. The logit model can be expanded to handle such situations. When applied to three or more discrete choices, the model is called *multinomial logit*. (The models discussed earlier are, by inference, *binomial* or *binary* logit.)

In the case of  $N \geq 3$  distinct events, the probability of the  $j^{\text{th}}$  event is

$$P_j = \frac{e^{\beta x_j}}{e^{\beta x_1} + e^{\beta x_2} + \dots + e^{\beta x_n}} \quad (6)$$

where  $j = 1, \dots, N$ ,  $\beta$  is a vector of regression parameters, and  $x_j$  is a vector of one by  $k$  explanatory variables associated with alternative  $j$ .

Multinomial logit analysis is an order of magnitude more difficult than binary logit, and requires a specialized computer program. The multinomial logit model was principally developed by McFadden (1973) and Theil (1969). Useful econometrics texts which cover this topic include Pindyck and Rubinfeld (1991), and Maddala (1983, chapter 3).<sup>24</sup>

#### Case #4 - Household Space Heating

Models to explain and predict a household's choice of space heating fuel and efficiency have been the subject of testimony in several state utility regulatory hearings, which

were concerned with setting energy prices and developing energy conservation policies. The operating cost is the primary factor that consumers consider in choosing space heating equipment. The cost is determined by the cost of the energy for each type of equipment, the efficiency of the equipment in using energy, and the winter climate, which can be summarized by heating degree days. In colder climates, operating costs should become even more important in determining the probabilities for each alternative, favoring those using the least expensive type of energy. Less important considerations are purchase and installation costs for each alternative, and preferences for certain fuels and equipment types. These other factors are more likely to prevail in warmer climates and in multifamily housing where there is less exposed outside wall area.

For a utility client, a model was developed to predict the type of space heating equipment that households will install. The utility had administered a survey to several thousand of its customers measuring their appliance stocks and demographic characteristics. Several space heating alternatives were identified:

1. Natural gas heater,
2. Propane heater,
3. Central electric resistance heater,
4. Non-central electric resistance heater, and
5. Heatpump.

Using the explanatory variables mentioned previously, a model was developed to estimate the prob-

abilities that a household will choose each alternative. For each household, the operating cost for each alternative was computed as a function of the energy price for the alternative, a 30-year average of heating degree days for the home's location, the dwelling type of the home, and the efficiency of the equipment. This cost was statistically significant (t-statistic 12.6). Additional explanatory variables represented preferences for electric and centralized systems and propane use in mobile homes. These were also highly significant. For example, a variable indicating a mobile home type was included in the set of explanatory variables for alternative 2 and revealed that propane was more likely to be used for heating this type of dwelling.

Because natural gas was the least expensive fuel, the operating cost variable always favored natural gas heating in this study, especially in the colder parts of the utility's service area. But this was offset by preferences for electric heating, and for propane gas in mobile homes. Since the set of operating costs depends on so many factors, and ranged widely over the sample, a significant portion of the sample favored each alternative.

The model was used to forecast penetration rates for each alternative based on projections of fuel prices and of home construction rates in different parts of the utility's service area.

Because it can quantify the influence of multiple factors on a set of alternatives, the multinomial logit model will find many applications in civil litigation and regulatory hearings.

## Summary

---

An attempt has been made in this paper to show how logistic regression (logit analysis) can be a valuable aid to the economist performing forensic consulting assignments. As shown by the cases used for illustration above, many issues in litigation can be analyzed as problems involving probabilities. Whenever the factors which affect probabilities need to be identified and measured with some degree of accuracy, logit analysis is frequently the technique of choice.

Logit analysis is, at bottom, merely a specialized statistical regression technique. Anyone who can interpret an ordinary least squares printout can use logit methods as well with only a little extra effort. Logit is perhaps less well known than more conventional sta-

tistical procedures because it is non-linear, and because many common computer statistical packages do not contain logit subroutines. It may be fair to say that whenever a forensic economist has employed Chi-square tests or contingency tables, he or she might have been better advised to recast the problem as a logit regression equation.

Thus, in choosing methods for analyzing probability questions, the forensic statistical practitioner should remember that logit techniques have the advantage of explanatory power and relative ease of use. They have the further advantage, of no small consequence when presenting findings to a jury, that the results are often amenable to striking graphic presentation. A picture is worth  $10^n$  words, and logit gives the picture!

## Endnotes

1. Its use in litigation has been previously discussed by Fisher (1980), and Finkelstein (1980).
2. The most commonly used regression routine is ordinary least squares (OLS). The results would also include summary statistics like  $R^2$  and other measures of "goodness of fit" which could be used to determine if the chosen model is adequate to explain the one phenomenon is interested in.
3. The Supreme Court in *Castaneda v Partida* 430 U.S. 482 and *Hazelwood v U.S.* 433 U.S. 299, decided a finding was statistically significant if it was two or three standard deviations from the value assumed by the null hypothesis. For the sample regression results under consideration here, a null hypothesis might be that  $\beta_3=0$ , which would imply that the number of children (N) does not matter. The regression estimated  $\hat{\beta}_3=1.6$ , and the corresponding t-ratio of 1.4 tells us that a coefficient value of 1.6 is 1.4 standard deviations away from 0. Since  $1.4 < 2$ , the estimated coefficient is determined to be not significant. For a discussion of the Court's use of this statistical test, see Meier, Sacks and Zabell (1986).
4. There are a host of good econometrics textbooks that cover these topics Two that also cover logit regression are Pindyck and Rubinfeld (1991), and Ramanathan (1989). See also, McFadden (1973), for an early application of logit analysis.
5. Note that  $e = 2.71828 \dots$  is the base of the natural logarithm.
6. A number of statistical software packages include logistic regression routines. The list of such packages includes BMDP, SPSS, SAS, AQD, RATS, CRUNCH, and others
7. An analogous model, called probit analysis, asserts that a transformation of the probability,  $P_i$ , to a cumulative normal value is linearly related to the explanatory variables. Probit and logit usually provide similar results. The use of regression methods, including logit and probit, to estimate probabilities is covered in Pindyck and Rubinfeld (1991). The interested reader may also wish to consult Maddala (1983), Hosmer and Lemeshow (1989), or Cox (1970).
8. See the examples presented in Cox (1970).
9. Performing logit analysis using grouped data has the same problem with heteroscedasticity that is present with the LPM. See Maddala (1983, Ch. 2).
10. Like President Garfield's physicians, of whom the President's assassin said (before his hanging), "I only shot the President, it was his doctors who killed him."
11. The model was estimated on a COMPAQ II personal computer using the AQD logit regression module. Of the original 197 cases in the hospital's

study, 19 failed to report a value for HEALTH, so only 178 observations were used in the final regression.

- 12 For example, the estimate of the coefficient of X turned out to be  $\hat{\alpha}_3 = 2.108$ . If the true value of  $\alpha_3 = 0$ , the probability of obtaining an estimate as large as 2.108 in absolute value by accident is only 0.027 (2.7%). Low P-values are associated with high levels of significance. It should also be mentioned that the overall measuring of goodness of fit for logit regressions, such as the "Latent  $R^2$ ", are different from similar  $R^2$  measures reported in more conventional regression routines. Seemingly low values of a latent or pseudo  $R^2$  may indicate a rather good overall fit. See Maddala (1983, Ch.2), for some of the conjecture on this issue.
- 13 See Hosmer and Lameshow (1989, Ch. 3).
- 14 The 95% confidence interval for the estimated odds ratio is 0.96 to 70.32.
- 15 The minor discrepancy is due to rounding. The log of the odds ratio ( $\ln 8.23$ ) = 2.108, which is the logistic regression coefficient for the variable X.
- 16 See Conway & Roberts (1986).
- 17 See Meier, Sacks and Zabell (1986) for a discussion of two-way statistical tests used in discrimination analysis.
- 18 See Fleiss (1981) for a rather complete review of these methods.
- 19 Computed by taking the ratio of the probabilities of retention and termination for those under 40, 0.48 and 0.027, respectively. The odds are most often reported as a value compared to 1.0.
- 20 Fleiss (1981) presents some methods for determining confidence intervals for the odds ratio.
- 21 The LPM also has the statistical problems not associated with the logit model that were previously described.
- 22 Earnings capacity includes non-wage economic benefits.
- 23 A PC version of the CPEA model is available from Legal Economic Software, P.O. Box 1288, Cardiff, CA 92007.
- 24 Estimation of multinomial logit parameters may be performed using SPSS-X.

## References

- Bureau of Labor Statistics. 1987 "Most Occupational Changes are Voluntary." U S Department of Labor, USDL (Oct.): 87-452
- Conway, Delores A and Harry V Roberts 1986. "Regression Analyses in Employment Discrimination Cases " *Statistics and the Law* New York: John Wiley & Sons
- Cox, D.R. 1970. *The Analysis of Binary Data*. London. Methaen & Co.
- Finkelstein, Michael O. 1980. "The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases " *Columbia Law Review* 80(May): 737-54.
- Fleiss, Joseph L 1981. *Statistical Methods for Rates and Proportions*. New York: Wiley & Sons.
- Fisher, Franklin M. 1980. "Multiple Regression in Legal Proceedings " *Columbia Law Review* 80(May): 702-36.
- Hosmer, David W. and Stanley Lameshow. 1989. *Quantitative Variables in Econometrics*. New York: John Wiley & Sons.
- Goldberger, Arthur S. 1964 *Econometric Theory*. New York: Wiley & Sons.
- Maddala, G. S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge, England: Cambridge Univ. Press
- Markey James P., and William Parks II. 1989. "Occupational Change. Pursuing a Different Kind of Work." *Monthly Labor Review* 112(Sept.): 3-12.
- McFaddan, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior," ed. P. Zarembka. *Frontiers in Econometrics*. New York: Academic Press.
- Meier, Paul, Jerome Sacks, and Sandy Zabell. 1986. "What Happened in Hazelwood." in *Statistics and the Law*. New York: John Wiley & Sons.
- Pindyck, Robert S., and Daniel L. Rubinfeld. *Econometric Models and Economic Forecasts, 3rd ed* New York: McGraw-Hill.
- Ramanathan, Ramu. 1989. *Introductory Econometrics with Applications*. San Diego: Harcourt, Brace & Jovanovich.
- Theil, Henri. 1969. "A Multinomial Extension of the Linear Logit Model." *International Economic Review*. 10:251-59.