Lecture 12: Generalized Linear Models for Binary Data

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Bernoulli Random Variables

- Many variables can have only 2 possible values.
- That is, they are Bernoulli random variables
- Recall, for $Y = 0, 1$
- π is the probability of $Y = 1$
- $E(Y) = \mu = \pi$
- $Var(Y) = \mu(1 - \mu) = \pi(1 - \pi)$

Binomial Distribution

When we have n independent trials and take the sum of the $Y's$, we have a **binomial distribution** with

- mean = $n\pi$
- variance =$n\pi(1-\pi)$

In general, we are interested in the parameter π

We will consider models for π, which can vary according to some values of an explanatory variable(s) (i.e., $x_1, x_2, \ldots, x_p$)

To emphasize that π changes with respect to (w.r.t.) the $x's$, we write

$$\pi(x) = P(Y = 1|x_1, x_2, \ldots, x_p)$$

Linear Probability Model

- One way to model $\pi(x)$ is to use a linear model.
- For simplicity, lets consider the case where we only have one explanatory variable
- Thus,

$$\pi(x) = \alpha + \beta x$$

- Using the terminology of GLMs,
  1. The random component follows a **binomial** distribution
  2. The link is the **identity** link
  3. The systematic component contains an intercept, α and one covariate, x along with its parameter, β.

Notes about the Linear Probability Model

The $E(Y) = \pi(x)$ changes with the value of x

If $\beta < 0$, then $\pi(x)$ decreases as x increases (monotonically decreasing)

If $\beta > 0$, then $\pi(x)$ increases as x increases (monotonically increasing)

However, since $\pi(x)$ is a probability, it must be such that $\pi(x) \in [0, 1] \forall x$

For a given α, β, there could be values of x that produce estimated probabilities out of range.

Example

Suppose you have an achievement score that ranges in value from 0 to 350 and you have data on $n = 600$ students.

You want to model the probability that an individual is accepted to a 4 year college based on the achievement score.

Then, attendance of college is a Bernoulli random variable with a 'success' ($Y_i = 1$) being student i is accepted and a 'failure' ($Y_i = 0$) being student i is not accepted.

x_i is the achievement score for the i^{th} individual.

Summary data

	Achievement Score						
	<200	201-225	226-250	251-275	276-300	301-325	326-350
Not Accepted (0)	40	69	66	62	38	14	3
Accepted (1)	8	20	37	80	73	63	27
P(Y=1)	.17	.22	.36	.56	.66	.82	.90

For summary, the achievement score has been grouped into blocks of 25.

$$P(Y = 1|x < 200) = 8/48 = .166666$$

Test of Independence

Suppose, prior to formulating a regression model for the data, lets consider the simple hypothesis of independence.

After inputting the data into SAS and using PROC FREQ (you should feel comfortable recreating this by now) you get the following summary results

Statistic	df	Value	$p-$value
Pearsons X^2	6	119.83	<0.001
Likelihood ratio G^2	6	129.00	<0.001

We reject the null hypothesis that attendance and score are independent.

We will develop a regression model to explain how they are **related**.

Linear Probability Model

For this data, I do not have the raw data, so we can choose the values of x such that they represent the midpoint of each interval.

That is, $x \in (175, 213, 239, 264, 289, 314, 339)$

We can implement the linear probability model in GENMOD by the following:

```
proc genmod descending;
 freq count;
 model attend = score /link=identity dist=bin;
run;
```

Using the following data structure

Data

```
data one;
 input attend score count;
 cards;
 0 175 40
 0 213 69
 0 239 66
 0 264 62
 0 289 38
 0 314 14
 0 339 3
 1 175 8
 1 213 20
 1 239 37
 1 264 80
 1 289 73
 1 314 63
 1 339 27
 ;
run;
```

Resulting Estimated Model

Parameter	DF	Estimate	Standard Error
Intercept	1	-0.7456	0.0812
score	1	0.0049	0.0003

Or in terms of $\pi(x)$,

$$\begin{aligned} \hat{\pi}(x) &= \hat{\alpha} + \hat{\beta}x \\ &= -0.7456 + 0.0049x \end{aligned}$$

Therefore, for each 10 point increase in the score, the probability of admission increases by 0.05 (=10*0.0049)

Notes about $\pi(x)$

The following table summarizes the observed and estimated (or "fitted" $\hat{\pi}(x)$)

x	n	y1	$\pi(x)$ (y1/n)	$\widehat{\pi(x)}$
175	48	8	0.167	0.112
213	89	20	0.225	0.298
239	103	37	0.359	0.426
264	142	80	0.563	0.548
289	111	73	0.658	0.671
314	77	63	0.818	0.793
339	30	27	0.900	0.916

For this data, the linear probability model seems to function rather well.

For the domain of x, all of the estimated or fitted values for $\pi(x)$ are in $[0, 1]$.

However, this need not always be the case.

Limitations of the Linear Probability Model

- Even though the parameters of the linear model are easily interpreted, there are limitations
- A major problem with a linear model for $\pi(x)$ is that it can yield predicted values of π less than 0 and/or greater than 1.

- Example: These data are from Agresti (1990). Look for the data in the course webpage. The explanatory variable is a “labeling index” (LI) which measures the proliferative activity of cells after a patient receives an injection of a drug for treating cancer. The response variable is whether the patient achieved remission.

The estimated equation is

$$\hat{\pi}(x) = -0.2254 + 0.0278\text{LI}$$

with a full tabulated fitted values of

LI	Number of Cases	Number of Remissions	$\hat{\pi}$
8	2	0	-0.003
10	2	0	0.053
12	3	0	0.190
14	3	0	0.164
...	...	...	...
38	3	2	0.832

Here, we observe an undefined fitted value for LI=8.

Why are we using GENMOD and Not GLM?

Recall the Attendance and Test Score example.

We fit the data using PROC GENMOD. Why?

Before we answer this question, could we have fit the model in PROC GLM?

```
proc glm;
  freq count;
  model attend = score;
  run;
  quit;
```

Selected Results

GLM Results

Parameter	Estimate	Standard Error
Intercept	-.8215401218	0.11310615
score	0.0051377844	0.00042956

GENMOD Results

Parameter	DF	Estimate	Standard Error
Intercept	1	-0.7456	0.0812
score	1	0.0049	0.0003

These look close, so what is wrong?

Non-constant Variance

The linear probability model for binary data is **not** an ordinary simple linear regression problem, because

1. Non-Constant Variance
   - The variance of the dichotomous responses Y for each subject depends on x.
   - That is, The variance is not constant across values of the explanatory variable
   - The variance is
   $$Var(Y) = \pi(x)(1 - \pi(x))$$
   - Since the variance is not constant, maximum likelihood estimators of the model parameters have smaller standard errors than least squared estimators.
   - Technically speaking, ML is more efficient than least squares when you have non-constant variance.
2. Y is Bernoulli and not Normal

GENMOD uses ML based on the distribution specified in the model statement. We'll cover this concept in more detail later.

Additional Examination of the Relationship among $\pi(x)$ and x

In many cases, we would expect to see a “non-linear” association among $\pi(x)$ and x.

For example, consider the probability of buying a new car as a function of household salary.

For changes in 10,000 dollar increments, we would expect large jumps in probability as salary increased from 10,000 to 20,000; 20,000 to 30,000, etc. However, if annual salary was in the neighborhood of $1,000,000, a change from 1,000,000 to 1,010,000 would result in only a small change in probability.

A linear model is not capable of this tendency.

Ideal Properties for a model of $\pi(x)$

Properties a curve should have for modeling the relationship between $\pi(x)$ and x

1. A fixed change in x should have a smaller effect when π is close to 0 or 1 than when it is closer to the middle of the range for π.
2. The relationship between $\pi(x)$ and x is usually monotonic.

Therefore, we want some sort of “S” curve as our model.

We will examine two common non-linear models: the logistic and the probit models

Picture of a Monotonically Increasing S Curve

(picture to be drawn in class)

Logistic Model

One of the most common non-linear model for the conditional expectation of a bernoulli variable is the **logistic model**.

Instead of a linear model, consider

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

For $\beta < 0$
As $x \to \infty, \pi(x) \downarrow 0$

For $\beta > 0$
As $x \to \infty, \pi(x) \uparrow 1$

As we discussed previously, the link for a logistic model is the logit transformation

$$log(\frac{\pi(x)}{1 + \pi(x)}) = \text{logit}(\pi(x)) = \alpha + \beta x$$

Snoring Example

	Heart Disease		Proportion
Snoring	Yes	No	Yes
Never	24	1355	0.017
Occasionally	35	603	0.055
Nearly every night	21	192	0.099
Every night	30	224	0.118

Our outcome is heart disease, and in order to use the ordinal levels of snoring, we need to select scores.

A set (0, 2 , 4, 5) seems to capture the relative magnitude of the differences among the categories.

Alternative Data Structure

- Previously, we looked at modeling the binomial outcome directly (so called “single trial” structure).
- You can in SAS use the “event/trials” syntax.

For Event/trial data, you would enter the data as

```
data two;
  input snoring hdyes hdno;
  hdtotal = hdyes + hdno;
  cards;
  0 24 1355
  2 35 603
  4 21 192
  5 30 224
  ;
run;
```

Advantages of Event/Trials

The advantages of the event/trial layout are:

1. If you have tabular data, you will need to type less data into your program
2. As we will see, you do not need to worry about the “descending” option
3. It is the most common layout for Epi data (“cases” are grouped by factor levels of x)

The “single trial” syntax is best if you have the raw data (data row i represents the bernoulli outcome for the i^{th} individual)

Implementation of Event/Trials in GENMOD

```
proc genmod;
 model hdyes / hdtotal = snoring /dist=bin link=logit;
run;
```

The results of this model are:

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept	1	-3.8662	0.1662	-4.1920	-3.5405
snoring	1	0.3973	0.0500	0.2993	0.4954

Or,

$$\text{logit}(\hat{\pi}(x)) = -3.87 + 0.40x$$

Here $\beta > 0$ so the fitted probabilities increase with x.

Table of Fitted Values

	Heart Disease		Proportion	Logit
Snoring Score	Yes	No	Yes	Fitted
0	24	1355	0.017404	0.020508
2	35	603	0.054859	0.044294
4	21	192	0.098592	0.093046
5	30	224	0.11811	0.132423

Alternative Data Input

```
data three;
 input snoring y count;

cards;
0 1 24
0 0 1355
2 1 35
2 0 603
4 1 21
4 0 192
5 1 30
5 0 224
  ;
run;
proc genmod descending data=three;
 freq count;
 model y = snoring /dist=bin link=logit;
run;
```

Alternative Data Results

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error
Intercept	1	-3.8662	0.1662
snoring	1	0.3973	0.0500

Note: these are the same results, as expected, as the event/trials coding.

Notes about Logit Transformation

Recall,
logit(π) = log($\pi/(1-\pi)$)

The term “logit” was coined to make the previous standard non-linear model, the Probit, based on the normal distribution.

The logit is the natural parameter of the binomial distribution and as such the logit link is the canonical link.

Whereas $0 \leq \pi \leq 1$, the range for logit(π) is all real numbers, $-\infty <$ logit$(\pi) < \infty$

The systematic component, $x\beta$, can be any real number and it will produce a fitted value for π within (0,1).

The greater the $|\beta|$, the greater the steeper the S-Curve

Probit Link

A monotonically increasing S-shaped curve is similar to the cumulative distribution function for some random variable.

Therefore, we could model $\pi(x) = F(x)$ for some cdf F.

To control the shape of the S-curve, we essentially need two parameters - the location and the "scale"

By selecting the cdf for a normal distribution, we have the flexibility of both the location (by selection of the mean) and the scale (by selection of the variance).

The probit link is defined as

$$\text{probit}(\pi) = F^{-1}(X \leq x)$$

where F is the standard normal distribution.

Standard Normal Probit

For example,

$$\begin{aligned}
\text{probit}(0.025) &= -1.96 \\
\text{probit}(0.05) &= -1.64 \\
\text{probit}(0.0) &= 0.0 \\
\text{probit}(0.95) &= 1.64 \\
\text{probit}(0.975) &= 1.96
\end{aligned}$$

or in terms of a GLM,

$$\text{probit}(\pi(x)) = \alpha + \beta x$$

where the random component is binomial and the link function is **probit**.

Probit Analysis in SAS

Recall our snoring example.

To fit the Probit link in SAS, all you need to do is specify the link as probit in the GENMOD model statement using either single trial or event coding.

```
proc genmod descending data=three;
 freq count;
 model y = snoring /dist=bin link=probit;
run;

---- or ------

proc genmod data=two;
 model hdyes / hdtotal = snoring /dist=bin link=probit;
 run;
```

Selected Results

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept	1	-2.0606	0.0704	-2.1986	-1.9225
snoring	1	0.1878	0.0236	0.1415	0.2341

Recall that

$$\pi(x) = \Phi(\alpha + \beta x)$$

so the fitted values are

$$\hat{\pi}(x) = \Phi(-2.0606 + 0.0236x)$$

Summary Slide

	Heart Disease		Proportion	Logit	Probit
Snoring Score	Yes	No	Yes	Fitted	Fitted
0	24	1355	0.017404	0.020508	0.01967054
2	35	603	0.054859	0.044294	0.04599426
4	21	192	0.098592	0.093046	0.09519951
5	30	224	0.11811	0.132423	0.13101632

We see that the Logit and Probit models produce similar results.

Informal Poll:
How many people, prior to this class, have heard of a logistic regression model? How many for the probit regression model?

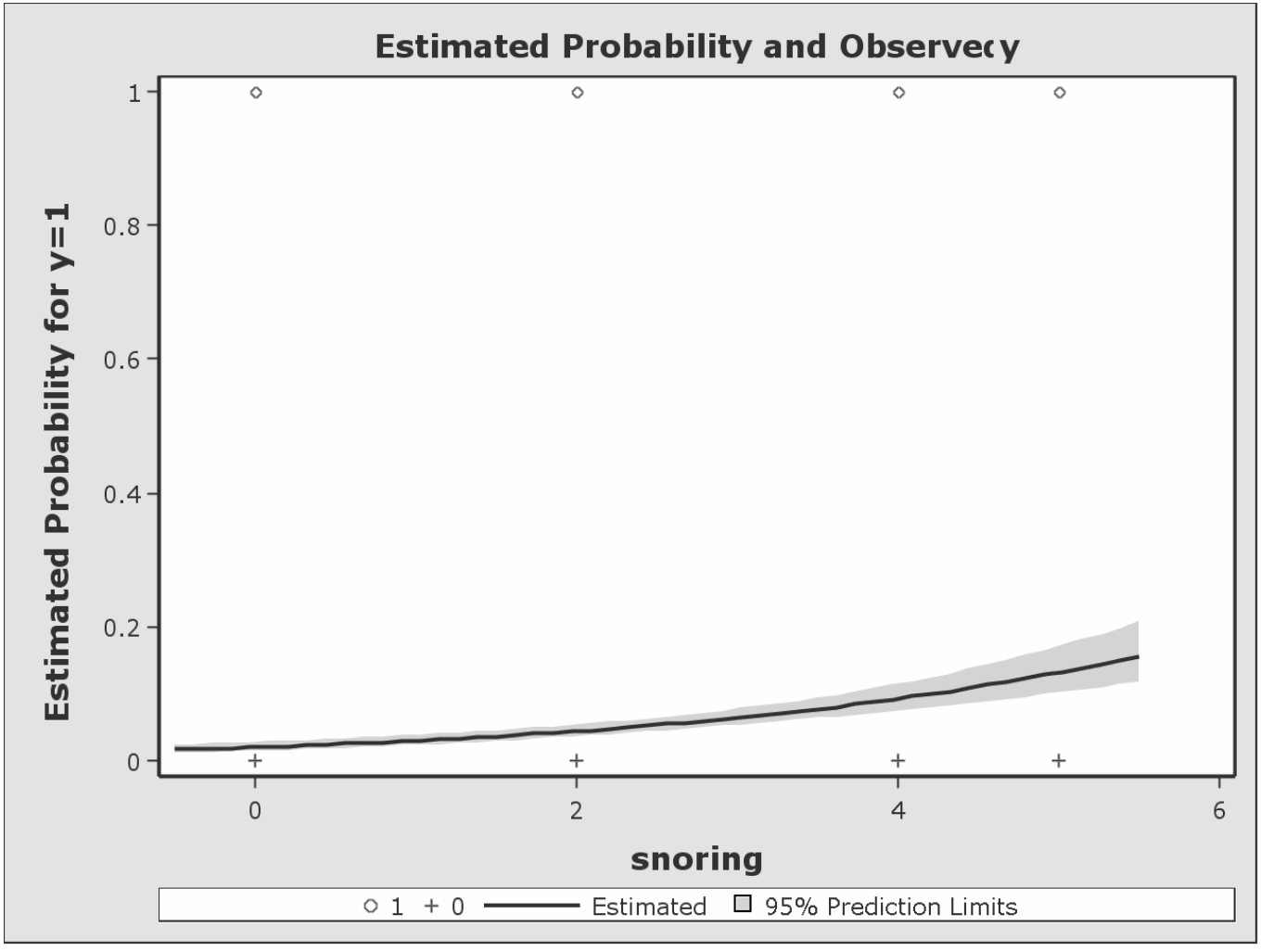
Notes

- Beginning in version 9.1, there is a preliminary release of SAS ODS Graphics
- This produces “publication ready” figures
- Let’s reanalyze the snoring data using PROC LOGISTIC (PROC GENMOD doesn’t yet have all of the graphics)
- In case you couldn’t guess it, PROC LOGISTIC fits a logistic regression model.

SAS to Latex

```
ods latex file="lecsas.tex"
          gpath="\\Dbe\teaching\11S Cat Data Analysis\l12"
          path="\\Dbe\teaching\11S Cat Data Analysis\l12"
          style=science; /* many more styles available */
ods graphics on;
proc logistic descending data=three;
 freq count;
 model y = snoring;
 graphics estprob;
run;
ods graphics off;
ods latex close;
```

Example figure



Recall the Labeling Index Study

- Recall the labeling index study
- The linear probability model did not fit the data well
- Lets consider the logistic model for the analysis

```
data labeling;
input li numcase numrem;
cards;
 8 2 0
 10 2 0
 12 3 0
 14 3 0
 16 3 0
 18 1 1
 20 3 2
 22 2 1
 24 1 0
 26 1 1
 28 1 1
 32 1 0
 34 1 1
 38 3 2
 ;
run;

ods latex file="lec12bsas.tex"
       gpath="\\Dbe_305c_a\teaching\05F Cat Data Analysis\l12"
       path="\\Dbe_305c_a\teaching\05F Cat Data Analysis\l12"
       style=statistical;
ods graphics on;
proc logistic  data=labeling;
 model numrem / numcase = li;
 graphics estprob;
run;
ods graphics off;
ods latex close;
```

SAS Results

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

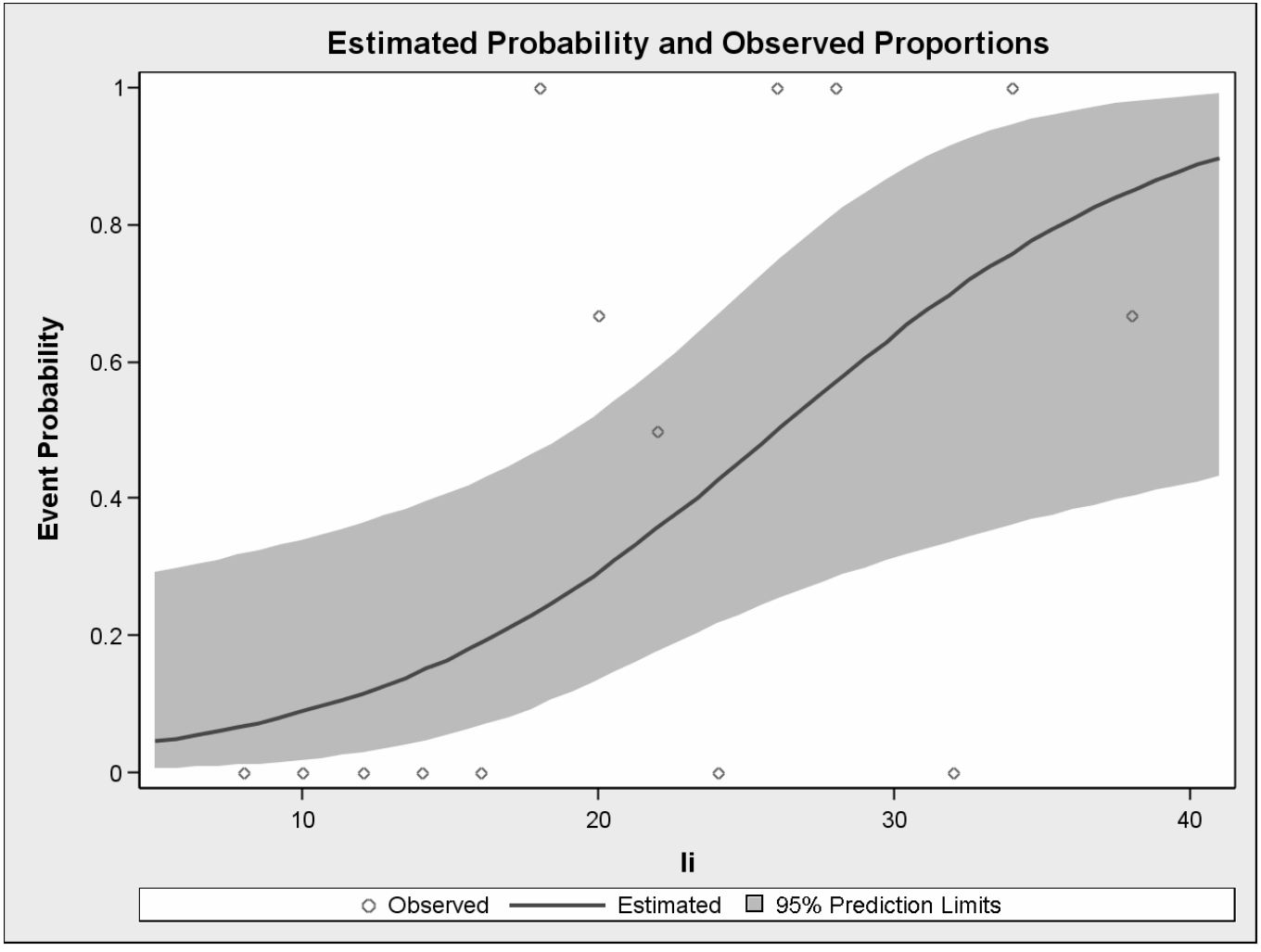
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.7771	1.3786	7.5064	0.0061
li	1	0.1449	0.0593	5.9594	0.0146

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
li	1.156	1.029	1.298

The odds of remission increase 1.16 times for every 1 unit increase in the labeling index.

Example figure



Popularity of the Logistic Model

The logistic model has several factors going for it:

1. It uses the canonical link: although not required, many purist favor it.
2. Parameter estimates are log-odds ratios.
3. Parameter estimates for probit models do not have a common meaning. Although, it is useful for predicting the success probability.

We will examine additional links for binomial data in the near future.

Modeling Binary data trivia

The earliest non-linear transformation of π dates to 1886 (Fechner)

The probit link was popularized by Gaddum (1933) and Bliss (1934, 1935) in toxicological experiments.

The term Probit was coined by Bliss

Fisher & Yates (1938) first suggested a logit link function

Berkson (1944) introduced the term “logit” because of the similarity between the logit and probit models