Introduction to Generalized Linear Models for Binary Data
Edpsy/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology

© Board of Trustees, University of Illinois

Spring 2017
Outline

Introduction to GLMs for binary data

Primary Example: High School & Beyond.

- The problem
- Linear model for $\pi$.
- Modeling Relationship between $\pi(x)$ and $x$.
- Logistic regression.
- Probit models.
- Trivia
The Problem

- Many variables have only 2 possible outcomes.
- Recall: Bernoulli random variables
  - \( Y = 0, 1 \).
  - \( \pi \) is the probability of \( Y = 1 \).
  - \( E(Y) = \mu = \pi \).
  - \( \text{Var}(Y) = \mu(1 - \mu) = \pi(1 - \pi) \).
- When we have \( n \) independent trials and take the sum of \( Y \)'s, we have a Binomial distribution with
  \[
  \text{mean} = n\pi \quad \text{and} \quad \text{variance} = n\pi(1 - \pi).
  \]
- We are typically interested in \( \pi \).
- We will consider models for \( \pi \), which can vary according to some the values of an explanatory variable(s) (i.e., \( x_1, \ldots, x_k \)).
- To emphasis that \( \pi \) changes with \( x \)'s, we write \( \pi(x) \).
Example: High School & Beyond

- Data from seniors (N=600).
- Consider whether students attend an academic high school program type of a non-academic program type (Y).
- We would like to know whether the probability of attending an academic program $\pi(x)$ is related to achievement ($x$).
- Scores on 5 standardized achievement tests are available (Reading, Writing, Math, Science, and Civics), so we’ll just take the sum as a measure of achievement ($x$).
Linear Model for $\pi$

- One way to model $\pi(x)$ is to use a linear model:
  \[ \pi(x) = \alpha + \beta x \]

- This is a “Linear Probability Model” — probability changes linearly with achievement ($x$).

- $\beta$ represents how much larger (smaller) the probability of attending an academic high school program for a unit change in achievement.

- GLM components of linear probability model:
  - **Random** — $Y$ is attending academic program and has a Binomial distribution.
  - **Systematic** — $X$ is the sum of achievement test scores.
  - **Link** — Identity.
HSB Data

Of the 600 students, there are 490 different values of $x$, so if we compute observed proportions for each of the observed values of $x$, many would be either 0 or 1 (i.e., $y = 0, 1$).

To get a look at the relationship, we can group the data (i.e., collapse $x$ into some number of categories).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>no (0)</td>
<td>39</td>
<td>68</td>
<td>67</td>
<td>62</td>
<td>37</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>82.98</td>
<td>79.07</td>
<td>63.21</td>
<td>44.60</td>
<td>33.04</td>
<td>20.25</td>
<td>9.68</td>
</tr>
<tr>
<td>yes (1)</td>
<td>8</td>
<td>18</td>
<td>39</td>
<td>77</td>
<td>75</td>
<td>63</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>17.02</td>
<td>20.93</td>
<td>36.79</td>
<td>55.40</td>
<td>66.96</td>
<td>79.75</td>
<td>90.32</td>
</tr>
</tbody>
</table>

Is there is relationship? Could it be linear?
Plot of Collapsed Data

- Observed Proportion at Academic
- MEAN Sum of Achievement Test Scores

C.J. Anderson (Illinois)  Introduction to GLMS for Binary Data  Spring 2017
Tests of Statistical Relationship

Statistical Tests of Independence and Linear relationship:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Value</th>
<th>p–value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>6</td>
<td>119.21</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>6</td>
<td>128.25</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>117.03</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

The scores used for the test of ordinal (linear) relationship were the mean values of the achievement test score categories.

Note: $r = .44$
Linear Probability Model for HSP

Estimated Linear Model for probability of attending an academic high school program:

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

$$= -0.8970 + 0.0054x$$

where $x$ is mean of sum of the 5 achievement scores.

The estimated expected values for $E(Y) (\hat{\pi})$ are “fitted values”.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Academic</th>
<th>Non-acad.</th>
<th>Proportion</th>
<th>Fitted</th>
<th>Std. Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>188.972</td>
<td>8</td>
<td>39</td>
<td>0.17</td>
<td>0.13</td>
<td>1.08</td>
</tr>
<tr>
<td>212.748</td>
<td>18</td>
<td>68</td>
<td>0.21</td>
<td>0.26</td>
<td>-1.17</td>
</tr>
<tr>
<td>236.868</td>
<td>39</td>
<td>67</td>
<td>0.37</td>
<td>0.39</td>
<td>-0.47</td>
</tr>
<tr>
<td>262.806</td>
<td>77</td>
<td>62</td>
<td>0.55</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>287.342</td>
<td>75</td>
<td>37</td>
<td>0.67</td>
<td>0.66</td>
<td>0.17</td>
</tr>
<tr>
<td>312.080</td>
<td>63</td>
<td>16</td>
<td>0.80</td>
<td>0.80</td>
<td>0.01</td>
</tr>
<tr>
<td>332.713</td>
<td>28</td>
<td>3</td>
<td>0.90</td>
<td>0.91</td>
<td>-0.14</td>
</tr>
</tbody>
</table>
Looking at fit of Linear Probability Model

![Graph showing the relationship between observed proportion at academic and mean sum of achievement test scores.](image)

- **Proportion**
- **Fitted**

- **Observed Proportion at Academic**
- **MEAN Sum of Achievement Test Scores**

- **C.J. Anderson (Illinois)**
- **Introduction to GLMS for Binary Data**

- **Spring 2017 10.1/45**
Structural Problem w/ Linear Probability Models

A linear model for $\pi(x)$ can yield predicted values $< 0$ and/or $> 1$.

Example: These data are from Lee (1974; Agesti, 1990). 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.

Below are the fitted values from a linear model fit to these data:

<table>
<thead>
<tr>
<th>LI</th>
<th>Number of Cases</th>
<th>Number of Remissions</th>
<th>$\hat{\pi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>−.003</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0</td>
<td>.053</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0</td>
<td>.109</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>0</td>
<td>.164</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>2</td>
<td>.832</td>
</tr>
</tbody>
</table>
Plot of LI–remission data & Fitted

Observed Proportions and Linear Prob Model
...and 95% Confidence Bands...

- Data
- Fitted
- Lower
- Upper

Fitted Proportion vs. Labeling Index (LI)
**Plot of LI–remission data & Fitted**

Typo somewhere... When I fit these data...

<table>
<thead>
<tr>
<th>li</th>
<th>nremit</th>
<th>ncases</th>
<th>LinFit</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>2</td>
<td>0.00003</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2</td>
<td>0.05339</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0.10676</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>3</td>
<td>0.16012</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>3</td>
<td>0.21349</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>0.26685</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>3</td>
<td>0.32021</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>2</td>
<td>0.37358</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>1</td>
<td>0.42694</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>1</td>
<td>0.48031</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>1</td>
<td>0.53367</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>1</td>
<td>0.64040</td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td>1</td>
<td>0.69376</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>3</td>
<td>0.80049</td>
</tr>
</tbody>
</table>
Linear Probability Model is a GLM

Linear probability model for binary data is **not** an ordinary simple linear regression problem.

- The variance of the dichotomous responses $Y$ for each subject depends on $x$.
- The variance is not constant across values of the explanatory variable, but rather it equals

$$\text{var}(Y) = \pi(x)(1 - \pi(x))n$$

- Since the variance is not constant, maximum likelihood estimators of the model parameters have smaller standard errors than least squares estimators.
Linear Probability Model is a GLM

Example: Cancer remission data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MLE Estimate</th>
<th>MLE Std Error</th>
<th>OLS Estimate</th>
<th>OLS Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $\alpha$</td>
<td>-0.2134</td>
<td>0.2768</td>
<td>-0.1613</td>
<td>0.2790</td>
</tr>
<tr>
<td>Slope, $\beta$</td>
<td>0.0267</td>
<td>0.0104</td>
<td>0.0268</td>
<td>0.0120</td>
</tr>
</tbody>
</table>
Modeling Relationship between $\pi(x)$ and $x$

First Property a curve should have

- A fixed change in $x$ should have a smaller effect when $\pi$ is close to 0 or 1 than when it is closer to the middle of the range for $\pi$.

- Generally, when $\pi(x)$ is close to 0 or 1, a fixed change in $x$ has less of an effect than when $\pi(x)$ is closer to the middle of it’s range.

Example: Probability of getting a moderately difficult item correct as a function of total number of items correct (without the item).

$$g(P(\text{item correct})) = \alpha + \beta x \quad \text{where } x \text{ is rest-score.}$$

- On 100 item test, we would would expect a larger increase in $P(\text{item correct})$ when $x$ goes from 50 to 60 than when when $x$ goes from 89 to 99.

- We would also expect to see a smaller decrease in $P(\text{item correct})$ when $x$ goes from 10 to 0 than when $x$ goes from 60 to 50.
Second Property for Curve

The relationship between $\pi(x)$ and $x$ is usually monotonic such that

\[ \pi(x) \text{ continuously increases as } x \text{ increases} \]

or

\[ \pi(x) \text{ continuously decreases as } x \text{ increases.} \]

Considering these two properties, an S-shaped curve:
Cumulative Distribution Functions (cdf)

Suppose that
- $Z$ is a random variable
- $z$ is a possible value of $Z$ (e.g., $-\infty < z < \infty$)

A cumulative distribution function for $Z$ is defined as

$$F(z) = P(Z \leq z) \quad -\infty < z < \infty$$

Some examples for discrete $Z$:
Symmetric pdf’s (bell shaped)

that have symmetric probability density functions (i.e., “bell-shaped” ones).

Two distributions that we will discuss are

- Logistic
- Normal
Logistic regression

The cumulative logistic distribution function is

$$F(x) = P(X \leq x) = \frac{\exp((x - \mu)/\tau)}{1 + \exp((x - \mu)/\tau)}$$

where

- $\mu$ is a mean (location)
- $\tau$ is a scaling parameter
- $-\infty < x < \infty$

Using the logistic $cdf$, the logistic regression function is

$$\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \log \left( \frac{\exp(x - \mu)\tau}{1 + \exp((x - \mu)/\tau)} \right)$$

$$= (x - \mu)/\tau$$

$$= -\mu/\tau + x/\tau$$

$$= \alpha + \beta x$$
Logistic regression model as a GLM

\[ \log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \alpha + \beta x \]

- **Random component**: Binomial
- **Link Function**: logit
  - “logit(\(\pi\))” = log(\(\pi/(1 - \pi)\)).
  - “logistic regression model” \(\equiv\) “logit model”
  - The logit is the *natural parameter* of the Binomial distribution; therefore, the logit link is the *canonical link* function.
  - \(0 \leq \pi \leq 1\), but \(-\infty < \text{logit}(\pi) < \infty\).
- **Systematic component**: A linear predictor such as \(\alpha + \beta x\)

which can be any Real number and yield a \(\pi\) within \((0, 1)\).
log \left( \frac{\pi(x)}{1-\pi(x)} \right) = \alpha + \beta x

**Interpretation of $\beta$:**

$\beta$ determines the rate that $\pi(x)$ changes with changes in $x$.

If $\beta > 0$ then $\pi(x)$ increases as $x$ increases.

$\alpha = 1$ and $\beta = 2$: 
The Problem

Linear Model for $\pi$

Relationship $\pi(x) \& x$

Logistic regression

Probit models

Triva

The relationship $\pi(x)$ can be modeled using the following logarithmic function:

$$\log \left( \frac{\pi(x)}{1-\pi(x)} \right) = 1 - 2x \beta$$

If $\beta < 0$ then $\pi(x)$ decreases as $x$ increases.
The Problem

Linear Model for $\pi$

Relationship $\pi(x) \& x$

Logistic regression

Probit models

Triva

\[
\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = 1 + 0x
\]

$\beta = 0$ means no relationship between $Y$ and $x$:
Changing Value of $\beta$

Larger value of $\beta$ leads to a steeper curve:
Changing Value of $\alpha$

Larger value of $\alpha$ leads to a vertical shift for the logit, but a horizontal shift for the $\pi$:
When $\alpha = 0$ with different $\beta$’s

The logits intersect at $x = 0$ with logit = 0 and the probabilities intersect at $x = 0$ with $\pi(0) = .5$:
HSB and Academic Programs (grouped data)

$Y = 1$ for academic program and $x =$ mean of total achievement test scores:

- Estimated equation:

$$\text{logit}(\hat{\pi}(x)) = -6.741 + .026x$$

- $\hat{\beta} = .026$ — as achievement scores go up, the probability that a student went to an academic high school program increases.

Is this a “large” value?

- **Statistically?** The estimated standard error of $\hat{\beta}$ is .0026, so

  $$.026 \pm 2(.0026) \rightarrow (.021, .031).$$

- **Important?** This is not a subjective judgement.
Fitted and Observed Values (grouped data)

The fitted values $\hat{\pi}(x)$ from logistic regression model:

<table>
<thead>
<tr>
<th>Achievement Category</th>
<th>Mean Achieve</th>
<th>Observed proportion</th>
<th>Fitted Values (probabilities)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Linear</td>
</tr>
<tr>
<td>162–200</td>
<td>188.97</td>
<td>.17</td>
<td>.13</td>
</tr>
<tr>
<td>201–225</td>
<td>212.75</td>
<td>.21</td>
<td>.26</td>
</tr>
<tr>
<td>226–250</td>
<td>236.87</td>
<td>.37</td>
<td>.39</td>
</tr>
<tr>
<td>251–275</td>
<td>262.81</td>
<td>.55</td>
<td>.53</td>
</tr>
<tr>
<td>276–300</td>
<td>287.34</td>
<td>.67</td>
<td>.66</td>
</tr>
<tr>
<td>301–325</td>
<td>312.08</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>326–350</td>
<td>332.71</td>
<td>.90</td>
<td>.91</td>
</tr>
</tbody>
</table>
Plot of Fitted and Observed Values

- Proportion
- Fitted

MEAN Sum of Achievement Test Scores

0.0 0.2 0.4 0.6 0.8 1.0

180 200 220 240 260 280 300 320 340
Comparison with Linear Prob Model

- Proportion
- Logistic
- Linear

Graph showing the comparison between the proportion, logistic, and linear models against the mean sum of achievement test scores.
Probit models (Normal cdf)

Rather than use the logistic cdf, we can use the (standard) Normal distribution.

When $F(z)$ is the normal cdf, the link is referred to as “probit”.

The probit link is defined as

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

For example,

- $\text{probit}(.025) = -1.96$
- $\text{probit}(.050) = -1.64$
- $\text{probit}(.500) = 0.00$
- $\text{probit}(.950) = 1.64$
- $\text{probit}(.975) = 1.96$
Probit model as GLM

The probit model for binary data:

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

This is a generalized linear model with

- Random component: Binomial distribution.
- Systematic component: linear function of explanatory variable(s).
- Link function: probit.
Example of Probit model: HSB

\[ \text{probit}(\hat{\pi}(x)) = -4.0828 + 0.0158x \]

- As achievement scores go up, so does the probability of having attended an academic program.
- Estimated standard error of \( \hat{\beta} \) is 0.0015, so \( \hat{\beta} = .0158 \) “large” in the sense that \( .0158 \pm 2(.0015) \rightarrow (0.013, 0.019) \).
- Fitted values are *extremely* close to those from the logit model.

<table>
<thead>
<tr>
<th>Category</th>
<th>( x )</th>
<th>( p )</th>
<th>Linear</th>
<th>Logistic</th>
<th>Probit</th>
</tr>
</thead>
<tbody>
<tr>
<td>162–200</td>
<td>188.97</td>
<td>.17</td>
<td>.13</td>
<td>.14</td>
<td>.14</td>
</tr>
<tr>
<td>201–225</td>
<td>212.75</td>
<td>.21</td>
<td>.26</td>
<td>.24</td>
<td>.24</td>
</tr>
<tr>
<td>226–250</td>
<td>236.87</td>
<td>.37</td>
<td>.39</td>
<td>.37</td>
<td>.37</td>
</tr>
<tr>
<td>251–275</td>
<td>262.81</td>
<td>.55</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>276–300</td>
<td>287.34</td>
<td>.67</td>
<td>.66</td>
<td>.68</td>
<td>.68</td>
</tr>
<tr>
<td>301–325</td>
<td>312.08</td>
<td>.80</td>
<td>.80</td>
<td>.81</td>
<td>.81</td>
</tr>
<tr>
<td>326–350</td>
<td>332.71</td>
<td>.90</td>
<td>.91</td>
<td>.88</td>
<td>.88</td>
</tr>
</tbody>
</table>
Observed and Fitted Values: HSB
Logit & Probit Models

The logistic regression model

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

and the probit model

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

often yield very similar fitted values.

- It is extremely rare for one of these models to fit substantially better (worse) than the other.

- The Probit model yields curves for $\pi(x)$ that look like normal $cdf$ with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$.
For the HSB data, the probit model corresponds to a normal cdf with mean $\mu = -(-4.0828)/.0158 = 258.41$ and standard deviation $\sigma = 1/.0158 = 63.29$.

For the probit model, the $x$ that yields $\hat{\pi}(x) = .5$ is the mean; that is,

$$\hat{\pi}(258.41) = .5$$

and for the logit model,

$$\hat{\pi}(258.41) = \logit^{-1}(-6.741 + .0262(258.41))$$

$$= \frac{1}{1 + \exp(-(-6.741+.0262(258.41)))}$$

$$= .5$$
SAS/PROC GENMOD

DATA hsb;
  INPUT achieve attend ncase;
  LABEL achieve=’Mean Achievement for Category’
    attend=’Number who attend academic’
    ncase=’Number who in Achievement Category’;
DATALINES;

Then to fit a **linear model** for binary data,
PROC GENMOD ORDER=DATA DATA=hsb;
  MODEL attend/ncase = achieve / LINK=identity
    DIST=BINOMIAL ;
SAS/PROC GENMOD

For a **logit** model for binary data,

```sas
PROC GENMOD ORDER=DATA DATA=hsb;
  MODEL attend/ncase = achieve / LINK=logit
  DIST=BINOMIAL ;
```

For a **probit** model for binary data,

```sas
PROC GENMOD ORDER=DATA DATA=hsb;
  MODEL attend/ncase = achieve / LINK=probit
  DIST=BINOMIAL ;
```
Example # 2: Cancer remission data

Example: These data are from Lee (1974; Agesti, 1990). 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.

- Linear probability model: $Y = -0.2134 + 0.0267(LI)$
  WARNING: The relative Hessian convergence criterion of 0.1210654506 is greater than the limit of 0.0001. The convergence is questionable.
  WARNING: The procedure is continuing but the validity of the model fit is questionable.

- Logit model: $\logit(Y) = -3.7771 + 0.1449(LI)$

- Probit model: $\probit(Y) = -2.3178 + 0.0878(LI)$
Comparison of Model Fitted Values

- linear
- logit
- probit

Observation & Fitted Proportions vs. Labeling Index (LI)
Comparison of Model Fitted Values & Data

![Graph showing comparison of model fitted values and data]

- Data
- Linear
- Logit
- Probit

Observed & Fitted Proportions vs. Labeling Index (LI)
Binary Data Modeling Trivia

▷ Probit Model.
  ▷ First person known to have suggested using the inverse of the normal \(cdf\) to transform probabilities was
  \[Fechner (1886).\]
  ▷ The probit model was popularized by Gaddum (1933) and Bliss (1934, 1935) in toxicological experiments.
  ▷ The term “probit” was introduced by
  \[Bliss — who \textit{used a normal cdf with } \mu = 5 \text{ and } \sigma = 1.\]

▷ Logit Model...
Binary Data Modeling Trivia

- Logit Model...
  - The term “logit” was proposed by Berkson (1944) because of the similarity between the logit and probit models.
  - Fisher & Yates (1938) first suggested a logit link function for binary data.

- Both the logit and probit model were derived from a “threshold model” where there is an underlying psychological quantity such that

  \[ y = \begin{cases} 
  1 & \text{if } \psi \geq \xi \\
  0 & \text{if } \psi < \xi 
  \end{cases} \]
To be covered Later

When we cover chapter 4 on logistic regression, we’ll talk about (among other things)

- More on the interpretation of logit/logistic regression model.
- Assessing model fit.