Introduction to Generalized Linear Models for Dichotomous Response Variables

Edps/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

Outline

Introduction to GLMs for binary data

Primary Example: High School & Beyond.

- The problem
- Linear model for π.
- Modeling Relationship between $\pi(x)$ and x.
- Logistic regression.
- Probit models.
- Trivia
- Graphing: jitter and loews

The Problem

- Many variables have only 2 possible outcomes.
- Recall: Bernoulli random variables
  - $Y = 0, 1$.
  - π 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

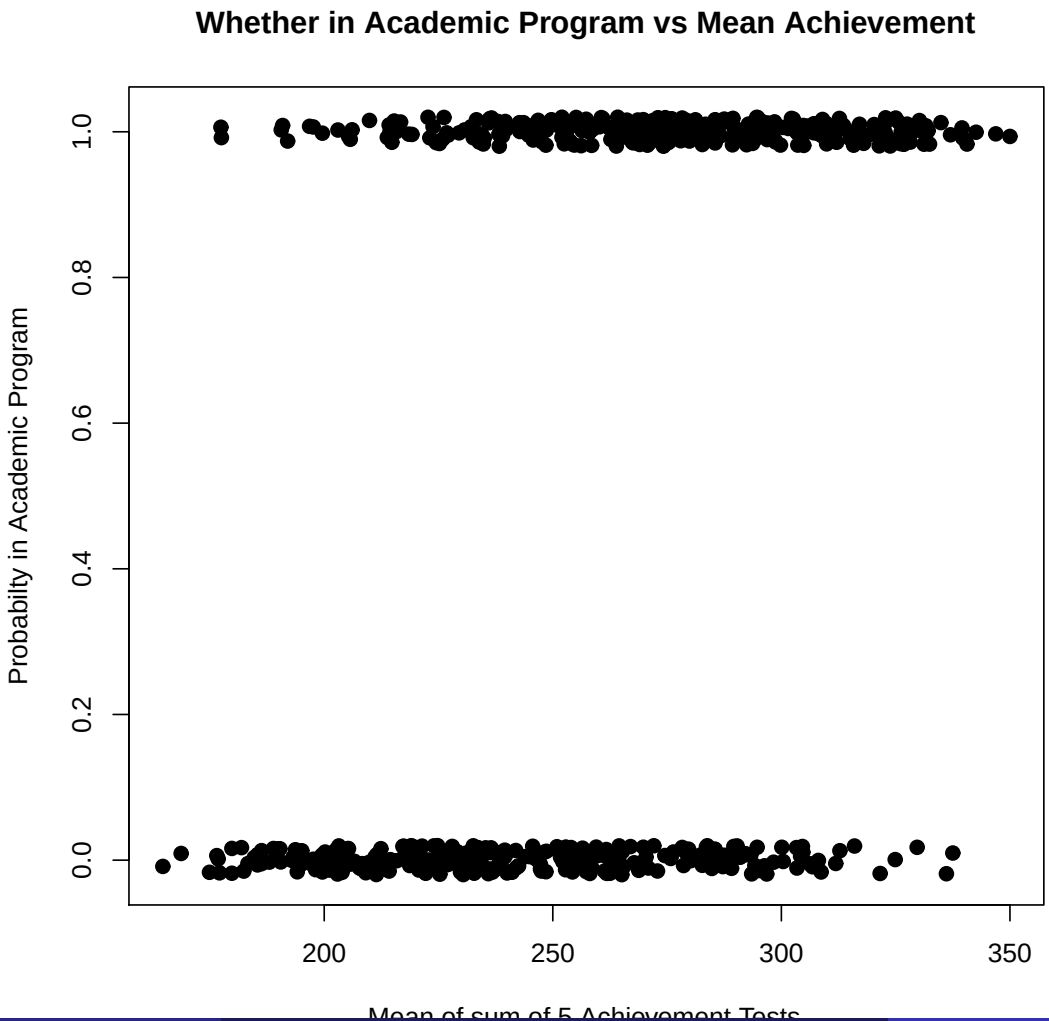
$$\text{mean} = n\pi \quad \text{and} \quad \text{variance} = n\pi(1 - \pi).$$

- We are typically interested in π.
- We will consider models for π, which can vary according to some the values of an explanatory variable(s) (i.e., $x_1, \ldots, x_k$).
- To emphasis that π changes with x's, we write $\pi(x)$

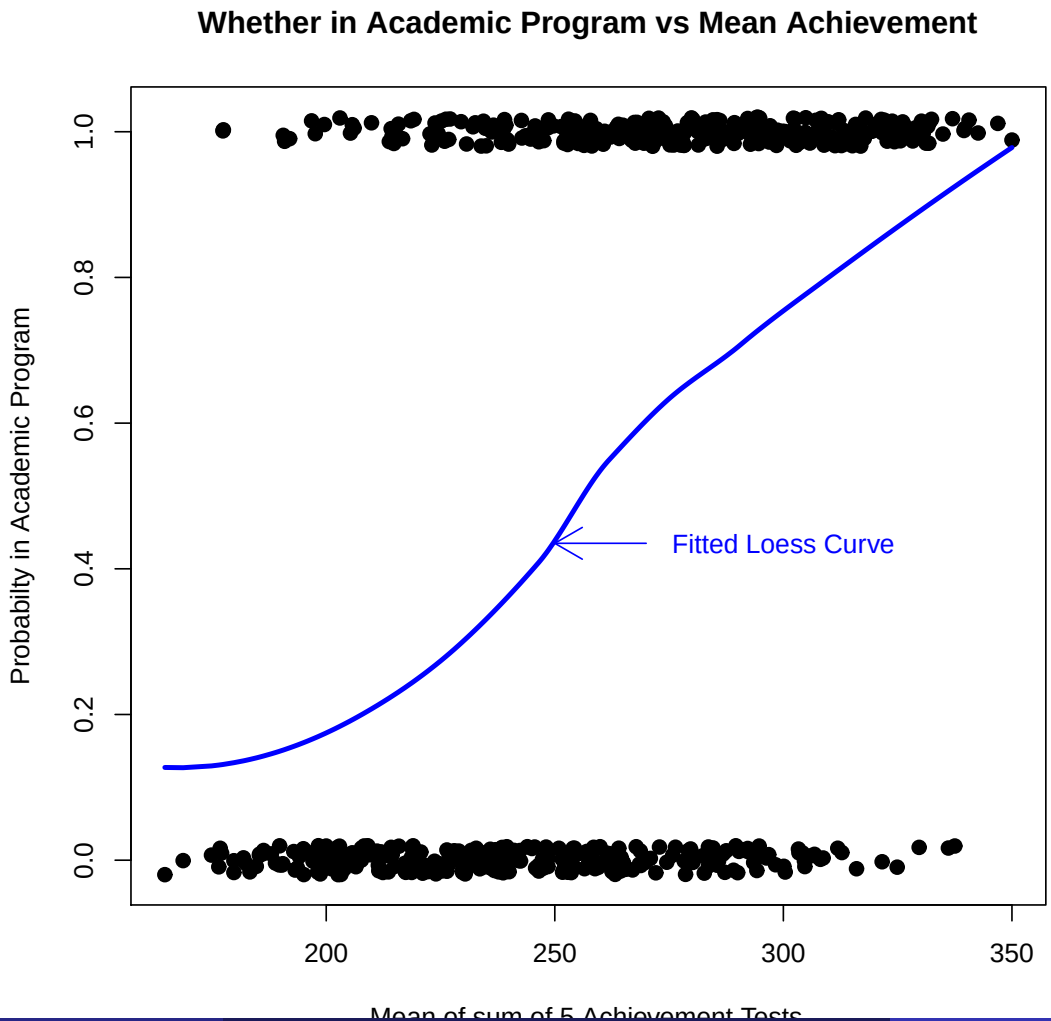
Example: High School & Beyond

- 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).

Graph of Data ("jittered")



Data with Smooth Curve



Linear Model for π

- 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).
- β 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

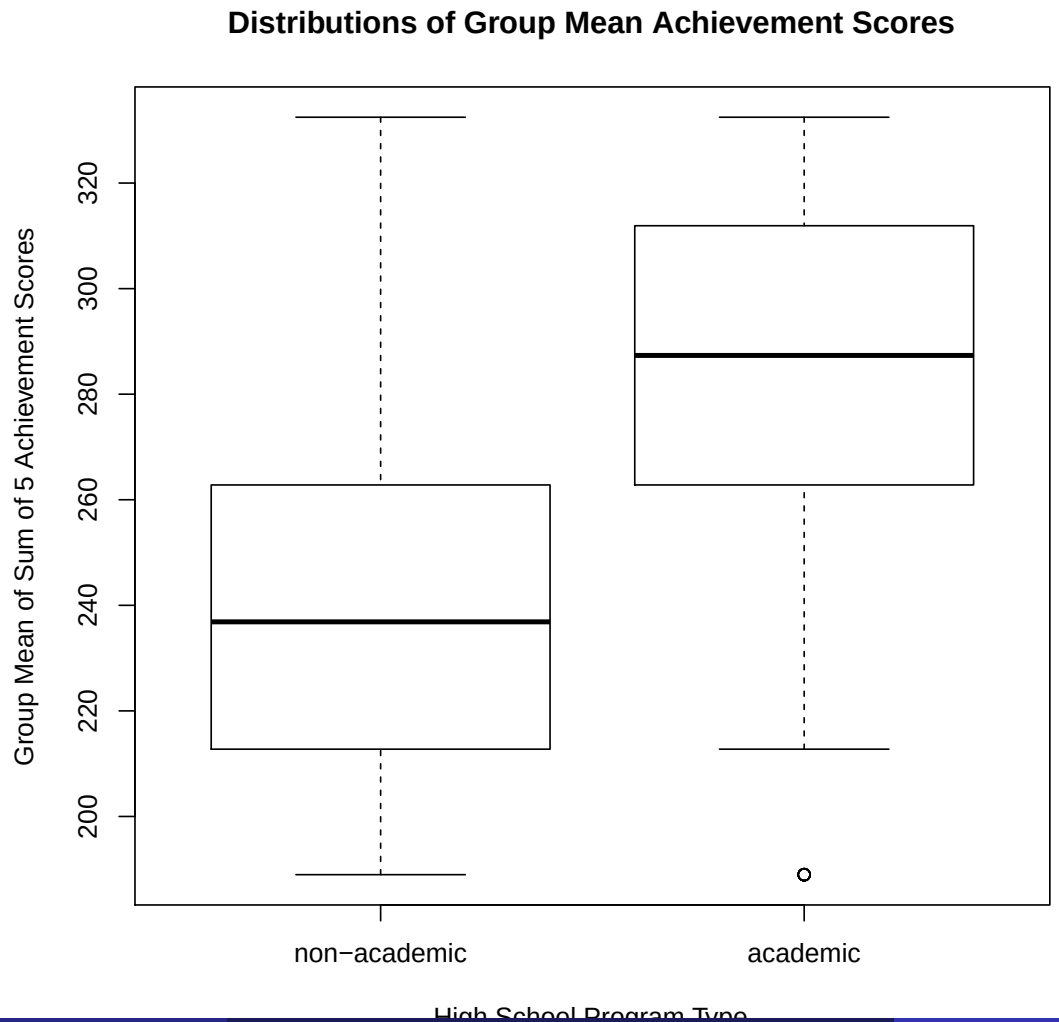
Of the 600 students, there are 490 different values of x, so if we compute observed proportions for each of the observed values of x, many would be either 0 or 1 (i.e., $y = 0, 1$).

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

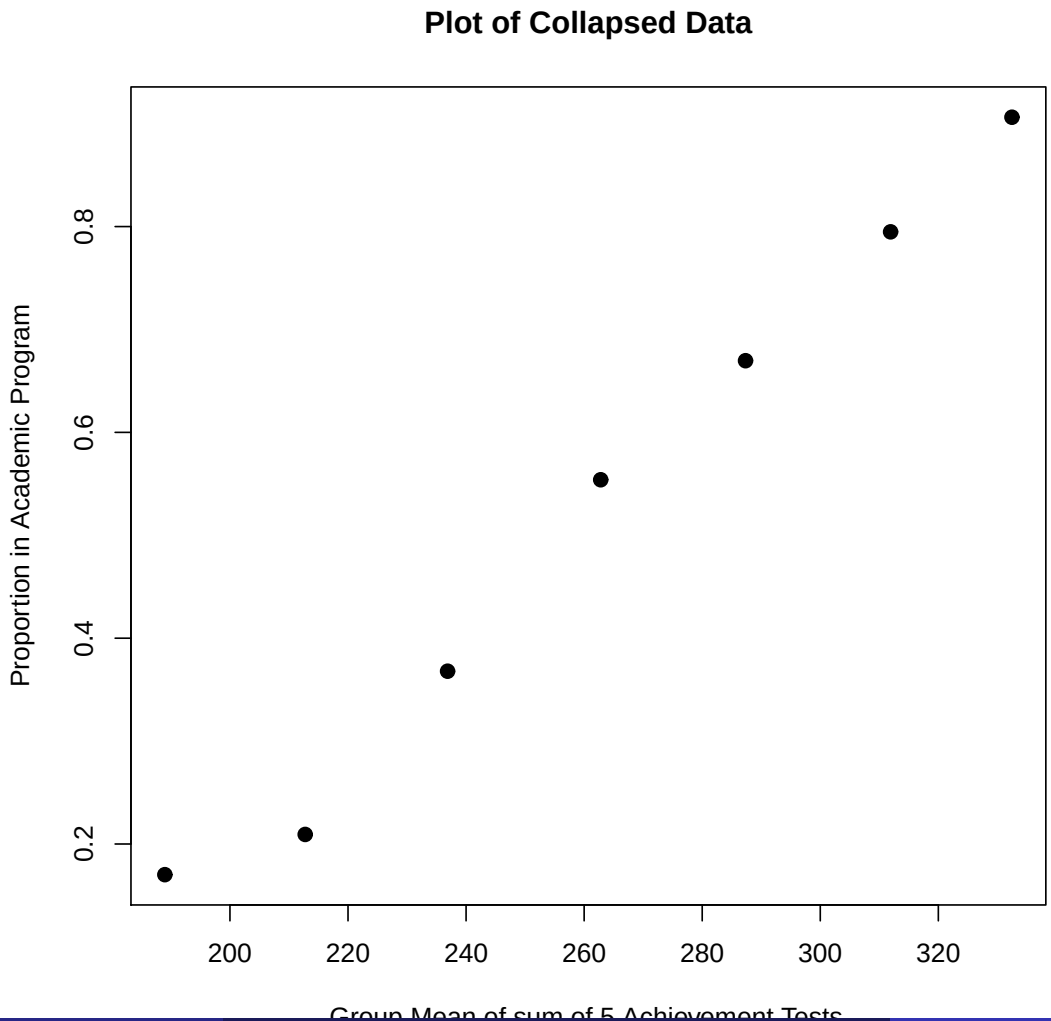
	Sum of Five Achievement Test Scores							
	162–200	201–225	226–250	251–275	276–300	301–325	326–350	
no (0)	39	68	67	62	37	16	3	← Count
	82.98	79.07	63.21	44.60	33.04	20.25	9.68	← Percent
yes (1)	8	18	39	77	75	63	28	
	17.02	20.93	36.79	55.40	66.96	79.75	90.32	
	47	86	106	139	112	79	31	

Is there is relationship? Could it be linear?

Side-By-Side Box Plots



Plot of Collapsed Data



Tests of Statistical Relationship

Statistical Tests of Independence and Linear relationship:

Statistic	df	Value	p–value
Chi-Square	6	119.34	$< .0001$
Likelihood Ratio Chi-Square	6	128.50	$< .0001$
Mantel-Haenszel Chi-Square	1	117.17	$< .0001$

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:

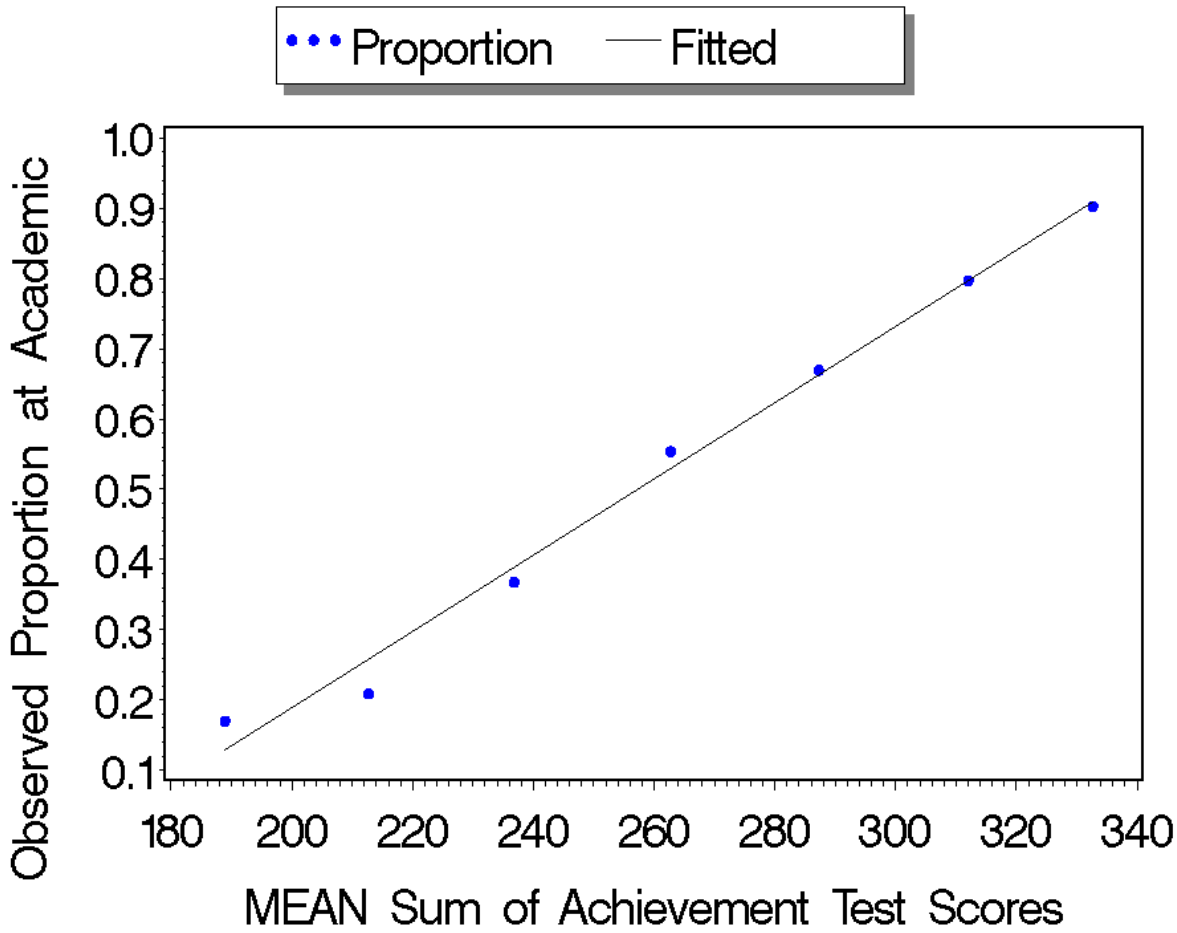
$$\begin{aligned}\hat{\pi}(x) &= \hat{\alpha} + \hat{\beta}x \\ &= -.8987 + .0054x\end{aligned}$$

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

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

	Observed Values			Linear Model	
Mean	Academic	Non-acad.	Proportion	Fitted	Std. Residual
188.972	8	39	0.17	0.13	1.08
212.748	18	68	0.21	0.26	-1.17
236.868	39	67	0.37	0.39	-0.47
262.806	77	62	0.55	0.53	0.61
287.342	75	37	0.67	0.66	0.17
312.080	63	16	0.80	0.80	0.01
332.713	28	3	0.90	0.91	-0.14

Looking at fit of Linear Probability Model



Structural Problem w/ Linear Probability Models

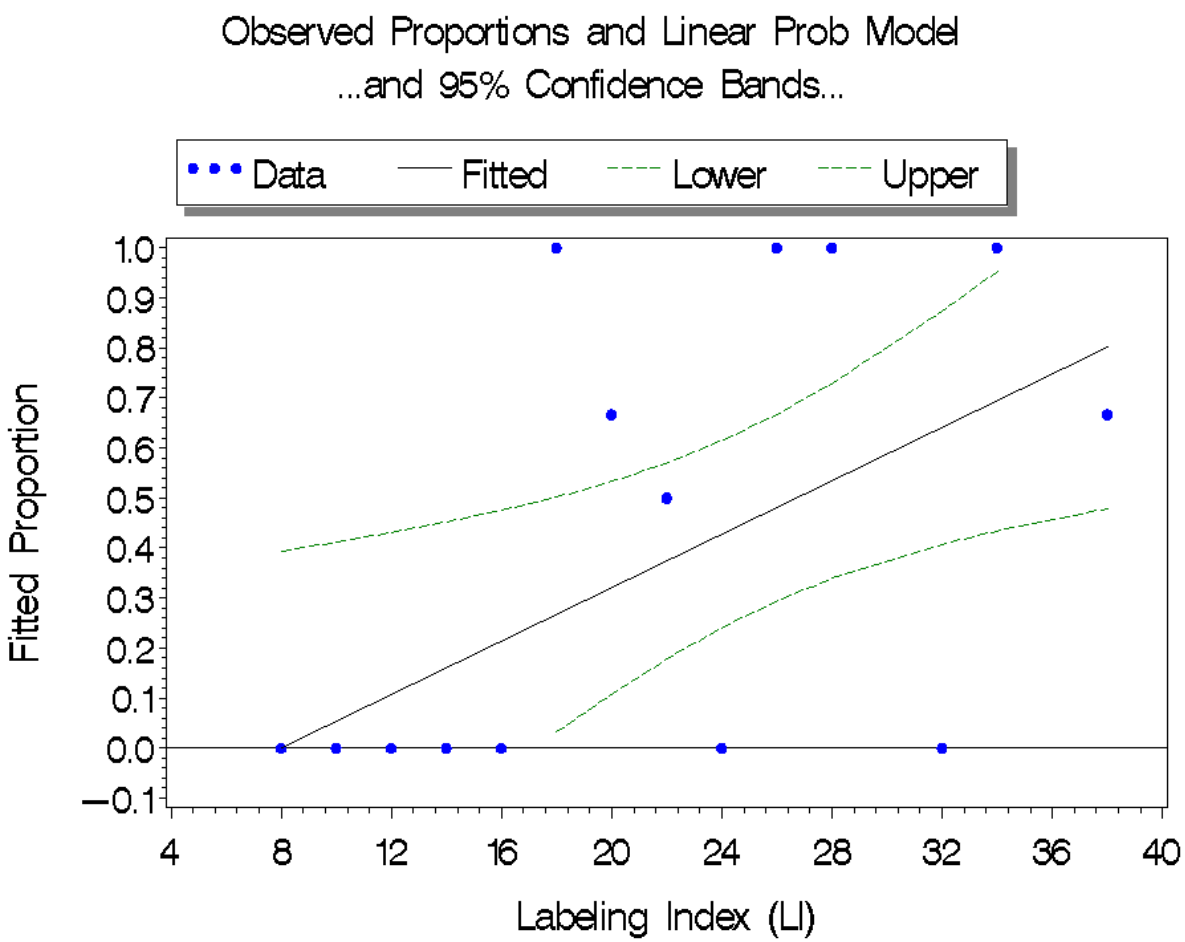
A linear model for $\pi(x)$ can yield predicted values < 0 and/or > 1.

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:

LI	Number of Cases	Number of Remissions	$\hat{\pi}$
8	2	0	$-.003$
10	2	0	.053
12	3	0	.109
14	3	0	.164
$\vdots$	$\vdots$	$\vdots$	$\vdots$
38	3	2	.832

Plot of LI–remission data & Fitted



Plot of LI–remission data & Fitted

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

li	nremit	ncases	LinFit
8	0	2	.00003
10	0	2	.05339
12	0	3	.10676
14	0	3	.16012
16	0	3	.21349
18	1	1	.26685
20	2	3	.32021
22	1	2	.37358
24	0	1	.42694
26	1	1	.48031
28	1	1	.53367
32	0	1	.64040
34	1	1	.69376
38	2	3	.80049

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

	MLE		OLS	
Parameter	Estimate	Std Error	Estimate	Std Error
Intercept, α	-0.2134	0.2768	-0.1613	0.2790
Slope, β	0.0267	0.0104	0.0268	0.0120

Modeling Relationship between $\pi(x)$ and x

First Property a curve should have

- 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.
- 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).

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

  - On 100 item test, we would would expect a larger increase in P(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(item correct) when x goes from 10 to 0 than when x goes from 60 to 50.

Second Property for Curve

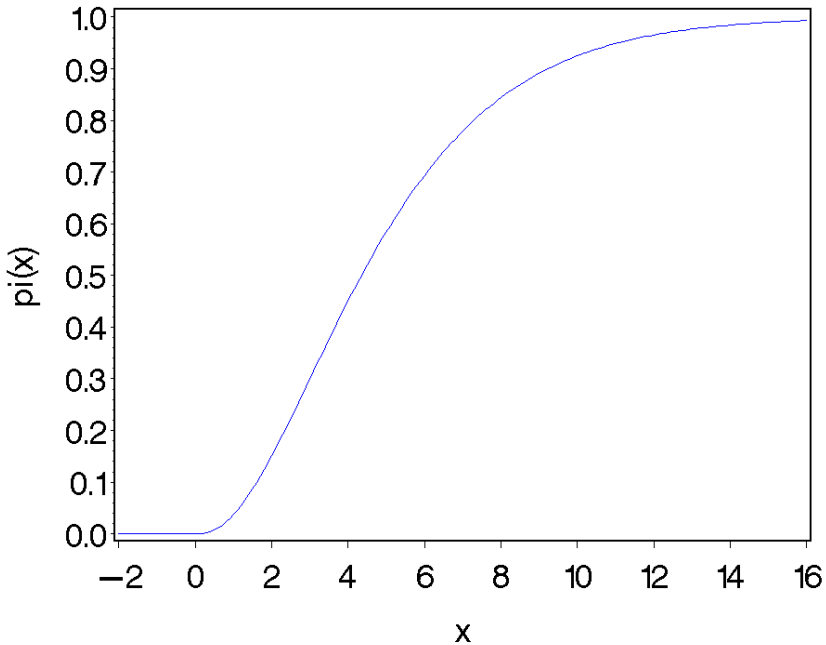
The relationship between $\pi(x)$ and x is usually monotonic such that

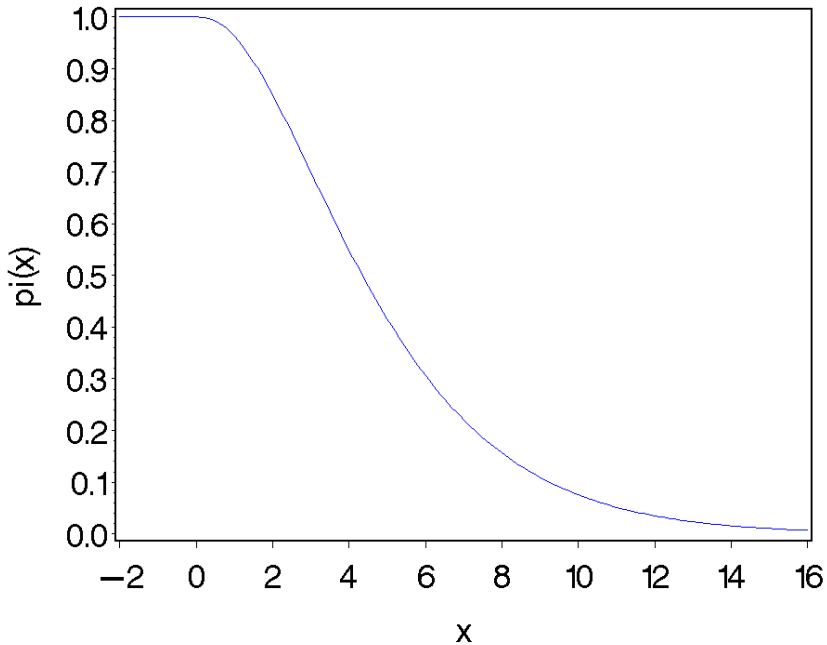
$\pi(x)$ continuously increases as x increases

or

$\pi(x)$ continuously decreases as x increases.

Considering these two properties, an S-shaped curve:





Cumulative Distribution Functions (cdf)

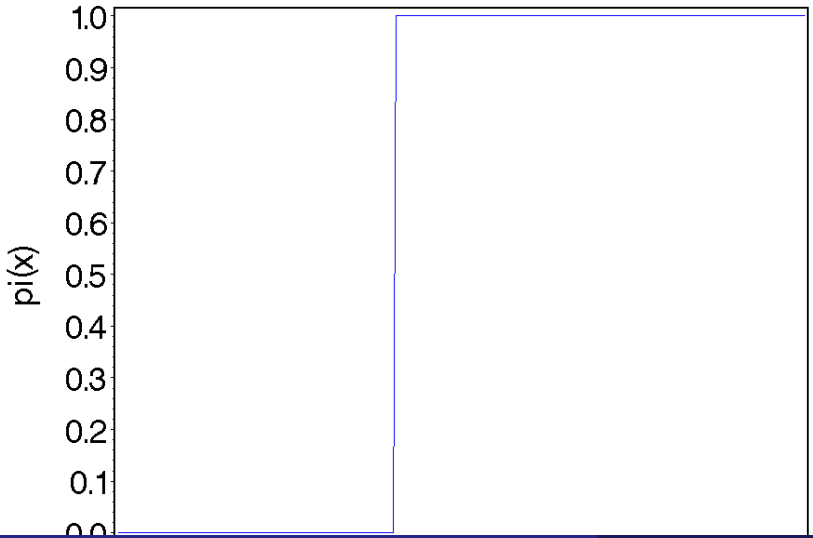
Suppose that

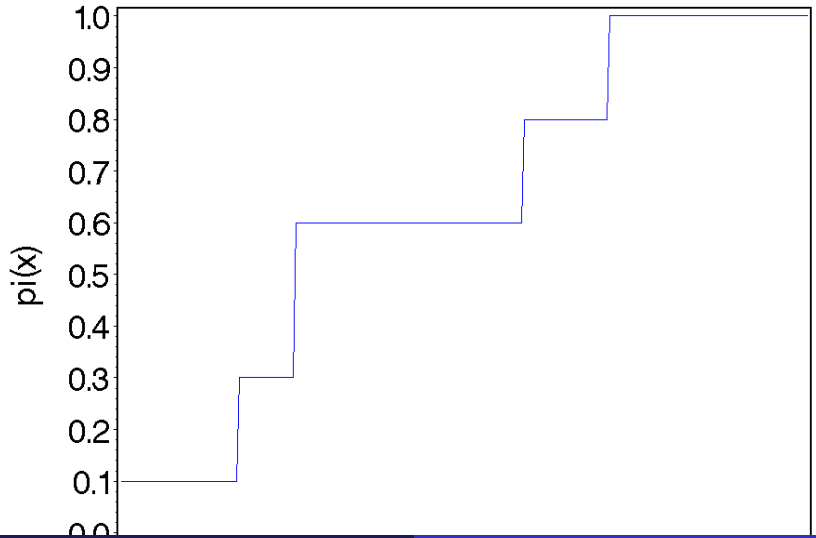
- Z is a random variable
- z is a possible value of Z (e.g., $-\infty < z < \infty$)

A cumulative distribution function for Z is defined as

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

Some examples for discrete Z:



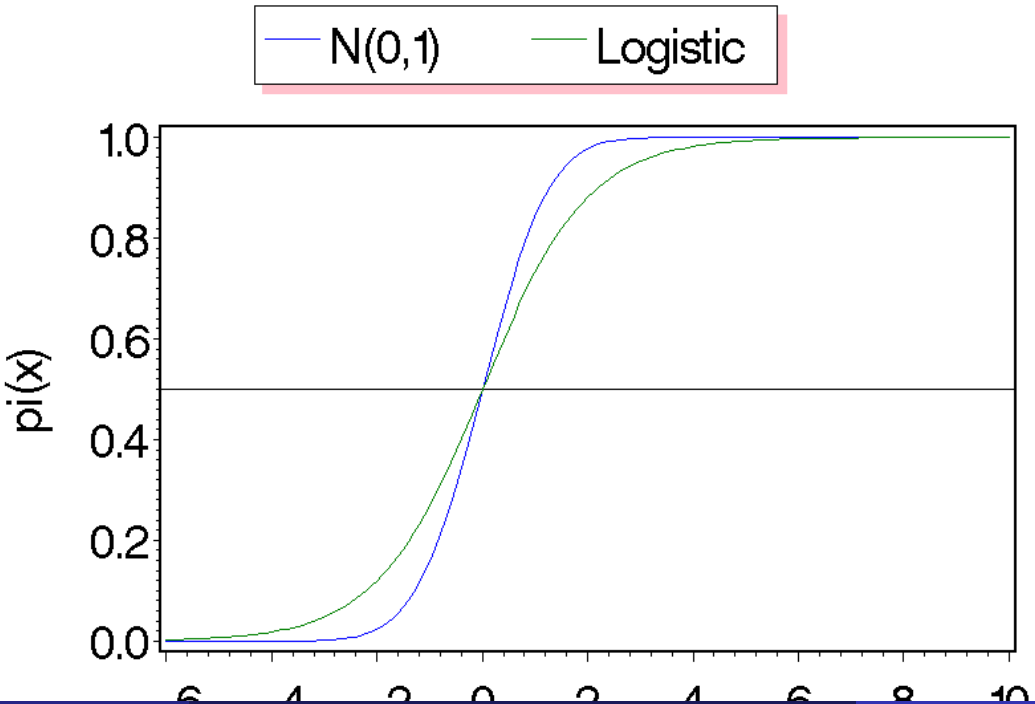


Symmetric pdf's (bell shaped)

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

- μ is a mean (location)
- τ is a scaling parameter
- $-\infty < x < \infty$

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

$$\begin{aligned}
\log\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \log\left(\frac{(\exp(x-\mu)\tau)/(1+\exp((x-\mu)/\tau))}{1/(1+\exp((x-\mu)/\tau))}\right) \\
&= (x-\mu)/\tau \\
&= -\mu/\tau + x/\tau \\
&= \alpha + \beta x
\end{aligned}$$

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(π)" $= \log(\pi/(1-\pi))$.
  - "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 $-\infty <$ logit(π) $< \infty$.
- Systematic component: A linear predictor such as

$$\alpha + \beta x$$

which can be any Real number and yield a π within $(0, 1)$.

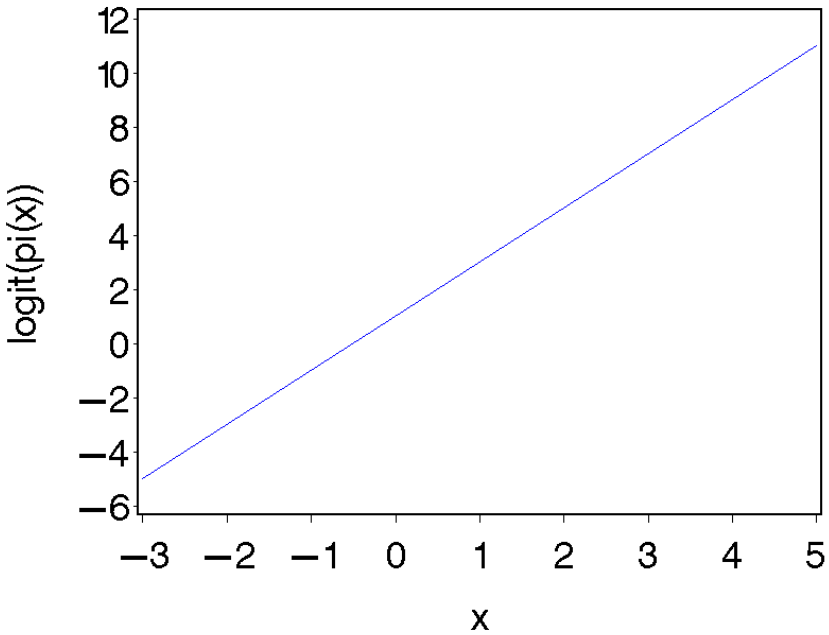
$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$

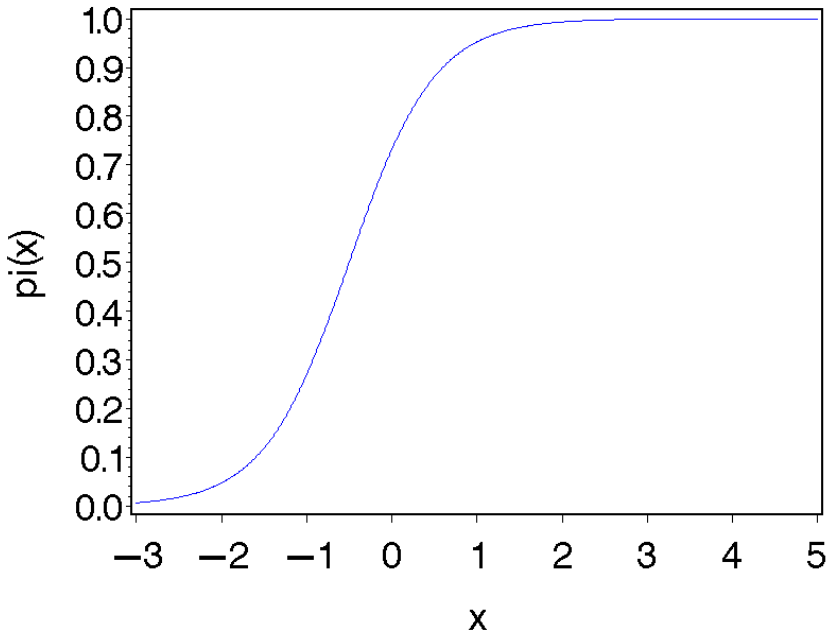
Interpretation of β:

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

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

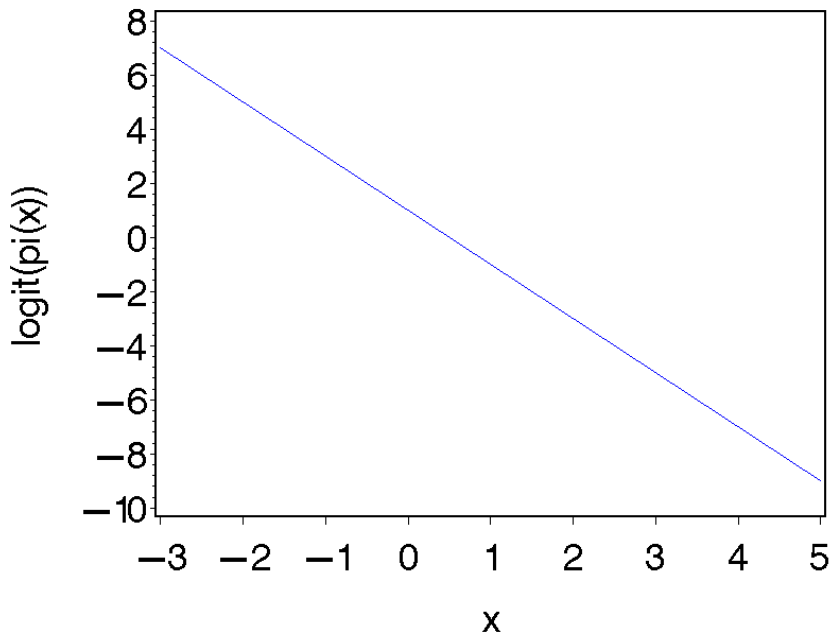
$\alpha = 1$ and $\beta = 2$:

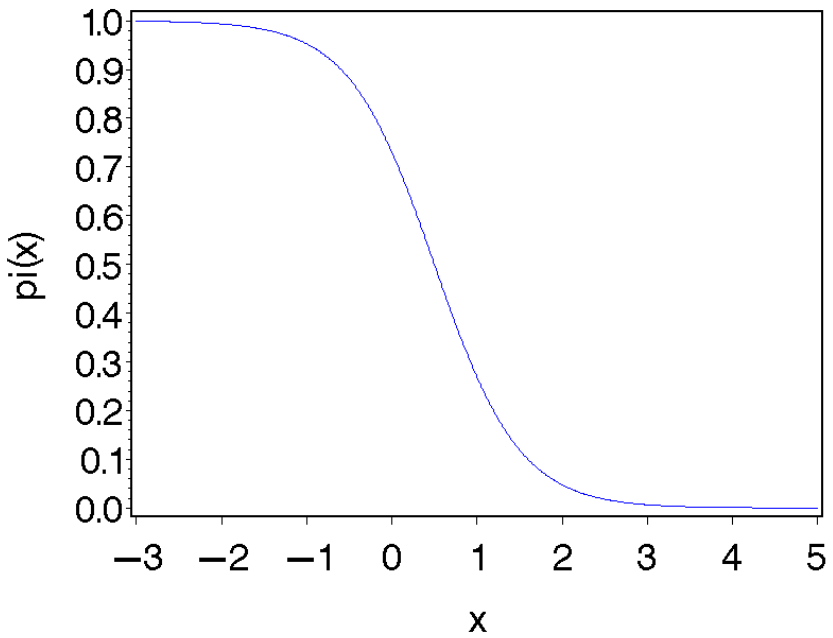




$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 1 - 2x\ \beta$

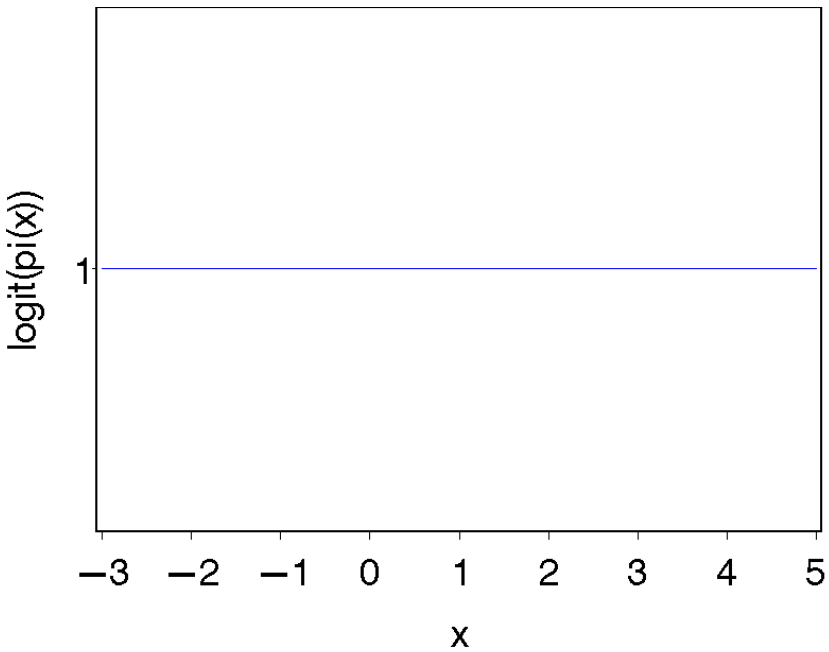
If $\beta < 0$ then $\pi(x)$ decreases as x increases.

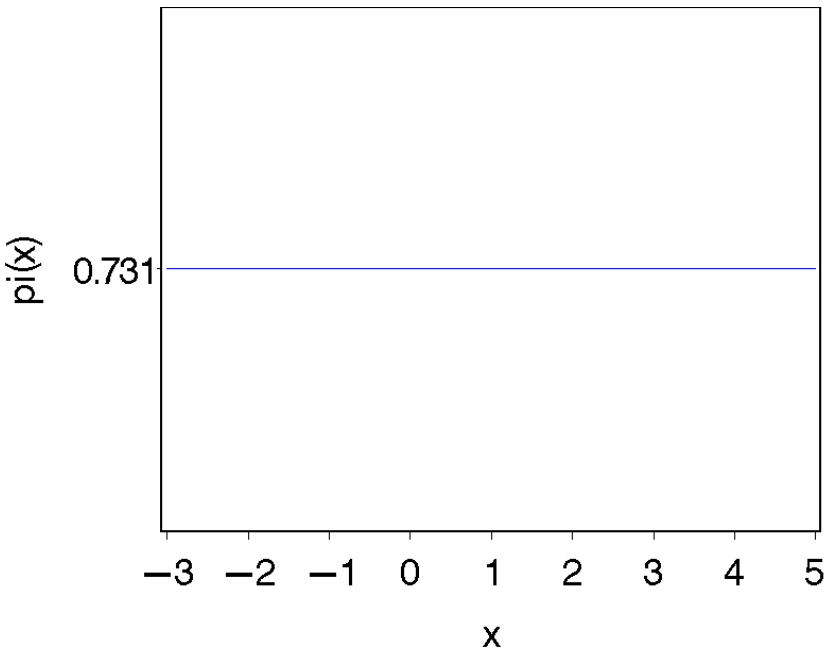




$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 1 + 0x$

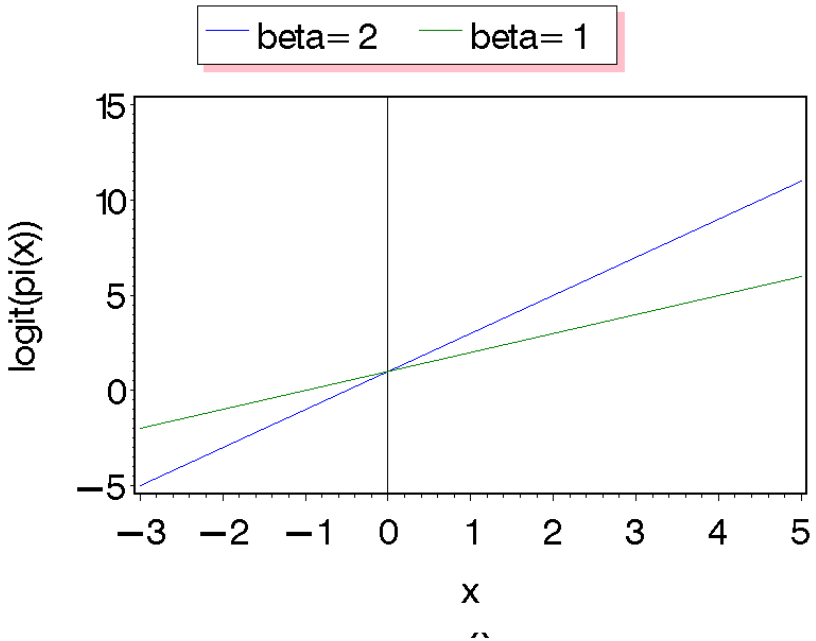
$\beta = 0$ means no relationship between Y and x:

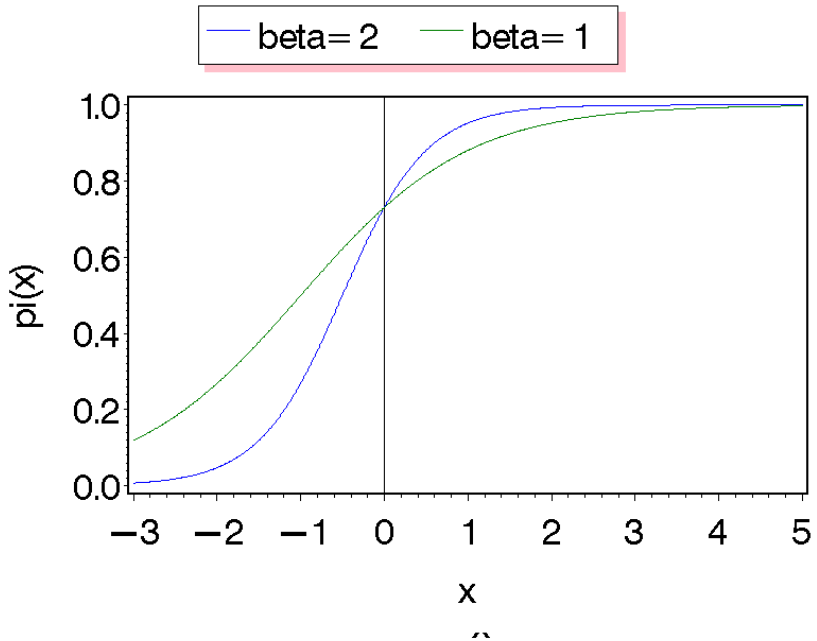




Changing Value of β

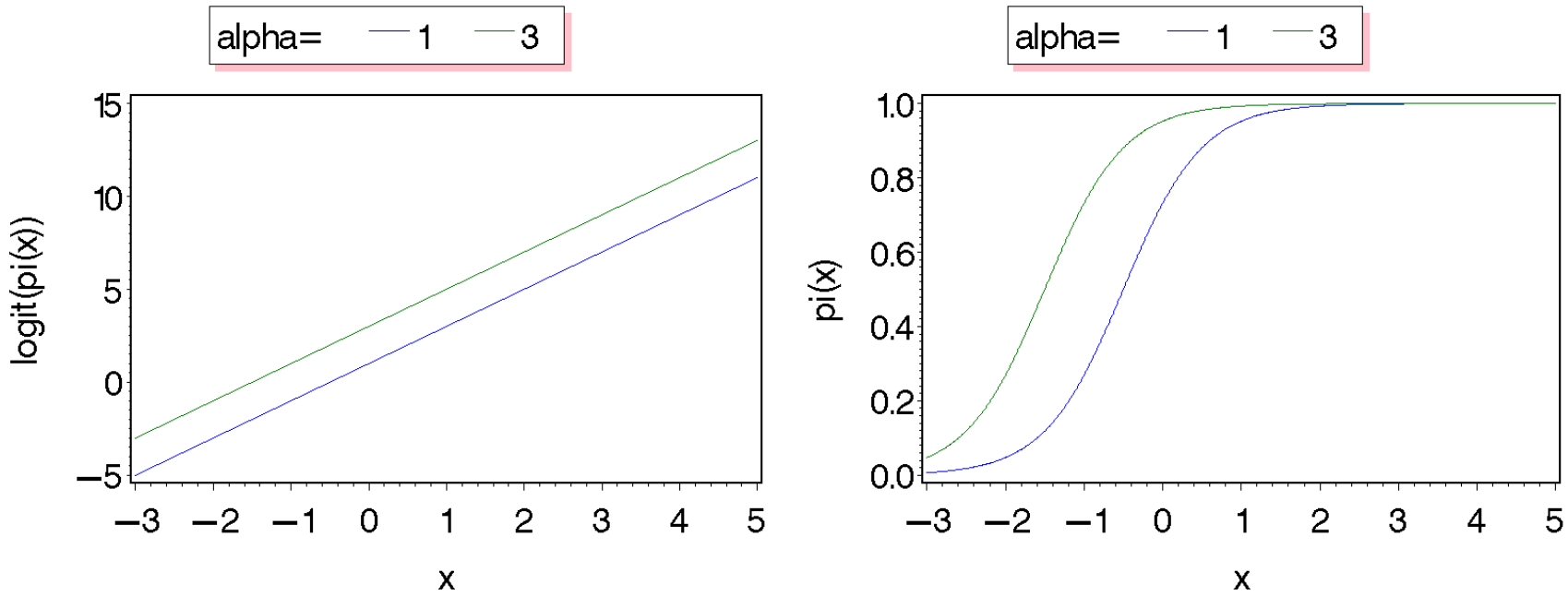
Larger value of β leads to a steeper curve:





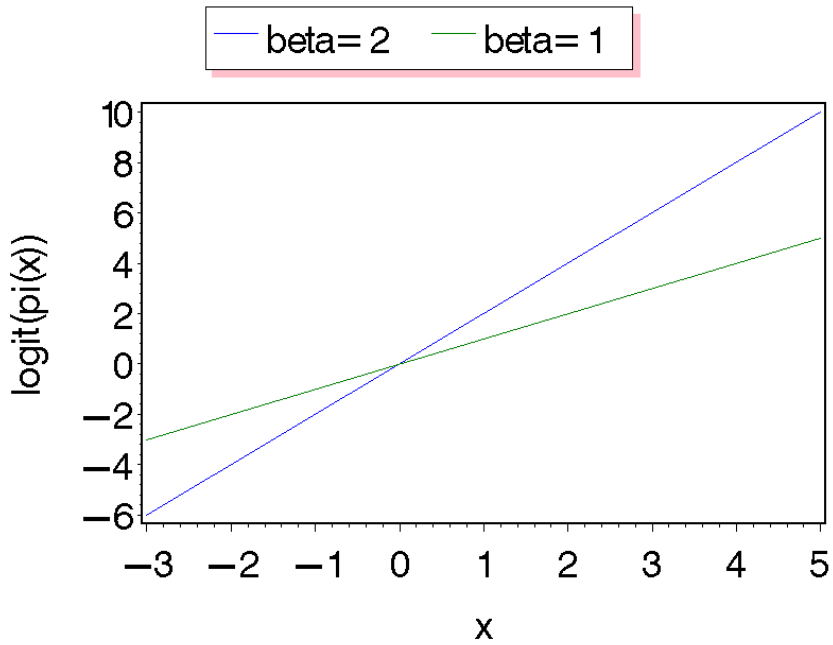
Changing Value of α

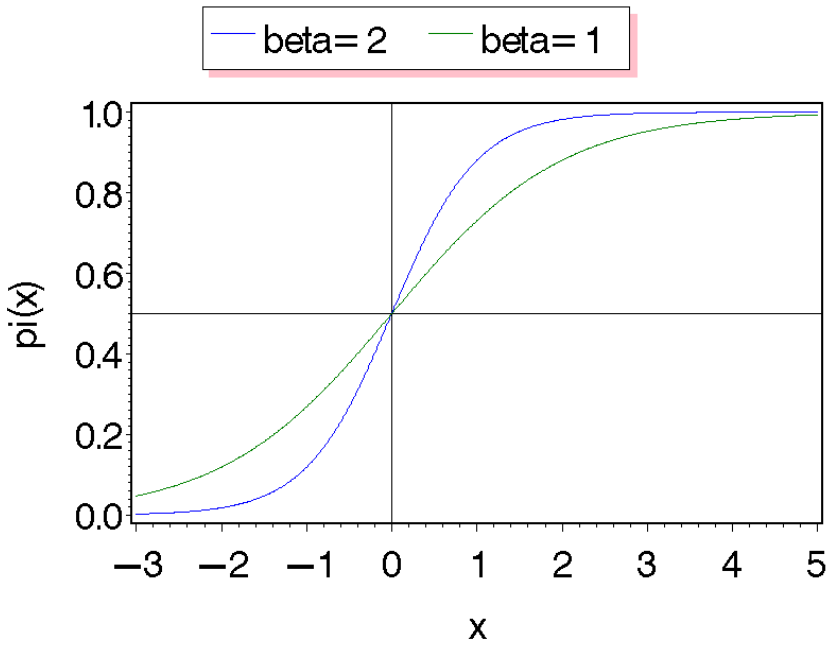
Larger value of α leads to a vertical shift for the logit, but a horizontal shift for the π:



When $\alpha = 0$ with different β'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:

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

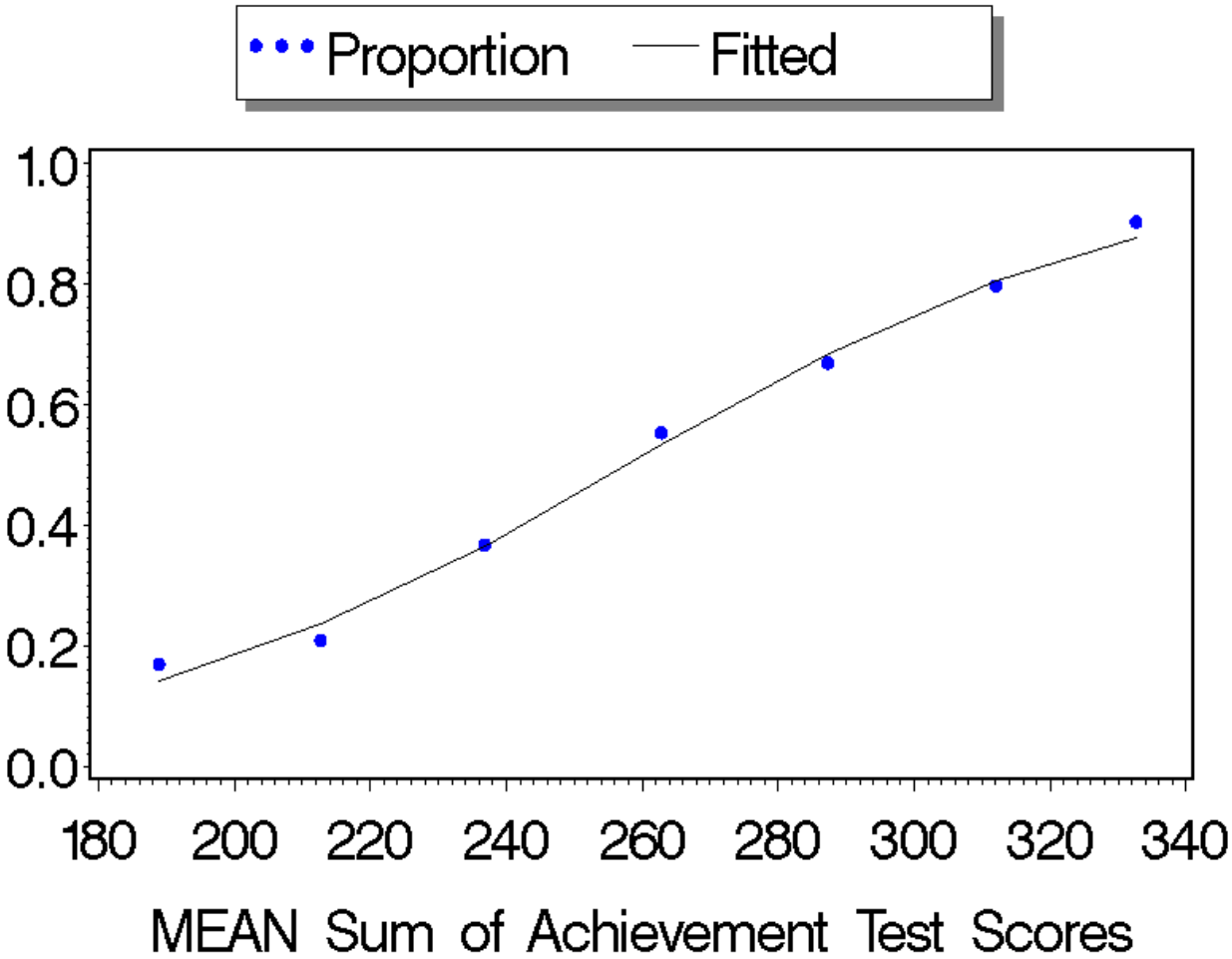
- $\hat{\beta} = .026$ — as achievement scores go up, the probability that a student went to an academic high school program increases. Is this a "large" value?
  - Statistically? The estimated standard error of $\hat{\beta}$ is .0026, so $.026 \pm 2(.0026) \longrightarrow (.021, .031)$.
  - Important? This is not a subjective judgement.

Fitted and Observed Values (grouped data)

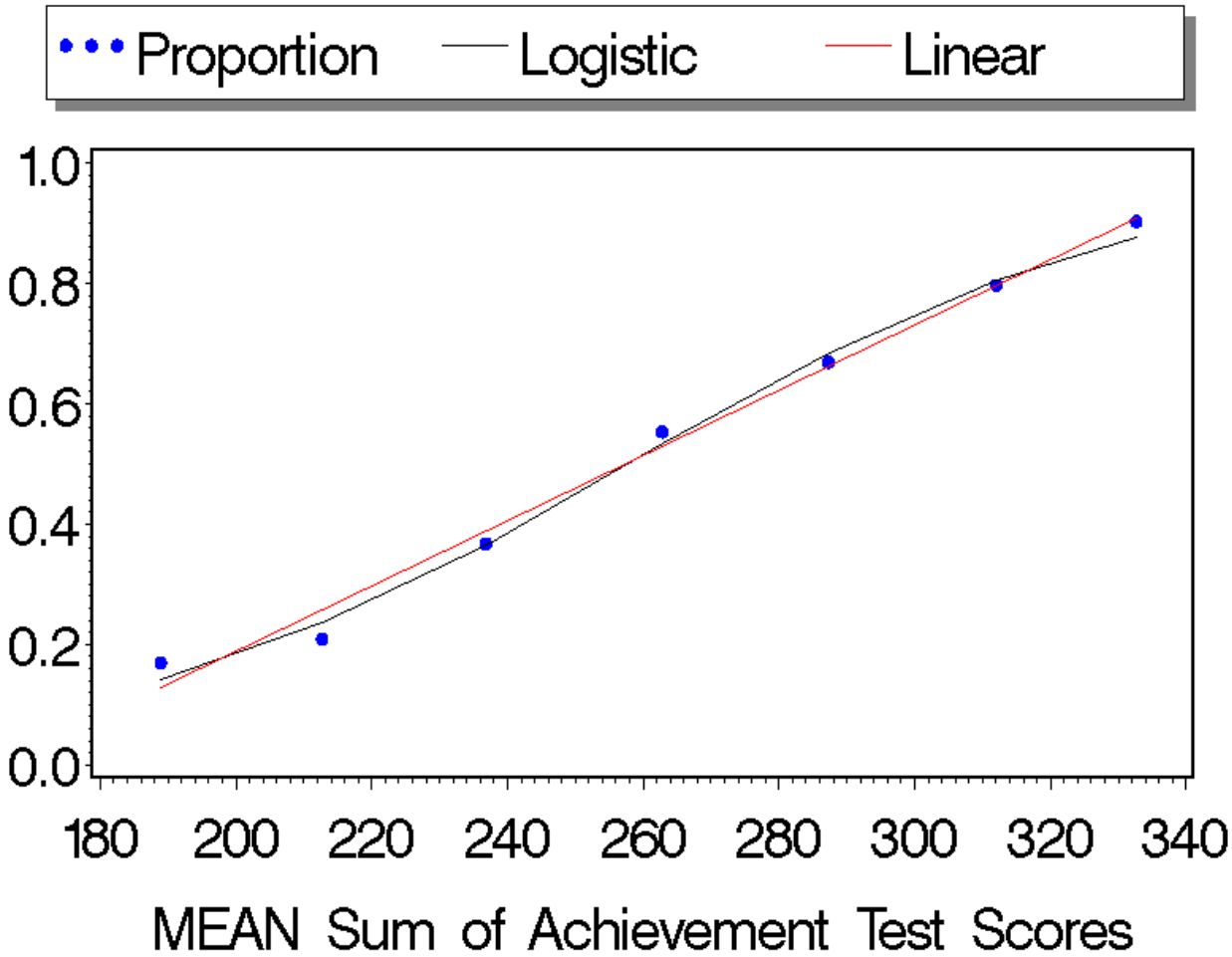
The fitted values $\hat{\pi}(x)$ from logistic regression model:

Achievement Category	Mean Achieve	Observed proportion	Fitted Values (probabilities) Linear	Logistic
162–200	188.97	.17	.13	.14
201–225	212.75	.21	.26	.24
226–250	236.87	.37	.39	.37
251–275	262.81	.55	.53	.53
276–300	287.34	.67	.66	.68
301–325	312.08	.80	.80	.81
326–350	332.71	.90	.91	.88

Plot of Fitted and Observed Values



Comparison with Linear Prob Model



Probit models (Normal *cdf*)

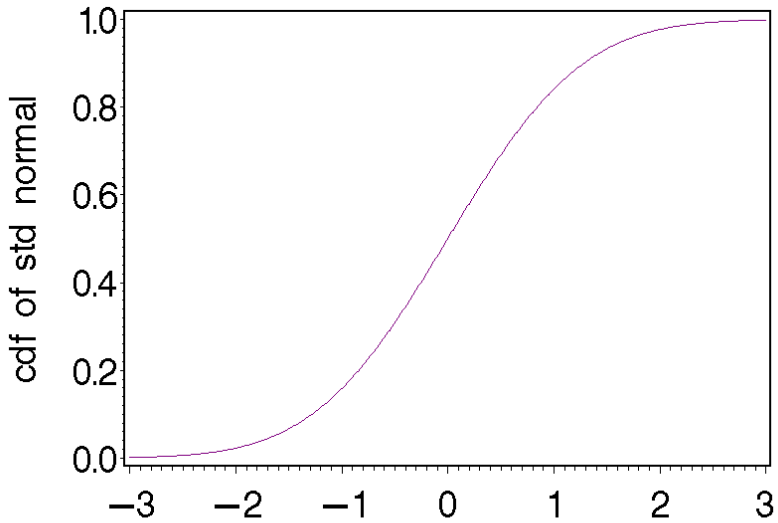
Rather than use the logistic *cdf*, we can use the (standard) Normal distribution.

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,



$$\begin{aligned}
\text{probit}(.025) &= -1.96 \\
\text{probit}(.050) &= -1.64 \\
\text{probit}(.500) &= 0.00 \\
\text{probit}(.950) &= 1.64 \\
\text{probit}(.975) &= 1.96
\end{aligned}$$

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.

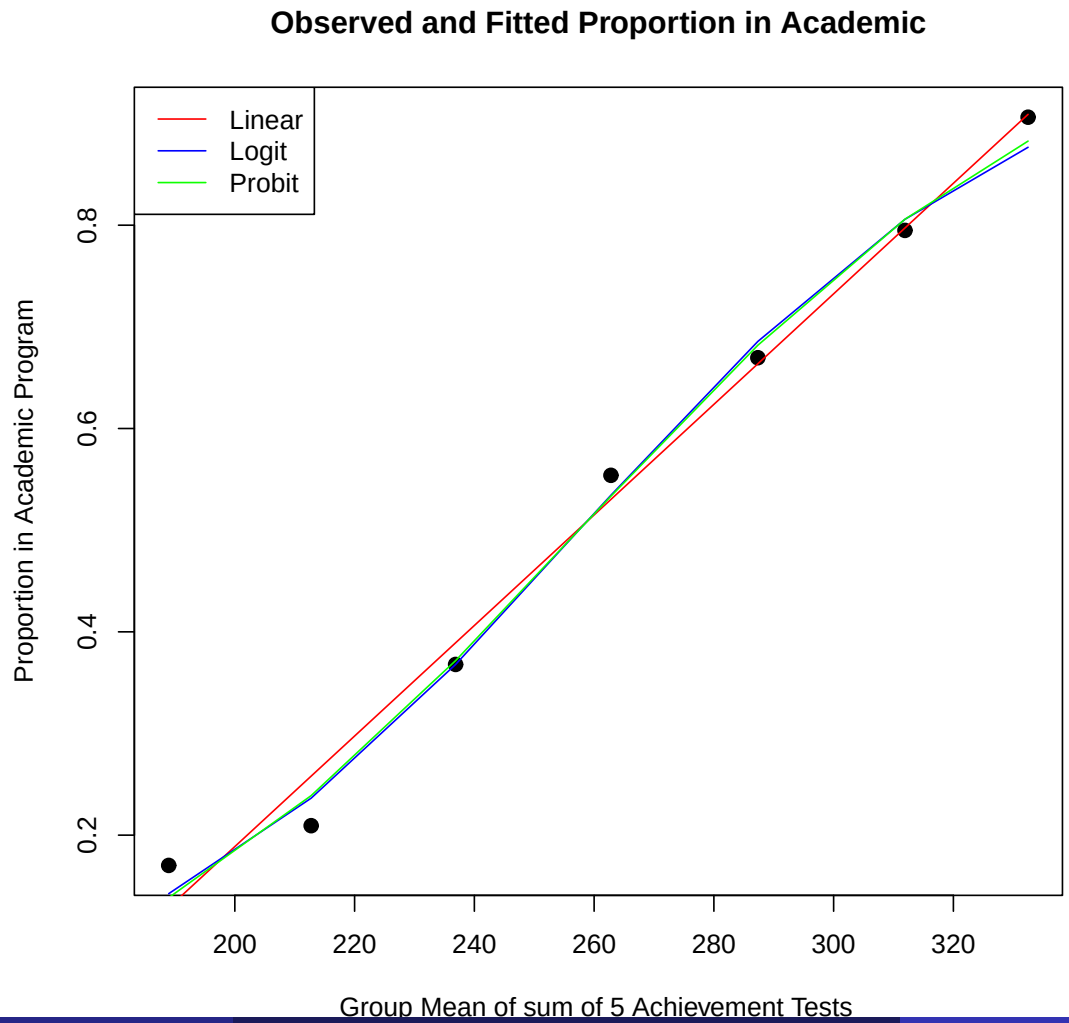
Example of Probit model: HSB

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

- 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 $.0158 \pm 2(.0015) \longrightarrow (0.013, 0.019)$.
- Fitted values are *extremely* close to those from the logit model.

Category	x	p	Linear	Logistic	Probit
162–200	188.97	.17	.13	.14	.14
201–225	212.75	.21	.26	.24	.24
226–250	236.87	.37	.39	.37	.37
251–275	262.81	.55	.53	.53	.53
276–300	287.34	.67	.66	.68	.68
301–325	312.08	.80	.80	.81	.81
326–350	332.71	.90	.91	.88	.88

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 $\sigma = 1/|\beta|$.

Logit & Probit Models (continued)

- For the HSB data, the probit model corresponds to a normal *cdf* with mean $= -(-4.0828)/.0158 = 258.41$ and standard deviation $= 1/.0158 = 63.29$.
- 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,

$$\begin{aligned} \hat{\pi}(258.41) &= \text{logit}^{-1}(-6.741 + .0262(258.41)) \\ &= \frac{1}{1 + \exp^{-(-6.741 + .0262(258.41))}} \\ &= .5 \end{aligned}$$

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,

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

For a probit model for binary data,

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

R glm

Note that parameter estimates differ slightly when dealing with groups hsb data and se's are quite different. Not apparent differences when data weren't grouped.

See R script of data input, basic statistics, and graphing

```
# Fit to uncollapsed data
linear.all ← glm(program ∼ achieve, data=hsb,
                 family=binomial(link="identity"))
summary(linear.all)


# Fit to collapsed data
hsb$p.acd ← hsb$program/hsb$ncase
linear.all ← glm(p.acd ∼ achieve, data=hsb, weights=ncase,
                 family=binomial(link="identity"))
```

R glm: logit

```
# Fit to uncollapsed data
logit.all ← glm(program ∼ achieve, data=hsb,
                family=binomial(link="logit"))
summary(logit.all)

# Fit to collapsed data
logit.all ← glm(p.acd ∼ achieve, data=hsb, weights=ncase,
                family=binomial(link="logit"))
summary(logit.all)
```

R glm

The probit:

```
# Fit to uncollapsed data
probit.all ← glm(program ∼ achieve, data=hsb,
                 family=binomial(link="probit"))
summary(probit.all)

# Fit to collapsed data
probit.all ← glm(p.acd ∼ achieve, data=hsb, weights=ncase
                 family=binomial(link="probit"))
summary(probit.all)
```

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: $Y = -0.2134 + 0.0267(\text{LI})$

  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: $\text{logit}(Y) = -3.7771 + 0.1449(\text{LI})$
- Probit model: $\text{probit}(Y) = -2.3178 + 0.0878(\text{LI})$

Example # 2: R Cancer remission data

```
# Linear probability model
linear.model ← glm(p.remit ∼ li, data=li, weights=ncases,
                   family=binomial("identity"))
summary(linear.model)
```

Error: no valid set of coefficients has been found: please supply starting values

Example # 2: R Cancer remission data

```
# Logit probability model
logit.model ← glm( ∼ li, data=li, weights=ncases,
            family=binomial("logit"))
summary(logit.model)

# logit model with quadratic term
li$lisq ← li$li**2
new ← glm(p.remit∼ li + lisq, data=li, weights=ncases,
            family=binomial("logit"))
summary(new)

# Probit probability model
probit.model ←glm(p.remit ∼ li, data=li, weights=ncases,
            family=binomial("probit"))
summary(probit.model)
```

Loess and Jitter

For code, see programs on web-site
In R:

- `plot(hsb$achieve,jitter(hsb$program,0.1) )`
- ```
  # fit loess to data
  p ← li$nremit/li$ncases
  lw1 ← loess(p ~ li$li)
  ```

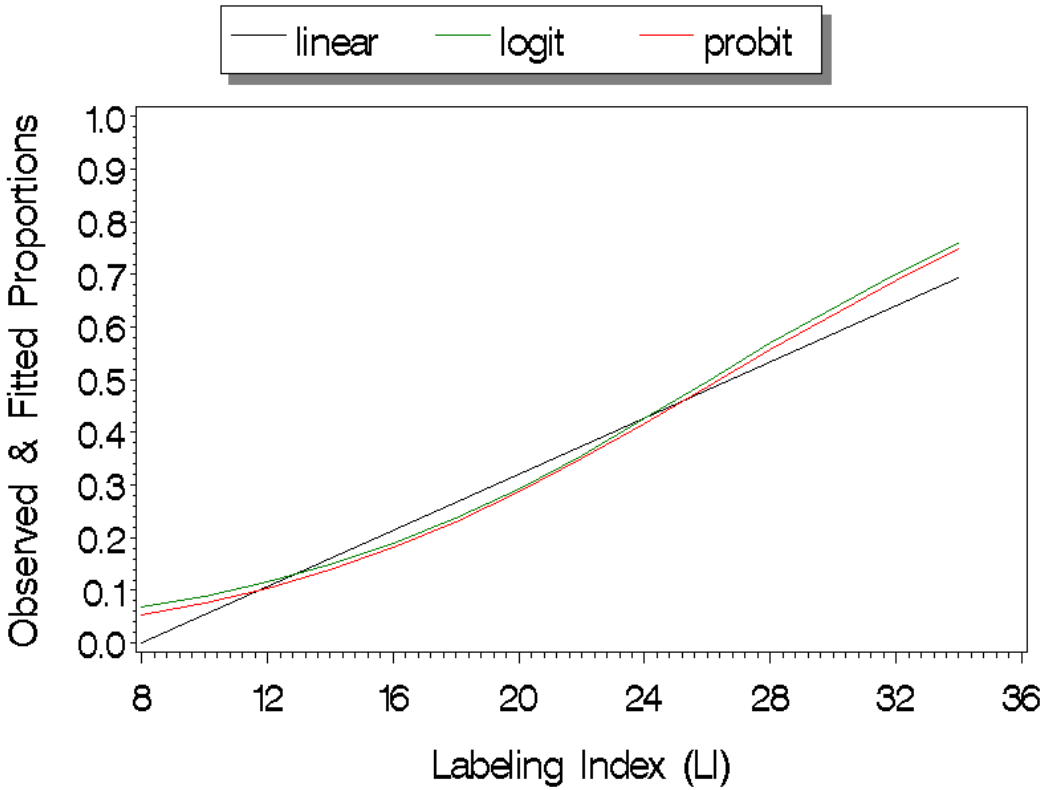
In SAS:

- PROC LOESS
- jitter option to scatter in sgplot (see HSB code for simple example)
  ```

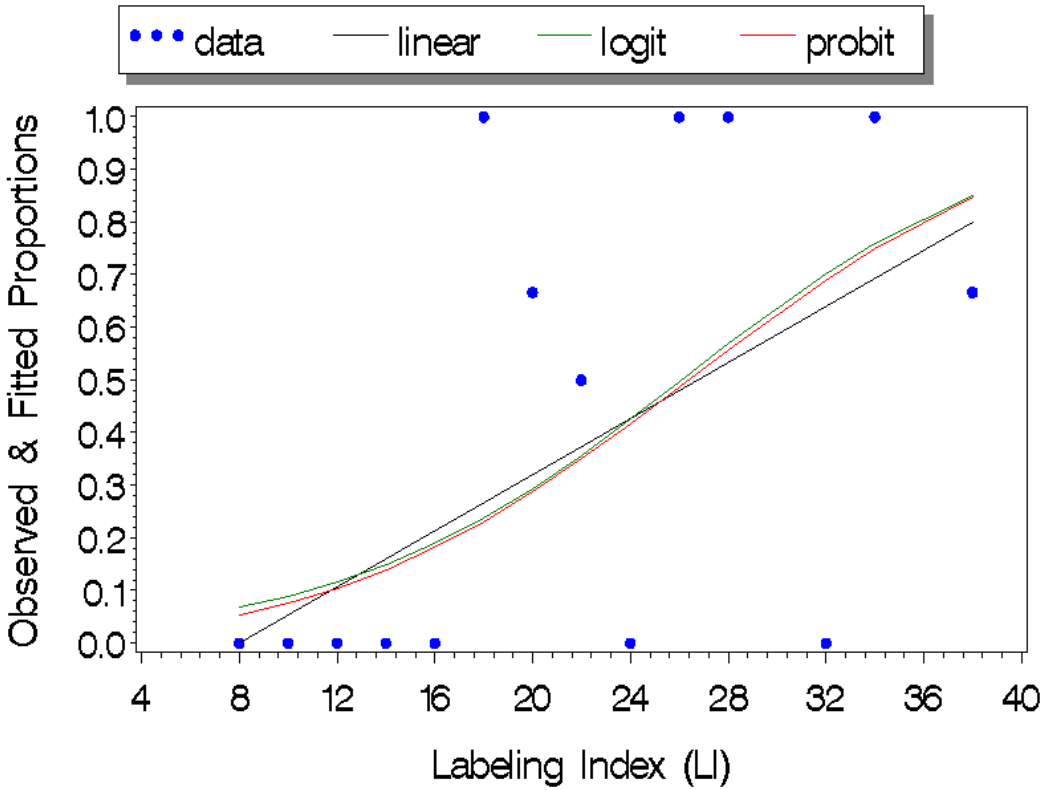
What is LOESS?

From https://www.statsdirect.com/help/nonparametric_methods/loess.htm: "LOESS Curve Fitting (Local Polynomial Regression). This is a method for fitting a smooth curve between two variables, or fitting a smooth surface between an outcome and up to four predictor variables. . . This is a nonparametric method because the linearity assumptions of conventional regression methods have been relaxed. Instead of estimating parameters like m and c in $y = mx + c$, a nonparametric regression focuses on the fitted curve. The fitted points and their standard errors represent are estimated with respect to the whole curve rather than a particular estimate. So, the overall uncertainty is measured as how well the estimated curve fits the population curve."

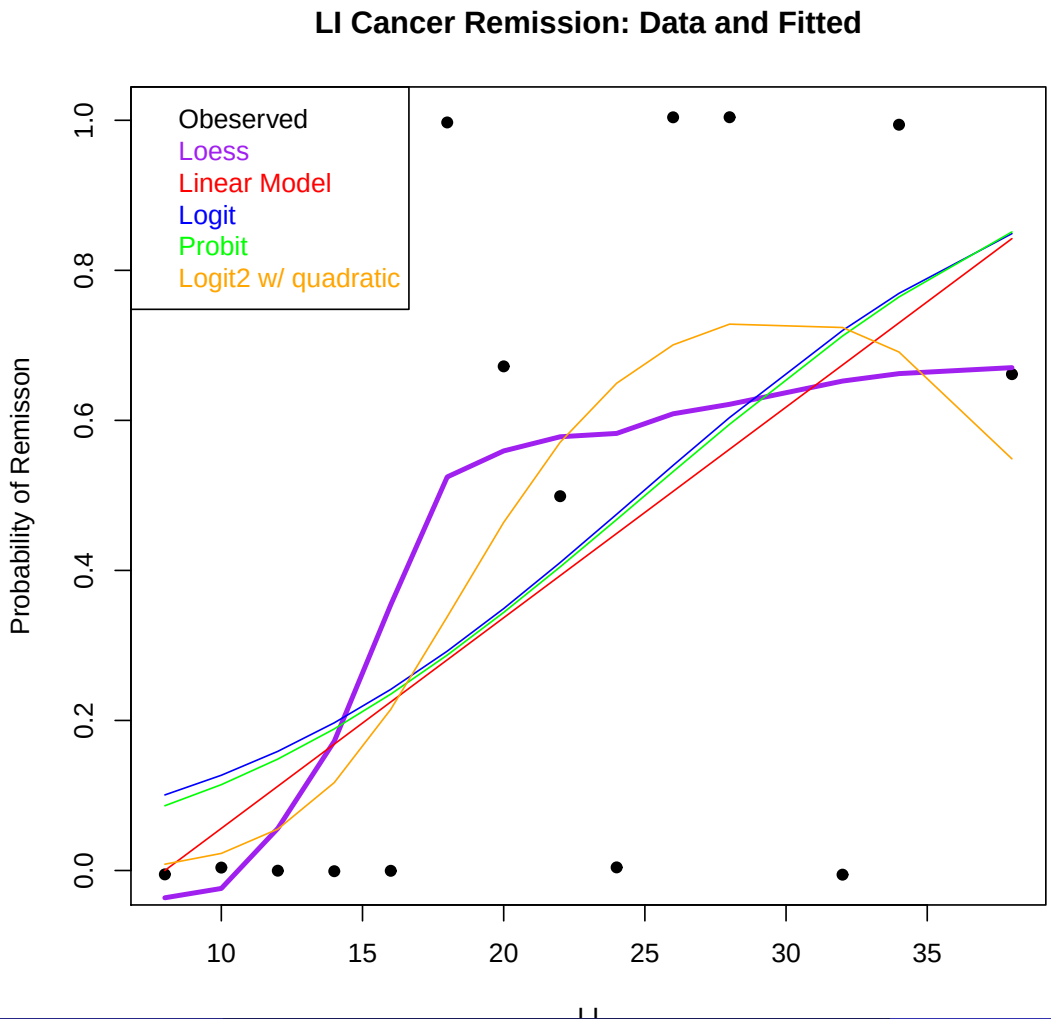
Comparison of Model Fitted Values



Comparison of Model Fitted Values & Data



R: Comparison of Model Fitted Values & Data



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 used a normal cdf *with* $\mu = 5$ *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.