Motivating Binary Response Regression Models

Kevin Quinn
Assistant Professor
Department of Political Science and
The Center for Statistics and the Social Sciences
Box 354322, Padelford Hall
University of Washington
Seattle, WA 98195-4322

October 5, 2000

1 Basic Parameterization Issues

- Consider the dataset in table 1.
- This dataset is composed of three variables collected on 30 students. The 3 variables are: an indicator as to whether the student passed a statistics class (0=fail, 1=pass), the student's grade in a prerequisite probability course (0=F, 1=D, 2=C, 3=B, 4=A), and the student's math SAT score.
- We'd like to model how the probability of passing the statistics course varies over the past grades and math SAT scores.
- How might we do this?
- So far we've only considered simple models for Bernoulli trials in which the Bernoulli parameter π is constant over all trials
- This implies the sampling density is

$$f(\mathbf{y}|\pi) \propto \prod_{i=1}^{30} \pi^{y_i}(1-\pi)^{1-y_i}$$

  where $\mathbf{y}$ is the vector of passing grade indicators for all 30 students.
- π doesn't vary across individuals
- We have reason to believe that π varies over individuals– We need a better model
- How can we parameterize our sampling density to allow π to vary across individuals?
- One possibility is to assume that π_i for $i = 1, 2, \ldots 30$ are all free parameters
  - Problem: as many parameters as data points
  - Perfect fit every time
  - Is this really explaining anything?
  - How do we interpolate and/or extrapolate to make predictions?
- We need to put some structure on the problem.
- Lets assume that π_i varies as a function of our covariates for the ith individual.
- The basic problem is to find a function that takes combinations of prerequisite grades and math SAT scores and returns a probability of passing.
- Figure 1 plots the raw data.

Passing Grade	Grade in Prereq.	Math SAT
0	3	525
0	2	533
1	3	545
0	4	582
1	2	581
1	1	576
1	3	572
1	4	609
1	2	559
1	1	543
1	3	576
1	4	525
1	0	574
1	1	582
1	2	574
0	3	471
1	3	595
0	2	557
0	4	557
1	4	584
1	3	599
0	2	517
1	4	649
1	2	584
0	1	463
1	3	591
0	2	488
1	3	563
1	3	553
1	4	549

Table 1: Hypothetical grades for a class of statistics students. From Johnson and Albert (1999, p. 77).

- What's the functional form?
- Might want to allow each combination of prerequisite grade and math SAT score to have its own probability of passing.
  - No real assumptions about functional form
  - Would require a lot of data ($5 \times 600 = 3000$ combinations of grades and math SAT scores)
- need more structure
- How about the simple linear model $\pi_i = \beta_0 + \text{grade}_i\beta_1 + \text{MSAT}_i\beta_2$?
  - This obviously won't work because it can result in values of π outside of the $[0, 1]$ interval.
  - However, the linear model does have one nice aspect: the effects are additive making interpretation relatively easy
- What if we transform the linear predictor so that it stays within $[0, 1]$?
- We need to find a function that maps the real number line onto $(0, 1)$.
- It would be nice if this function were also monotone so that we could interpret the signs of coefficients
- Cumulative distribution functions are one set of functions that have these properties
- If $F(\cdot)$ is some cumulative distribution function then we can model the probability of passing as:

$$\pi_i = F(\beta_0 + \text{grade}_i\beta_1 + \text{MSAT}_i\beta_2)$$

  - π_i is always in $[0, 1]$
  - a negative coefficient implies that the probability of success is decreasing in that variable
  - a positive coefficient implies that the probability of success is increasing in that variable
- Once we estimate β_0, β_1, and β_2 we can plot the probability of passing as function of prerequisite grades and math SAT scores.
- Figure 2 displays such a plot
- One commonly used cumulative distribution functions are the standard logistic distribution function

$$F(x) = \frac{e^x}{1 + e^x}$$

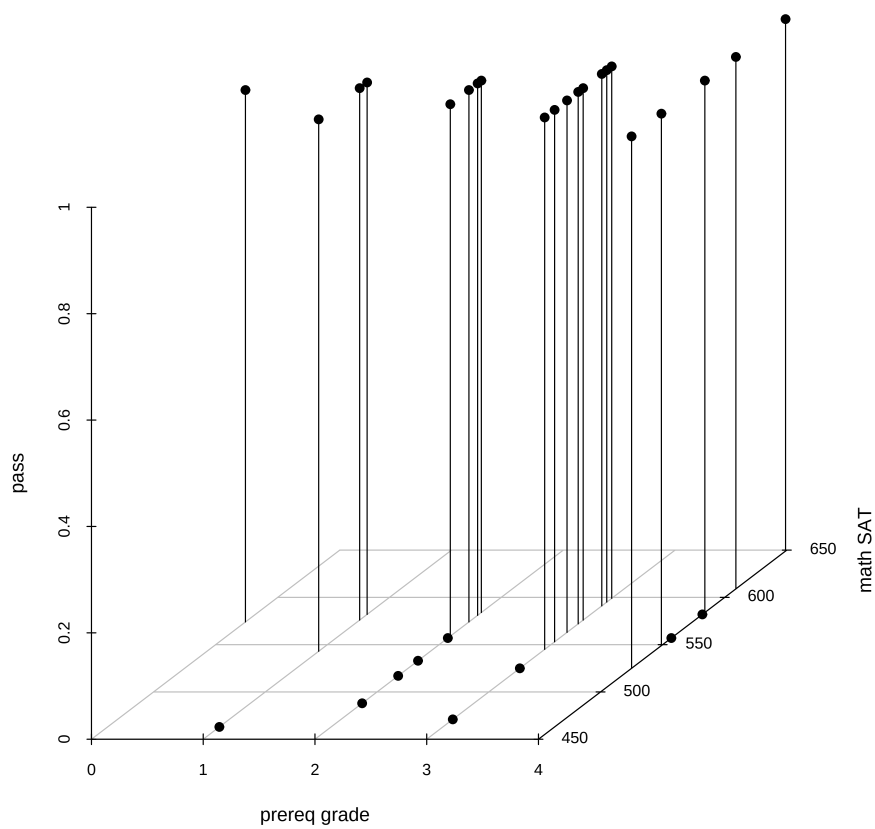

Figure 1: 3-D scatterplot of the statistics course data.

- Using the standard logistic cdf results in the ***logistic regression*** or ***logit*** model

$$\pi_i = F(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})}$$

$$p(\mathbf{y}|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

$$p(\mathbf{y}|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} F(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1 - F(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}$$

- Another is the standard normal cumulative distribution function

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)dz$$

- Using the standard normal cdf results in the ***probit*** model

$$\pi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$$

$$p(\mathbf{y}|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

$$p(\mathbf{y}|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}$$

2 Latent Variable/Random Utility Motivations

2.1 Simple Latent Variable Motivation

- Above we transformed our linear predictor using a cumulative distribution function largely for expediency
- Can we motivate such a transformation from first principles?
- Yes
- Suppose that the binary response y_i is the result of dichotomizing an unobserved (latent) continuous response y_i^*

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

where ϵ_i is drawn from a symmetric density $f(\cdot)$ with mean 0 and fixed variance

- While we can't directly observe y_i^* we do know how it is distributed given any values of $\mathbf{x}_i'$ and $\boldsymbol{\beta}$.
  - It has mean $\mathbf{x}_i'\boldsymbol{\beta}$ and fixed variance
- This implies that

$$\Pr(y_i = 1|\boldsymbol{\beta}) = \Pr(y_i^* > 0|\boldsymbol{\beta}) = \int_0^\infty f(y_i^*|\mathbf{x}_i'\boldsymbol{\beta})dy_i^*$$

where $f(\cdot|\mathbf{x}_i'\boldsymbol{\beta})$ is the density function of latent continuous variable with mean $\mathbf{x}_i'\boldsymbol{\beta}$

- Does $\int_0^\infty f(y_i^*|\mathbf{x}_i'\boldsymbol{\beta})dy_i^*$ equal $F(\mathbf{x}_i'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta}} f(y_i^*|0)dy_i^*$? In other words is this model equivalent to the logit and probit models discussed previously?
- Yes– Couple ways to show this
  - Geometrically– draw densities and areas under the curve
  - Algebraically–

$$\begin{aligned}
\Pr(y_i^* > 0|\boldsymbol{\beta}) &= \Pr(\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i > 0) \\
&= \Pr(\epsilon_i > -\mathbf{x}_i'\boldsymbol{\beta}) \\
\text{Assuming that } f(\cdot) \text{ is symmetric} & \\
&= \Pr(\epsilon_i < \mathbf{x}_i'\boldsymbol{\beta}) \\
&= F(\mathbf{x}_i'\boldsymbol{\beta})
\end{aligned}$$

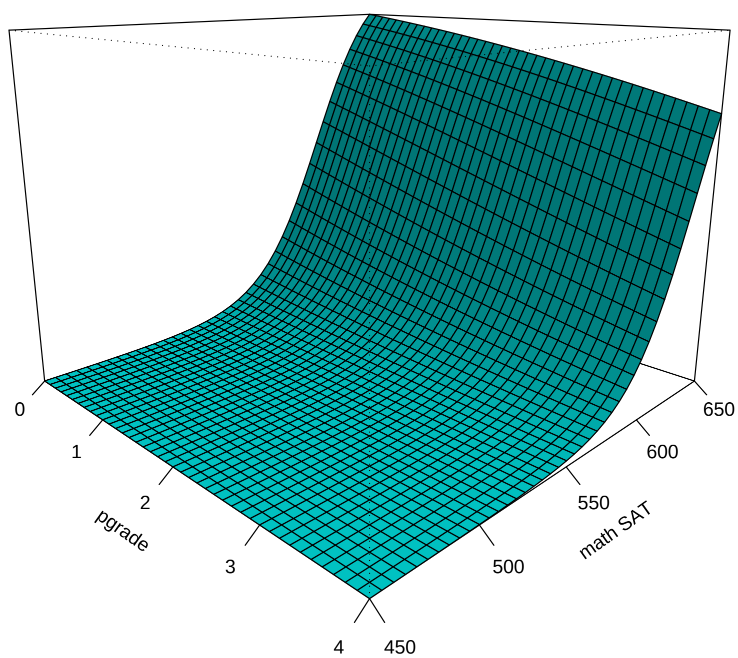

Figure 2: Logistic regression surface fitted to the statistics course data.

2.1.1 Identification Issues

- Note that the threshold that determines what values of y_i^* produce values of $y_i = 1$ is not identified and must be fixed at some constant value if a constant term is included in the model.
- fixing the threshold at 0 is an innocent normalization
  - Changing the threshold from 0 to c and adding c to the constant term doesn't change the values of π_i and consequently doesn't change the value of the sampling density
- Similarly the variance of the underlying distribution of y_i^* is not identified
- fixing the variance at 1 (in the case of the probit model) is an innocent normalization
  - if the standard deviation of the latent variable increases by a factor c we can just multiply all of our coefficients by c and the the values of π_i will remain the same leading to the same sampling density

2.2 Random Utility Motivation

- Oftentimes dichotomous response variables indicate observed choices made by individuals (voting, entering the labor force, etc)
- In these situations it is often natural to think of a binary response regression model arising from a ***random utility model***
- Let's look at the example of an individual i's decision to vote ($y_i = 1$) in an election or abstain ($y_i = 0$)
- Each individual attaches some utility to voting and to abstaining.
- The option with the higher utility is chosen
- Problem: we don't observe utility
- However, we do observe characteristics of the options as well as the individuals
- We can think of utility as a latent variable and then model it in terms of the attributes of the choice options and the individuals:

$$u_i(\text{vote}) = \mathbf{w}_v'\alpha + \mathbf{z}_i'\gamma_v + \eta_{iv} \tag{1}$$

$$u_i(\text{abstain}) = \mathbf{w}_a'\alpha + \mathbf{z}_i'\gamma_a + \eta_{ia} \tag{2}$$

where $\mathbf{w}_v$ and $\mathbf{w}_a$ are vectors of characteristics specific to voting and abstaining that affect the average person's utility attached to voting and abstaining respectively; $\mathbf{z}_i$ is a vector of characteristics specific to individual i; and η_{iv} and η_{ia} are random disturbances.

- the disturbances arise because we can't perfectly observe utility– we can only model utility based on observed choices
- By assumption, individual i chooses to vote if the utility she attaches to voting is greater than the utility she attaches to abstaining. Similarly, individual i chooses to abstain if the utility she attaches to abstaining is greater than the utility she attaches to voting.
- This implies

$$y_i = \begin{cases} 1 & \text{if } (u_i(\text{vote}) - u_i(\text{abstain})) > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Subtracting equation 2 from equation 1 gives us:

$$
\begin{aligned}
y_i^* &= \mathbf{w}_v'\alpha - \mathbf{w}_a'\alpha + \mathbf{z}_i'\gamma_v - \mathbf{z}_i'\gamma_a + \eta_{iv} - \eta_{ia} \\
&= (\mathbf{w}_v' - \mathbf{w}_a')\alpha + \mathbf{z}_i'(\gamma_v - \gamma_a) + (\eta_{iv} - \eta_{ia}) \\
&= \mathbf{x}_i'\beta + \epsilon_i
\end{aligned}
$$

- if the original disturbances in the utility functions (the η_is) follow normal distributions, the ϵ_is will also follow a normal distribution, which gives us a probit model
- if the if the original disturbances in the utility functions (the η_is) follow type I extreme value distributions, the ϵ_is will follow a logistic distribution, which gives us a logistic regression model