MLE for a logit model

Susan Thomas
IGIDR, Bombay

September 4, 2008

Goals

- The link between workforce participation and education
- Analysing a two-variable data-set: bivariate distributions
- Conditional probability
- Expected mean from conditional probabilities
- The logistic function
- MLE for the logistic problem

The problem of workforce participation

Economic problem: A model for workforce participation

- Variable, Y_i: the participation of women in the workforce.
- $Y_i = 0$ if the woman does not participate in the workforce.
- $Y_i = 1$ if the woman is a part of the workforce.
- Y_i is binary, which means a Bernoulli distribution

Economic problem: Dataset on eduction and workforce participation

- For every observation i, Y_i is observed along with $X_i =$ education of the woman.
  This is measured by years of schooling.
- X_i is an integer, taking values from "0" – no years of school, to "12" – High School, to "13" and beyond – "College".
- The data on education is grouped into 7 categories and has the following frequency distribution:

Y_i/X_i	0-7	8	9–11	12	13–15	16–19	≥ 20
0	256	180	579	1228	463	219	7
1	143	127	560	1858	858	665	41

- There are a total of 7184 observations.

Economic problem: predicting workforce participation

- Question: Is there a relationship between a woman's workforce participation and her eduction (Y_i and X_i)?
- Question: If we know how many years of education a woman has had (X_i), what is her *expected workforce participation* $\hat{E}(Y_i)$?
- We would like to model expected workforce participation, conditional on her education.
  We need statistical models of conditional expectations, $\hat{E}(Y|X)$.

Statistical underpinnings

Creating conditional probabilities

- $E(Y|X) = \sum_{i=1}^{I} Y_i f(Y|X)$ where $f(Y|X)$ is called the conditional probability of Y given X.
- In order to calculate conditional expectations, we need to know conditional probabilities.
- In order to estimate conditional probabilities, we need to understand conditional frequency distribution/densities.

Creating conditional probabilities for the dataset

- We change from number of observations to frequencies:

Y_i/X_i	0-7	8	9–11	12	13–15	16–19	≥ 20
0	0.04	0.03	0.08	0.17	0.06	0.03	0.00
1	0.02	0.02	0.08	0.26	0.12	0.09	0.01

- The sum of all the elements add up to 1. These are the *joint frequencies* or *joint probabilities* of observing workforce participation and education.
- We get the distribution of workforce participation (or education) by summing the row elements (or column elements) as:

Y_i/X_i	0-7	8	9–11	12	13–15	16–19	≥ 20	$\hat{f}(Y)$
0	0.04	0.03	0.08	0.17	0.06	0.03	0.00	0.41
1	0.02	0.02	0.08	0.26	0.12	0.09	0.01	0.59
$\hat{f}(X)$	0.06	0.04	0.16	0.43	0.18	0.12	0.01	1

- The last row is the *marginal frequency distribution* of X. The last column is the marginal of Y.

Conditional probabilities/frequencies

- Conditional information: making a statement about Y given we know X.
- Example, what is the frequency of workforce participation of women who have 8 years of education?
- This is written as $f(X=8)$.
  The sample frequency is $\hat{f}(X=8)$.
- We use the relationship:

$$f(Y|X) = \frac{f(Y,X)}{f(X)}$$

Sample conditional probabilities

- The data is given as:

Y_i/X_i	0-7	8	9–11	12	13–15	16–19	≥ 20	$\hat{f}(Y)$
0	0.04	0.03	0.08	0.17	0.06	0.03	0.00	0.41
1	0.02	0.02	0.08	0.26	0.12	0.09	0.01	0.59
$\hat{f}(X)$	0.06	0.04	0.16	0.43	0.18	0.12	0.01	1

- The data set gives

$$\begin{aligned} \hat{f}(y=0, x=8) &= 0.025 \\ \hat{f}(y=1, x=8) &= 0.017 \\ \hat{f}(x=8) &= 0.042 \end{aligned}$$

- Then, the sample conditional frequency distribution of workforce participation of women with 8 years of education is:

Y	0	1
$\hat{f}(Y\|X=8)$	0.59	0.41

- This is a Bernoulli with a “success” rate of 0.41%.

Sample conditional probabilities for all X

- We can calculate the conditional frequency distribution for each value of X as:

$\hat{f}(Y\|X)/X_i$	0-7	8	9–11	12	13–15	16–19	≥ 20
$\hat{f}(Y=0\|X)$	0.64	0.59	0.51	0.40	0.35	0.25	0.15
$\hat{f}(Y=1\|X)$	0.36	0.41	0.49	0.60	0.65	0.75	0.85

- Conditional frequencies add up to 1 for a given X.
- In the above table, we notice that as years of education increase, the probability of workforce participation increases also.

Calculating the expected value

- As always, the focus of our analysis/estimation is the expected value of workforce participation.
- Generally, expectation $\hat{E}(Y)$ is:

$$\hat{E}(Y) = \sum_{k=1}^{K} Y_k \hat{f}(Y_k)$$

- Here, the problem is different, because we observe how many years of education the person has had. We want to calculate the expectation of Y conditional on observed X. This is:

$$\hat{E}(Y|X = x_j) = \sum_{k=1}^{K} Y_k \hat{f}(Y_k|x_j)$$

Calculating the expected value

- The unconditional expectation of Y then becomes:

$$\begin{aligned}\hat{E}(Y) &= \sum_{j=1}^{J} \hat{E}(Y|X = x_j)\hat{f}(x_j) \\ &= \sum_{j=1}^{J} \left(\sum_{k=1}^{K} Y_k \hat{f}(Y_k|x_j) \right) \hat{f}(x_j)\end{aligned}$$

  This is the Law of Iterated Expectations.
- Given the sample conditional expectation of Y for every value of X, *and*
  the marginal frequency of X,
  we can calculate the unconditional expectation of Y.

Recap on independence

- Conditional density of Y is

$$f(Y|X) = \frac{f(X,Y)}{f(X)}$$

- Indepence:

$$f(X,Y) = f(X)f(Y)$$

  Joint is a product of the marginals.
- This implies that under independence, the conditional density of Y is the same as the marginal density of Y.

$$f(Y|X) = \frac{f(X,Y)}{f(X)} = \frac{f(X) * f(Y)}{f(X)} = f(Y)$$

- Under independence, the conditional expectation of Y is the same as the unconditional expectation of Y.

$$E(Y|X) = E(Y)$$

The logistic function for a binary dependent variable

Defining the **odds**

- **Odds** are a term that can be defined for a binary variable as follows:
$$\frac{\hat{f}(Y_i = 1)}{\hat{f}(Y_i = 0)}$$
- More generally, we say it is the ratio of conditional frequencies:
$$\frac{\hat{f}(Y_i = 1|X_i = X)}{\hat{f}(Y_i = 0|X_i = X)}$$
- Example, what are the odds of a woman being in the workforce, given that she has 8 years of education?

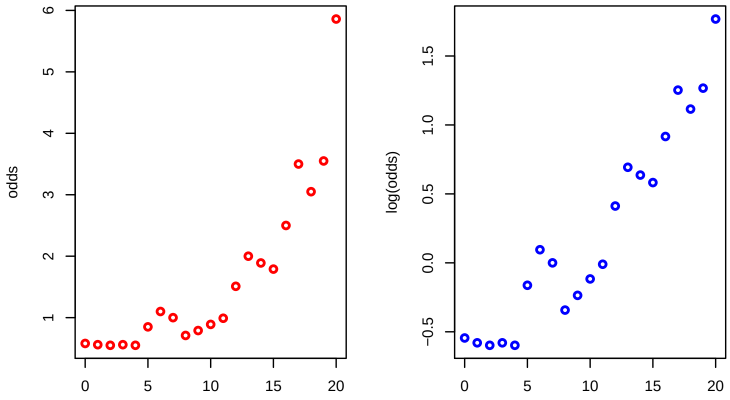
$$\frac{\hat{f}(Y_i = 1|X_i = X)}{\hat{f}(Y_i = 0|X_i = X)} = \frac{0.59}{0.41} = 1.44$$

Analysing the data

- We wish to analyse the relationship between education and probability of workforce participation.
- The table of conditional sample frequencies showed a positive relationship between the two.

Plotting the data

- We plot a graph of years of education vs. odds. We also plot the graph of years vs. log(odds).



- The plot of education vs. log(odds) is a more “linear” relationship than education vs. odds.

Econometric model for workforce participation

- The data is available in pairs of Y_i, X_i – for every woman, we observe her education and her workforce participation. Thus, every (X_i, Y_i) comes from a joint density function:

$$f(X_i, Y_i) = f(Y_i|X_i)f(X_i)$$

- Assume that the (X_i, Y_i) are independent across observations, i.
- $Y_i|X_i$ is Bernoulli distributed.
- The focus of our estimation is the "success" parameter of the Bernoulli.

Econometric model for workforce participation

- However, the success parameter for workforce participation Y is likely different for different levels of education, X.
- Model: $f(Y = 1|X) = 1 - f(Y = 0|X) = p(X)$.
  The Bernoulli success parameter for workforce participation is a function of education X
- However, we need to restrict $p(X)$ to fall between values 0 and 1, no matter what is the value of X.
- One distribution function that creates $0 \leq p(x) \leq 1$ for any value of X is the *logitistic* or the *logit* function.

Econometric model for workforce participation

- The logit function is defined as *logit(p)* as:

$$\log\left(\frac{p}{1-p}\right)$$

  where $p/(1-p)$ is the odds of success.
- This function takes a shape similar to the CDF of the normal for different values of p.
- We model success "conditional on X", which makes the logit form:

$$\log\left(\frac{p(X)}{1-p(X)}\right)$$

  Here, $\frac{p(X)}{1-p(X)}$ is the odds of success given X.

Econometric model for workforce participation

- We want to model log(odds) as a linear function of X_i:

$$\text{logit}(p(x)) = \beta_0 + \beta_1 X$$

- This gives: $p(X) = f(Y = 1|X) = \frac{exp(\beta_0+\beta_1 X)}{(1+exp(\beta_0+\beta_1 X))}$
- Thus, to know the expected conditional workforce participation of a woman given her years of education, we need to estimate β_0 and β_1.

Interpreting the logit model for workforce participation

- $p(X) = f(Y = 1|X)$ is the likelihood of observing a success given X.
$$log\left(\frac{f(Y = 1|X)}{f(Y = 0|X)}\right) = \beta_0 + \beta_1 X$$
- If we set $X_i = 0$, then $log(odds(X = 0)) = \beta_0$.
Thus, the probability that a woman with no education participates in the workforce is β_0 of the logit model.
- We can calculate that $\beta_1 = log\left(\frac{f(Y=1|(X_i+1))}{f(Y=0|(X_i+1))} / \frac{f(Y=1|(X_i)}{f(Y=0|(X_i)}\right)$
- β_1 becomes the change in the log(odds) of participation in the workforce when the amount of education shifts from $X = 0$ to a positive value of X.

Summary: Econometric model for workforce participation

- Independence of Y_i, X_i pairs across i.
- Conditional distribution of Y_i is Bernoulli with success parameter $p(X_i)$.
- Exogeniety of X_i: observed externally.
- We need to estimate β_1, β_0.
- We can use the MLE.

Setting up the MLE for the logit

L and log(L) for workforce participation

- The likelihood of observing Y_i is conditional and is as follows:

$$f_\theta(y_i) = \left(\frac{\exp \beta_0 + \beta_1 X}{(1 + \exp(\beta_0 + \beta_1 X))}\right)^{Y_i} \left(\frac{1}{(1 + \exp(\beta_0 + \beta_1 X))}\right)^{1-Y_i}$$

- We assume independence of the observed pairs (Y_i, X_i). Therefore, the likelihood, $L(Y, X)$ is:

$$\begin{aligned} L &= \prod_{i=1}^{N} \left(\frac{\exp(\beta_0 + \beta_1 X_i)}{(1 + \exp(\beta_0 + \beta_1 X_i))}\right)^{Y_i} \left(\frac{1}{(1 + \exp(\beta_0 + \beta_1 X_i))}\right)^{1-Y_i} \\ &= \left[\prod_{i=1}^{N} \left(\frac{1}{(1 + \exp(\beta_0 + \beta_1 X_i))}\right)\right] \exp\left(\beta_0 \sum_{i=1}^{N} Y_i + \beta_1 \sum_{i=1}^{N} Y_i X_i\right. \\ l &= -\sum_{i=1}^{N} log(1 + \exp(\beta_0 + \beta_1 X_i)) + \beta_0 \sum_{i=1}^{N} Y_i + \beta_1 \sum_{i=1}^{N} Y_i X_i \end{aligned}$$

First derivatives of l

- There are two parameters to differentiate l with:

$$\partial l(\beta_0, \beta_1)/\partial \beta_0 = -\sum_{i=1}^{N} \frac{\exp(\beta_0 + \beta_1 X_i)}{(1 + \exp(\beta_0 + \beta_1 X_i))} + \sum_{i=1}^{N} Y_i$$

$$\partial l(\beta_0, \beta_1)/\partial \beta_1 = -\sum_{i=1}^{N} \frac{\exp(\beta_0 + \beta_1 X_i)}{(1 + \exp(\beta_0 + \beta_1 X_i))} X_i + \sum_{i=1}^{N} X_i Y_i$$

- Solutions:

$$\sum_{i=1}^{N} \frac{\exp(\beta_0 + \beta_1 X_i)}{(1 + \exp(\beta_0 + \beta_1 X_i))} = \sum_{i=1}^{N} Y_i$$

$$\sum_{i=1}^{N} \frac{\exp(\beta_0 + \beta_1 X_i)}{(1 + \exp(\beta_0 + \beta_1 X_i))} X_i = \sum_{i=1}^{N} X_i Y_i$$

- There is no analytical close-form solution to find β_0, β_1. We use numerical methods instead.