

Why use logits?
 Logit regression diagnostics

Why use logits?

When we have a dummy dependent variable that can take on *only* the values 0 and 1, if we used OLS to relate that variable to predictors, we encounter several problems:

1. What are we predicting? The variable itself can *only* be 0 or 1. What does it mean to have some predicted number? In OLS, the predicted value is the estimate of the mean Y for all individuals with a particular set of values on the X (predictor) variables. When we were considering the binomial distribution, looking at Heads and Tails in a coin throw, our variable of interest was the *probability* of getting a Head. This is exactly the situation in dummy variable regression: the predicted value is an estimate of the *probability* that the value of the variable will be the category we've called 1, e.g. Head.

2. We know that probabilities should not be negative and should not be greater than 1. If we use OLS, there is no restriction on the possible values of the dependent variable, so we could be getting something that's negative (e.g. a negative probability of dying) or something greater than 1. Such estimates don't make sense.

The statistician's solution is to use some *transform* that allows p to be restricted to values between 0 and 1. The

$$\text{logit } P = L = \log \frac{P}{1-P}$$

logit is just such a transform. Remember:

$$P = \frac{1}{1 + e^{-L}}$$

and, if we know L, we can convert back to P. Look then at the largest and smallest values we can get for P. If L is so large that it equals 4, $\exp\{-4\} = 0$. Then $P=1$. If L is so small that it equals -4, $\exp\{4\} = 4$. Then $P= 1/4 = 0.25$.

What we have just seen is that, if we estimate logit P instead of estimating p itself, then we are restricting the possible values of P to just the range we want - namely $0 < P < 1$.

For math types: Proof that P is given by the equation above: If

$$\text{logit } P = L = \log \frac{P}{1-P} = \log P - \log(1-P)$$

then

$$e^L = \log \frac{1-P}{P}$$

and

$$e^{e^L} = \frac{1-P}{P}$$

It follows that $P + Pe^{-L} = 1$, so that $P = 1 / (1 + e^{-L})$.

Logit regression diagnostics

Multicollinearity

As in OLS, we are concerned about multicollinearity among the predictor variables. The same methods for diagnosing multicollinearity that were introduced for OLS are applicable here.

Discrimination

In the example we were looking at last time, of child survival, we saw that death rates declined with age of child. Suppose the death rate of 4 year-olds was 0 - no deaths were observed. Then age 4 is a *perfect discriminator*. If we know the child is 4, we know the dependent variable Died is 0. In such a case, we can't estimate the coefficient or the odds. Most logit regression programs will simply drop all observations in which one of the variables is a perfect discriminator.

Curvilinearity

Here the recommendation for looking at whether the relationship between X and logit Y is curvilinear is as follows: divide the X scale into a number of categories, i.e. group the X variable. Calculate the mean Y for all observations in a particular group and then find logit (mean Y). Plot the results.

You may want to transform X into a symmetric variable even before doing this type of analysis.

Another possibility, of course, is to try adding an X^2 term to see if it detects curvilinearity.

Influence statistics

In order to look at influence of particular observations, we must look at new definitions of residuals. In this case, the approach is based on the fact that the predicted probability of being a "1" is the same for all individuals who have the same set of X values -- referred to as the same *X pattern*. For example, using the child survival data from last time, we can say that 1 year old girls whose mothers have some education and live in the program area constitute pattern 1. The number of children who have pattern 1 X values is referred to as m_1 .

Pattern	Progarea	Mothed	Female	Y2	Y3	Y4
1	1	1	1	0	0	0
32	1	1	1	1	0	0

etc There are 2^6 possible patterns, although not all necessarily are represented in our dataset

But consider the girls who have pattern 32. There are 92 of them. To find that number, I did the following.

```
. quietly logistic Died Prog Mothed Female Y2-Y4
. lpredict pat, number
```

The option, number, used with the `lpredict` command, asks STATA to look at each observation in turn and look at

its X values. If that specific combination of X values hasn't occurred previously, set pat = a new *pattern number*. If the combination of X values has been encountered earlier, set pat = to that pattern number.

```
. tab pat
  covariate |
  pattern |      Freq.    Percent    Cum.
-----|-----
```

pattern	Freq.	Percent	Cum.
1	233	5.25	5.25
2	185	4.17	9.43
3	199	4.49	13.92
4	180	4.06	17.97
5	240	5.41	23.39
6	195	4.40	27.79
7	184	4.15	31.94
8	206	4.65	36.58
9	105	2.37	38.95
10	103	2.32	41.27
11	93	2.10	43.37
12	111	2.50	45.87
13	95	2.14	48.02
14	85	1.92	49.93
15	74	1.67	51.60
16	86	1.94	53.54
17	202	4.56	58.10
18	165	3.72	61.82
19	188	4.24	66.06
20	175	3.95	70.00
21	160	3.61	73.61
22	128	2.89	76.50
23	160	3.61	80.11
24	188	4.24	84.35
25	121	2.73	87.08
26	72	1.62	88.70
27	87	1.96	90.66
28	97	2.19	92.85
29	91	2.05	94.90
30	69	1.56	96.46
31	65	1.47	97.93
32	92	2.07	100.00
Total	4434	100.00	

```
-----|-----
```

We have 32 patterns in our dataset -- only half the possible number. To find out which pattern is girls in the program area who are 2 and have mothers with some education:

```
. keep if Prog==1      (2374 observations deleted)
. keep if Moth==1      (1366 observations deleted)
. keep if Female==1    (377 observations deleted)
. keep if Y2==1        (225 observations deleted)
```

I then tabulated the dependent variable

```
. tab Died
Died during|
   year |      Freq.      Percent      Cum.
-----|-----
       0 |          88      95.65      95.65
       1 |           4       4.35     100.00
-----|-----
   Total |          92     100.00
```

Therefore, of the $m_{32} = 92$ children with this pattern, 4 died, so that $Y_{32} = 4$.

These numbers, along with the predicted probability of dying, from the logistic regression, are the bases of the two *residuals* used in measuring the influence of particular *X patterns*. The difference between the examination of influence for OLS and logistic regression is this use of the *group* of people who have the same X pattern. Of course, if no two people have the same set of X's, you have as many groups as individuals.

We first need to define two measures that actually look at the *deviation of our model from perfect fit*. Perfect fit would occur if Y_j , the number of 1's among people with pattern j in their X variables, is exactly equal to m_j (the number of people with pattern j) times the estimated probability of a 1 for people with pattern j, P_j .

The Pearson residual

$$r_j = \frac{Y_j - m_j \hat{P}_j}{\sqrt{m_j \hat{P}_j (1 - \hat{P}_j)}}$$

This residual is based on the familiar binomial: the expected number of 1's in m_j trials with probability of a 1 given by \hat{P}_j is the product of those terms, given on the right of the numerator. The sd is the denominator. So r_j is simply (observed number of 1's - predicted number of 1's)/ sd.

This statistic is normally distributed, if the sample size is large. It's square has a chi-square distribution. If the Pearson residual is calculated for each observed pattern, squared, and summed, it is called the *Pearson χ^2 statistic*. It also has a chi-square distribution.

One of the influence statistics asks how much that statistic changes if all observations with a particular pattern of X values are dropped. We'll return to that in a moment.

The deviance residual

Another residual is based on the log of the observed *divided* by the predicted number of 1's and the log of the observed number of 0's divided by the predicted number of 0's in m_j trials. The *deviance residual* is defined as

$$d_j = (2 [Y_j \log_e \left(\frac{Y_j}{m_j \hat{P}_j} \right) + (m_j - Y_j) \log_e \left(\frac{m_j - Y_j}{m_j (1 - \hat{P}_j)} \right)])^{1/2}$$

where the sign of d_j is the same as that of $Y - m\hat{P}$.

The *deviance* is the sum of the squared deviance residuals.

The *leverage* statistic is again based on the hat matrix, but it is modified. The X matrix is weighted by the weights

$$w_i \hat{P}_i (1/\hat{P}_i)$$

given above. What this means is that extreme P's are not counted as much -- since w_i is greatest at $P=.5$. The leverage for a *pattern* is simply the sum of the leverage for each person who has that X pattern, so, for pattern j, it is

$$h_j' m_j h_i$$

where h_i is the leverage for person i (who has pattern j).

It is this set of residuals and leverage that is used in the influence statistics given in Hamilton. They are found by using the command *lpredict* after a logistic regression --- analogous to the way *predict* is used after an OLS regression.

Computing the influence statistics

As one might expect, the influence statistics are based on the *change* in some of the statistics defined above. In this case, however, it is the influence of dropping an entire *pattern*, e.g. *all people with pattern 2*, that is measured.

All the influence statistics are found by using versions of the *lpredict* command following a logistic command.

```
.lpredict newvar           predicted probability that Y =1 (comparable to phat)
.lpredict newvar, dbeta   aB influence statistic, analogous to Cook's D
.lpredict newvar, deviance deviance residual for jth X pattern, dj
.lpredict newvar, dx2     change in Pearson  $\chi^2$  when observations with pattern j
are omitted
.lpredict newvar, ddeviance change in deviance when observations with pattern j
are omitted
.lpredict newvar, hat     leverage of the jth pattern, hj
.lpredict newvar, number  assigns numbers to the X patterns, j=1,2,..., J
.lpredict newvar, resid   Pearson residual for the jth X pattern, rj
.lpredict newvar, rstandard standardized Pearson residual
```

The dbeta statistic, ^aB_j, if large, indicates that pattern j is exerting a large *overall* influence. If ^aB > 1, we worry.

For the **change** statistics, since they are based on χ^2 values, a large value is one greater than 4.