
NOTICE. Unless otherwise indicated, all materials on this page and linked pages at the blue.temple.edu, astro.temple.edu, and rbtaylor.net addresses are the sole property of Ralph B. Taylor and © 1999-2001 by Ralph B. Taylor. All users have the right to freely access and copy these pages provided that they: acknowledge the source, do not make changes on any pages, and do not charge more than copying costs for distribution.

PROBIT AND LOGIT ANALYSIS:
Extending the multiple regression model

Ralph B. Taylor
Department of Criminal Justice
Temple University
(c) 1999-2001 by Ralph B. Taylor

OVERVIEW

In criminal justice numerous research problems abound where the outcome is binary: Will the person released on bail show at trial or not? Will the parolee recidivate? Will the sex offender recidivate or not? Will the officer be found guilty of corruption or not? Yet, these binary (0/1) outcomes call for some special treatment. They cannot be addressed with multiple regression techniques. These notes introduce the problem of regression with a binary outcome. Passing mention is made of why regular regression is inappropriate. Further details are provided on how the probit model conceptualizes the problem. I take you through an example problem. I assume you have read the Aldrich and Nelson book already.

THE BASIC PROBLEM

Simple and multiple regression presume a linear relationship between the predictors and the outcome. There are a range of problems with this assumption (Aldrich and Nelson 13-30, espec. 26-30). Perhaps most importantly, the OLS model assumes marginal impacts of X on Y that are constant. Stated differently, the assumption is that the impact of X1 on Y, after controlling for the impacts of X2 through Xn on Y, are the same regardless of the predicted score on Y. By contrast, with a logit or probit model, the assumption is that the strength of the partial impact of X1 on Y will depend on the probability that the case scores 0 or 1 on the outcome. To put the point a bit more simply, if the chances are already strong that the case is predicted to score very close to 0 or very close to 1 on the outcome, the marginal impact of the additional predictor is small.

There are problems with the error terms as well, although WLS (weighted least squares) versions of regression help somewhat here. But these models do not capture the “true” relationship. “The only positive statement is that the OLS and WLS estimates will tend to indicate the correct sign of the effect of X on my. But none of the distributional properties holds, so statistical inferences will have not statistical justification” (A&N: 27).

THE REFORMULATION

PROBABILITY

The problem can be recast as a probability problem. You are trying to predict the probability that $Y=0$ or the probability that $Y=1$. Both these probabilities must sum to 1.

If the probability that it will rain today is .2 or 20%, the probability that it will not rain today is .8 or 80%.¹

ODDS RATIO

An odds ratio is a ratio of two probabilities (Hamilton: 220). The odds favoring $Y=1$

$O(y=1) =$

$p(Y=1)$

$1-[p(y=1)]$

(Eq. 12.1)

LOG OF THE ODDS RATIO

A log of an odds ratio is called a logit; together they make up a logistic curve. Logits can go from - infinity to + infinity as p goes from 0 to 1. See TABLE 1 below. The first row shows your odds ratio if you are playing the Pennsylvania lottery and the odds are 17 million to one.

Since these odds ratios can get staggeringly HUGE or infinitesimally TINY, to make them more workable we need to transform them. We already have talked about transformations - remember the ladder of powers from early in the semester?

So what we will do here is apply a NATURAL LOGARITHMIC transformation.

Odds ratio = as defined above (Eq 12.1)

The log of the odds ratio is the natural log of the ratio:

$L = \log_e [p/(1-p)]$ (Eq 12.2 from Ham. 7.2)

As you can see the “log odds” can go from large negative numbers to large positive numbers. These specific values (L) are logits. But at least these logits do not have the unmanageable range that the odds ratios did.

The log of the odds ratio is our new, dependent variable. So instead of having Y_i as our individual outcome scores on y , now we have L_i , the log of the odds ratio, which is the logit of Y

This transformed ratio we will think about as a linear function of X_1 , X_2 , and so on (A&N: 32, Eq. 2.3):

¹ Have you ever wondered what to do about a 20% chance of rain? Carry one fifth of your umbrella? Carry your umbrella 1/5 of the days when the probability is 20%. Only go out 4/5 as much as you usually would?

$$L = \sum(b_k X_{i,k}) = Z \tag{Eq. 12.3}$$

in other words, the log of the two odds ratio for case i are a function of the regression coefficients for each X, multiplied by i's score on each X, summed across all the predictors. This Z is very nice, except that it ranges from extremely high scores to extremely low scores. We want to transform this back so that it corresponds to the P_i; the probability that case i scores 1 rather than 0 on the outcome. In other words, we need to get it back to a 0 to 1 scale.

Since $\log(e^x) = x$ [the antilog of x is e to the x power where e is a strange and irrational number (2.78 something)]. TABLE 1 below shows this in a couple of ways.

- Column E uses A&N's formula in Eq. 2.4 (p. 32) to get P_i.
- Column F uses Eq. 7.4 (p. 221) in Hamilton to get predicted probabilities. The results are the same.

Predicted probabilities (p hat) for individual cases from a probit or logit model should reach, but never quite exceed 1, at the top end, and never go below 0 at the lower end. This new distribution of **predicted probabilities** is a **logistic function**.

Theoretically, what we have done at this point is allowed X to affect Y in a very specific type of nonlinear way, where at very high and very low values of p-hat (predicted probability), x effects are weak, and at middling values of p-hat, X effects are strong. **See Figure 7.3 on p. 221 in Hamilton**. Rarely, however, do we see X effects in our regressions over such a broad range of p-hat. More likely with real data we see a situation like **Figure 7.4 on p. 222 in Hamilton**.

In logit regression we are predicting the logit of case i (L_i):

$$L_i = A + B_1X_{i1} + B_2X_{i2} \dots + B_nX_{in} \tag{Eq. 12.4 from Hamilton Eq. 7.3}$$

where the logit is the natural log of the odds ratio.

To understand the specific coefficients, we will need to transform them by exponentiating them - this will be explained below.

PROBIT FUNCTION

The probit function is different from the logistic function. It is built on the cumulative normal distribution (CND) (see A&N, p. 34, Eq. 2.6).

With probit the function you are getting at is:

$$\text{prob}(y=1) = \Phi(Xb) \tag{Eq 12.4 from Steinberg (1997 p. 162)}$$

where Φ (phi pronounced "fee") stands for CND and you are summing across a number of coefficients and predictors.

A&N tell you that the choice between the logit curve (logistic regression) and the probit curve (probit) is fairly “arbitrary” (p. 35).

I disagree.

If you look carefully at the figure on p. 33 you will see that probit approaches the asymptote faster than logit. **My rule of thumb:** whenever possible with a binary outcome use probit. But logit will work almost as well, and in the cosmic scheme this does not matter that much. The problem is that since SPSS implements probit in an almost unworkable way, we are pretty much stuck with logistic regression in this package.

GENERALIZATION

What happens when you have more than two outcome categories? And there is no inherent ordering to these categories? For example, we are trying to predict if a case resulted in a plea bargain, a guilty verdict, or an innocent verdict.

When you have more than two unordered dependent outcome categories, you can transform each category (save one) into a binary outcome. For example, if your outcome categories are i, j, k you can make up dummies IVSNOTI (i=1; other 0) and JVSNOTJ (j=1; other 0). This is called the polytomous logit model or multinomial logit model (A&N: 39).

SPECIFICS OF THE MODEL

Questions of fit

First, there is no R squared. There is a “pseudo” R squared A&N talk about, but you want to stay away from that. See:

Forde, D. R. (1990). Overall Model Fit In Logistic Regression. Teaching Sociology, 19, 419.

What you have is relative fit measures using maximum likelihood estimation (MLE). You are trying to maximize a likelihood function, simultaneously maximizing the log of the likelihood function. [The log of the likelihood function you will see in the printout as LL.]

“A likelihood function expresses the probability of obtaining the observed sample as a function of model parameters ... What parameter values make our sample most likely?” (Hamilton: 223).

So your coefficients that you see are the results of that maximization. The program will **iterate** to get to this best fit - it tries repeatedly with different values until the change in the LL ratio is so small it thinks it is doing the best it can and stops.

You can use the LL to compare simpler to more complex models, and see if the more complex ones help. See Hamilton Eq. 7.12 on p. 225.

There is a chi squared statistics associated with each LL. It tests the null hypothesis that ALL coefficients in the model are 0. This is analogous to the F test of R squared.

Interpreting coefficients

The coefficients, to be interpreted need to be exponentiated. See, for example, Table 7.2 in Hamilton (p. 227) . The coefficient for female is -.05. You exponentiate this and you get .95

The outcome is whether or not they are in favor of school closings because of toxic contamination. 1 = Y; 0 = N. This says that women as compared to men were .95/1 times LESS likely to favor closings. In shorthand – **but be careful here** – you could say the women as compared to the men, after controlling for other predictors in the model, were 5% less likely than the men to be in favor of the school closings.

Look at the coefficient for CONTAM - whether R thought his own property had been affected by contamination. It is 1.3. You exponentiate this and you get 3.67. Those who HAD contaminated property (x=1) were 3.67 times more likely (3.67/1) to be in favor of school closings compared to those who did not have contaminated property (x=0). “The odds of favoring school closings are 3.7 times higher if the respondent’s own property or water is contaminated” (Hamilton: 231) after controlling for other predictors in the model.

SEE PARA 2 ON P. 231 OF HAMILTON FOR INTERPRETATION OF COEFFICIENTS. He writes in part:

If X is a measurement variable, then e^B describes the effect of a one unit change [on X; B is the coefficient produced by the model] If Y and X_k are unrelated, the coefficient on X_k equals 0 and $e^0 = 1$. The stronger the relation, the farther the odds ratio will be from 1. (Hamilton, 231).

This is pretty easy when we have dummy predictors (0,1), but it gets more complicated if we have something like age. **SEE HAMILTON** and the example below

AN EXAMPLE

From the Guns In American Gunowning-Households-Only subsample here is a logistic regression. The outcome is whether the person is involved in hunting (HUNTDUM)

Q 31 Thinking about just the last twelve months, have you used any of the guns in this household to go hunting? (1=no; 2=yes; recoded to 0=no; 1=yes)

We will use three predictors:

Female (0=male, 1=female); we expect women to be less likely to be involved in hunting
Nonwhite (0=white; 1=nonwhite); we expect nonwhites to be less involved in hunting
Age coded in years; we expect older respondents to be less likely to be involved in hunting

HERE ARE THE RESULTS WITH SOME COMMENTS INTERSPERSED

Total number of cases: 1154 (Unweighted)
Number of selected cases: 1154
Number of unselected cases: 0

Number of selected cases: 1154
Number rejected because of missing data: 21
Number of cases included in the analysis: 1133

Dependent Variable Encoding:

Original Value	Internal Value
.00	0
1.00	1

Dependent Variable.. HUNTDUM

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 828.71186

This describes the match between the data and the model - actually the lack thereof - at the very beginning, before any predictors have been entered. You expect this number to go down as you add in predictors and can make better predictions.

* Constant is included in the model.

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number

1..	AGE
	FEMALE FEMALE RESPONDENT
	NONWHITE Nonwhite Respondent

Estimation terminated at iteration number 4 because
Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood 711.005

This is your final -2 x LL; notice that it is smaller than when we started. Program has stopped iterating because the changes in LL are so small. It thinks it has fit the data as best it can, using the coefficients you see below.

Goodness of Fit	779.787
Cox & Snell - R ²	.144
Nagelkerke - R ²	.216

Strongly suggest you ignore the pseudo R squareds.

	Chi-Square	df	Significance
Model	117.707	3	.0000
Block	117.707	3	.0000
Step	117.707	3	.0000

This chi square test tests the null hypothesis:

In the population of gun owning households from which these households were sampled, the coefficients for all the predictors on the outcome are equal to zero

Classification Table for HUNTDUM
The Cut Value is .50

Observed		Predicted		Percent Correct
		.00	1.00	
		0	1	
.00	0	I 579	I 0	100.00%
1.00	1	I 179	I 0	.00%
Overall				76.39%

----- Variables in the Equation -----

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
AGE	-.0159	.0063	6.4090	1	.0114	-.0729	.9842
FEMALE	-2.1085	.2386	78.0822	1	.0000	-.3030	.1214
NONWHITE	-.9147	.3215	8.0948	1	.0044	-.0858	.4006
Constant	.2618	.2936	.7949	1	.3726		

The B is your coefficient. The Exp (B) is your exponentiated coefficient.

The Wald test is equivalent to the t test. You can use t-test tables to approximate the critical Wald ratio. If under significance you see a number that is smaller than .05 your results are significant at $p < .05$.

Each Wald ratio is testing the null hypothesis:

In the population of gunowning households from which these households were sampled - i.e., all gunowning households in the continental U.S. - the impact of X on the probability that the respondent had gone hunting in the last 12 months with a gun in the household, after controlling for the impacts of the other predictors in the model, is zero

How much less likely were women to go hunting, after controlling for race and age? The odds of women going hunting, compared to men were about one eighth the odds of men going hunting, after controlling for the other predictors; women were about one eighth as likely to go hunting

.121/1 = (about) 1/8

How much less likely were nonwhites to go hunting?

Now we get to age. We can say:

For each additional year of age, after controlling for the other predictors, the odds of going hunting declined by about 1.6% ($1 - .984 = .016$).

But suppose we are trying to figure out what the impact of ten years of age would be? How much less likely is a 40 year old as compared to a 30 year old?

$$(e^{-.0159})^{10} = (e^{-.159}) ; \text{Exp (B)} = .853;$$

So ten years age difference would reduce the likelihood of being victimized by almost 15 percent

THIS IS HOW YOU GET BEYOND A ONE UNIT DIFFERENCE

The matrix below tells you how closely correlated the estimates of the coefficients are; if they are too highly correlated you have problems. The only potentially problematic item here seems to be between the constant and age, but since the constant is not of substantive interest here, we will move on.

Correlation Matrix:

	Constant	AGE	FEMALE	NONWHITE
Constant	1.00000	-.93067	-.21951	-.20885
AGE	-.93067	1.00000	.06723	.10621
FEMALE	-.21951	.06723	1.00000	.07284
NONWHITE	-.20885	.10621	.07284	1.00000

TABLE 1

A	B	C	D	E	F
p(y=1)	p(y=0)	Odds ratio	Log of the Logistic odds ratiofunction	Logistic	p hat
	1-p(y=1))				
0.00000006	0.99999994	0.00000006	-16.6487	0.00000006	0.00000006
0.0000010	0.999999	0.000001	-16.6487	0.00000006	0.00000006
0.0000100	0.99999	0.000010	-13.8155	0.00000100	0.00000100
0.0001000	0.9999	0.0001	-11.5129	0.00001000	0.00001000
0.0010000	0.999	0.0010	-9.2102	0.0001	0.0001
0.0100000	0.99	0.0101	-6.9068	0.0010	0.0010
0.025	0.975	0.0256	-4.5951	0.0100	0.0100
0.05	0.95	0.0526	-3.6636	0.0250	0.0250
0.1	0.9	0.1111	-2.9444	0.0500	0.0500
0.2	0.8	0.2500	-2.1972	0.1000	0.1000
0.3	0.7	0.4286	-1.3863	0.2000	0.2000
0.4	0.6	0.6667	-0.8473	0.3000	0.3000
0.5	0.5	1.0000	-0.4055	0.4000	0.4000
0.6	0.4	1.5000	0.0000	0.5000	0.5000
0.7	0.3	2.3333	0.4055	0.6000	0.6000
0.8	0.2	4.0000	0.8473	0.7000	0.7000
0.9	0.1	9.0000	1.3863	0.8000	0.8000
0.95	0.050000000000000000000001	19.0000	2.1972	0.9000	0.9000
0.99	0.01	99.0000	2.9444	0.9500	0.9500
0.999	0.001	999.0000	4.5951	0.9900	0.9900
0.9999	0.0001000	9999.0000	6.9068	0.9990	0.9990
0.99999	0.0000100	99999.0000	9.2102	0.99990000	0.9999

References

- Aldrich, JH & Nelson, FD (1984). Linear probability, logit, and probit models. Beverly Hills: Sage.
- Hamilton, J. (1992). Regression with graphics. Monterey: Brooks/Cole.
- Steinberg, D. (1985). Logit: A Supplementary module for SYSTAT. Salford Systems.
- Steinberg, D. (1985). Probit: A Supplementary module for SYSTAT. Salford Systems
- Steinberg, D. (1997). "Probit analysis." In Systat 7.0 New Statistics. Chicago: SPSS Inc. (pp, 161-168).