STATISTICAL HORIZONS

Logistic Regression

Paul D. Allison, Ph.D.

Upcoming Seminar:

April 22-24, 2021, Remote Seminar

Table of Contents

Lecture Notes

Appendix

Logistic Regression

Logistic regression has become the standard method for modeling a dichotomous outcome in virtually all fields.

- It can accomplish virtually everything that is possible with linear regression, but in a way that is appropriate for a dichotomous outcome. And it can be generalized in many different ways.
- Many modeling strategies for linear regression will also work for logistic regression.
- Nevertheless, there are many special features of logistic regression that need to be carefully considered.

What's wrong with OLS linear regression of a dichotomous outcome?

Let y_i be a dependent variable with values of 1 and 0 and $\mathbf{x}_i$ a vector of covariates.

Linear regression with a dummy dependent variable implicitly assumes a linear probability model (LPM)

$$\begin{aligned}\pi_i &= \boldsymbol{\beta}\mathbf{x}_i \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\end{aligned}$$

where π_i is the conditional probability that y=1, $\boldsymbol{\beta}$ is a vector of coefficients and $\mathbf{x}_i$ is a vector of predictor variables (covariates).

If the LPM is correct, ordinary least squares (OLS) is unbiased for $\boldsymbol{\beta}$. But there are three problems:

1. Heteroscedasticity.
2. Non-normality
3. Possible non-linearity.

If the linear probability model is true, then heteroscedasticity is implied:

$$Var(y_i|\mathbf{x}_i) = \pi_i(1 - \pi_i) = \boldsymbol{\beta}\mathbf{x}_{i(}1 - \boldsymbol{\beta}\mathbf{x}_i), \text{ not a constant}$$

Consequently, OLS is not efficient and standard errors are biased.

Since the dependent variable is dichotomous, it can't possibly be normal.

How serious are these problems?

If the sample is moderately large, lack of normality is rarely a problem. Central limit theorem tells us that test statistics will be approximately normal.

Heteroscedasticity is more serious, but in many applications it makes little difference. There is also an easy way to correct for heteroscedasticity.

Example: Women's Labor Force Participation

Panel study of income dynamics (PSID) for 753 married women.

Mroz, T. A. 1987.
"The sensitivity of an empirical model of married women's hours to work economic and statistical assumptions." *Econometrica* 55: 765–799.

Data file can be downloaded at http://www.stata.com/texts/eacsap/
Data set is mroz.dta.

Description: The file contains data on labor force participation of 753 married women. The file includes the following variables:

inlf	=1 if in labor force in 1975, otherwise 0
hours	hours worked, 1975
kidslt6	number of kids less than 6 years
kidsge6	number of kids 6-18 years
age	woman's age in years
educ	years of schooling
wage	estimated hourly wage from earnings
repwage	reported wage at interview in 1976
hushrs	hours worked by husband, 1975
husage	husband's age
huseduc	husband's years of schooling
huswage	husband's hourly wage, 1975
faminc	family income, 1975
mtr	federal marginal tax rate facing woman
motheduc	mother's years of schooling
fatheduc	father's years of schooling
unem	unemployment rate in county of residence
city	=1 if living in a metropolitan area, else 0.
exper	actual labor market experience

OLS regression with **inlf** as the dependent variable:

<u>Stata</u>

```
use c:\data\mroz.dta, clear
reg inlf kidslt6 age educ huswage city exper
```

```
      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  6,   746) =   41.80
       Model |  46.4800152     6  7.7466692            Prob > F      =  0.0000
    Residual |   138.24774   746  .185318687           R-squared     =  0.2516
-------------+------------------------------           Adj R-squared =  0.2456
       Total |  184.727756   752  .245648611           Root MSE      =  .43049

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     kidslt6 |  -.2769332   .0334097    -8.29   0.000    -.3425214    -.211345
         age |  -.0189357   .0022871    -8.28   0.000    -.0234257   -.0144458
        educ |   .0381819   .0073786     5.17   0.000     .0236966    .0526672
     huswage |  -.0074076   .0041026    -1.81   0.071    -.0154616    .0006463
        city |  -.0006648   .0348912    -0.02   0.985    -.0691615    .0678319
       exper |   .0227591   .0021086    10.79   0.000     .0186195    .0268986
       _cons |   .7844792   .1348688     5.82   0.000     .5197117    1.049247
```

SAS

```
PROC REG DATA=my.mroz;
MODEL inlf=kidslt6 age educ huswage city exper;
RUN;
```

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	46.48002	7.74667	41.80	<.0001
Error	746	138.24774	0.18532		
Corrected Total	752	184.72776			

Root MSE	0.43049	**R-Square**	0.2516
Dependent Mean	0.56839	**Adj R-Sq**	0.2456
Coeff Var	75.73747		

Parameter Estimates

Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	**1**	0.78448	0.13487	5.82	<.0001
kidslt6	kidslt6	**1**	-0.27693	0.03341	-8.29	<.0001
age	age	**1**	-0.01894	0.00229	-8.28	<.0001
educ	educ	**1**	0.03818	0.00738	5.17	<.0001
huswage	huswage	**1**	-0.00741	0.00410	-1.81	0.0714
city	city	**1**	-0.00066481	0.03489	-0.02	0.9848
exper	exper	**1**	0.02276	0.00211	10.79	<.0001

If LPM is true, these should be unbiased estimates of the true coefficients. And the sample size is large enough that we don't have to worry about non-normality of the error term (because of central limit theorem).

But heteroscedasticity could be a problem, leading to biased standard errors and p-values. This can be easily fixed by using robust standard errors, also known as the Huber-White method or the sandwich method.

<u>Stata</u>

```
reg inlf kidslt6 age educ huswage city exper,
  robust
```

```
             |               Robust
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     kidslt6 |  -.2769332   .0312716    -8.86   0.000     -.338324   -.2155423
         age |  -.0189357   .0021187    -8.94   0.000    -.0230951   -.0147764
        educ |   .0381819   .0072138     5.29   0.000     .0240202    .0523436
     huswage |  -.0074076   .0041662    -1.78   0.076    -.0155864    .0007712
        city |  -.0006648   .0343583    -0.02   0.985    -.0681153    .0667857
       exper |   .0227591    .002025    11.24   0.000     .0187837    .0267344
       _cons |   .7844792   .1336087     5.87   0.000     .5221854    1.046773
```

SAS

```
PROC REG DATA=my.mroz;
MODEL inlf=kidslt6 age educ huswage city exper /
 HCC;
RUN;
```

HCC stands for heteroscedasticity consistent covariance matrix.

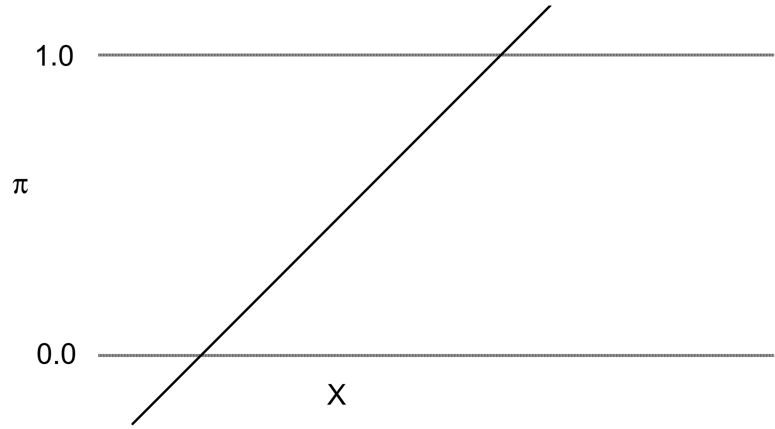
Parameter Estimates

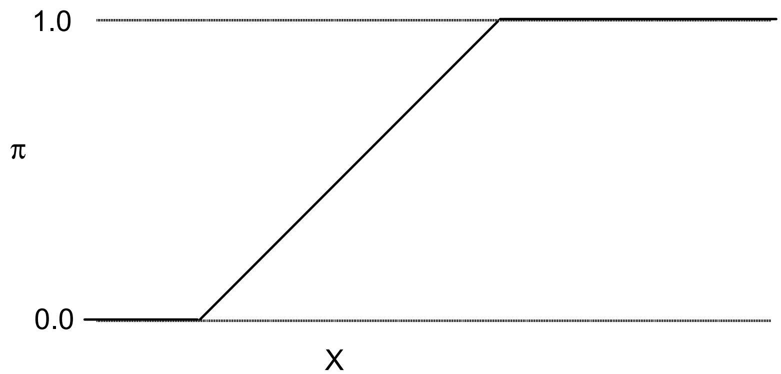
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Heteroscedasticity Consistent Standard Error	Heteroscedasticity Consistent t Value	Heteroscedasticity Consistent Pr > |t|
Intercept	Intercept	**1**	0.78448	0.13487	5.82	<.0001	0.13299	5.90	<.0001
kidslt6	kidslt6	**1**	-0.27693	0.03341	-8.29	<.0001	0.03113	-8.90	<.0001
age	age	**1**	-0.01894	0.00229	-8.28	<.0001	0.00211	-8.98	<.0001
educ	educ	**1**	0.03818	0.00738	5.17	<.0001	0.00718	5.32	<.0001
huswage	huswage	**1**	-0.00741	0.00410	-1.81	0.0714	0.00415	-1.79	0.0744
city	city	**1**	-0.00066481	0.03489	-0.02	0.9848	0.03420	-0.02	0.9845
exper	exper	**1**	0.02276	0.00211	10.79	<.0001	0.00202	11.29	<.0001

What else is wrong with the LPM?

$$\pi_i = \beta x_i$$

The left hand side is constrained to lie between 0 and 1, but the right hand side has no such constraints. For any values of the β's, we can always find some values of x that give values of π that are outside the permissible range. (See picture on page 9). A strictly linear model just isn't plausible.





Let's generate predicted values:

<u>Stata</u>

```
predict yhat
summarize yhat
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        yhat |       753    .5683931    .2486132  -.2686827   1.101222
```

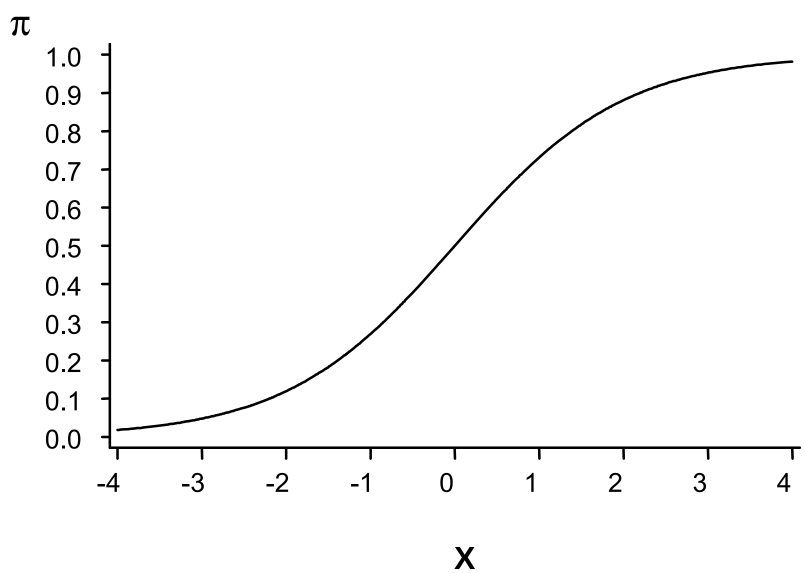
SAS

```
PROC REG DATA=my.mroz;
MODEL inlf=kidslt6 age educ huswage city exper;
OUTPUT PRED=yhat;
PROC MEANS; VAR yhat; RUN;
```

Analysis Variable : yhat Predicted Value of inlf

N	Mean	Std Dev	Minimum	Maximum
753	0.5683931	0.2486132	-0.2686827	1.1012222

A broken line is more reasonable (see picture), but is still awkward, both theoretically and computationally.



What makes most sense is an S-shaped curve like the one above. We want such a curve to be smooth, and possibly symmetrical as well.

A variety of S-shaped curves are possible, but only three used widely in practice:

1. Logit – logistic curve
2. Probit – cumulative normal distribution
3. Complementary log-log (asymmetrical).

We'll look first and primarily at the logit, but will consider the others as well.

The Odds

One component of the logistic model is the "odds", an alternative way of representing the likelihood of an event. It's often used by gamblers. If π is the probability of an event, then

$$\text{Odds} = \frac{\pi}{1-\pi} \quad .$$

This varies between 0 and $+\infty$ as π varies between 0 and 1.

Here's another way of thinking about the odds. Let S be the expected number of individuals who experience the event, and let F be the expected number who do not experience the event.

Then odds=S/F.

For example, if in a given population 728 people have blood type O and 431 people have other blood types, the odds of blood type O are 728/431=1.69.

If π = .75 then the odds is 3, or "3 to 1". If π = .6, odds = 3/2, or "3 to 2".

Probability	Odds
.1	.11
.2	.25
.3	.43
.4	.67
.5	1.00
.6	1.50
.7	2.33
.8	4.00
.9	9.00

Conversely,

$$\pi = \frac{\text{odds}}{1 + \text{odds}}$$

If the odds are 3.5, π = 3.5/(1+3.5) = .78.

Important to get used to thinking in terms of odds. Odds are a more natural scale for multiplicative comparisons. For example, if I have a probability of .60 of voting in an election, it would be absurd to say that someone else's probability of voting was twice as great. No problem on the odds scale, however.

Odds Ratios

We can measure the “effect” of a dichotomous variable by taking the ratio of the odds of the outcome event for the two categories of the independent variable. Consider the following 2 x 2 table:

	Alive	Dead
Drug	90	10
Placebo	70	30

For those who got the drug, the estimated odds of surviving are 90/10=9

For those who got the placebo, the estimated odds of surviving are 70/30=2.33.

The odds ratio is 9/2.33=3.86. This says that the effect of getting the drug is to multiply the odds of survival by 3.86.

An odds ratio of 1.00 corresponds to “no effect”. An odds ratio between 0 and 1 corresponds to a negative effect.

We often work with the log odds ratio, which is positive for a “positive effect”, zero for no effect, and negative for a “negative” effect.

The effect of drug on death is 1/(3.86)=.26. Similarly, the effect of placebo on survival is 1/(3.86)=.26. So we either work with the odds ratio or the reciprocal of the odds ratio, depending on what categories we’re comparing.

The Logistic Regression Model

We want a transformation of π that varies between – ∞ and + ∞ instead of between 0 and 1. We already have a transformation that varies between 0 and ∞, the odds. The logarithm of the odds varies between – ∞ and + ∞.

So take the logarithm of the odds and set that equal to a linear function of the x variables:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\beta}\mathbf{x}_i$$

For simplicity and generality, we use vector notation:

$$\boldsymbol{\beta}\mathbf{x}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

The left hand side is called the logit or the “log-odds"

Solving for π yields a model for the probability:

$$\pi_i = \frac{1}{1+e^{-\boldsymbol{\beta}\mathbf{x}_i}}$$

If we graph this (with a single x and β =1), we get the curve shown earlier.

Maximum Likelihood Estimation of Logistic Regression Model (Basics)

ML: Choose parameter estimates which, if true, would make the observed data as likely as possible.

Properties:

1. Consistent – as the sample gets larger, estimators converge in probability to the true values. Implies that estimates are approximately unbiased.
2. Asymptotically efficient – In large samples, estimators have (approximately) minimum sampling variation.
3. Asymptotically normal – similar to central limit theorem. Justifies use of a normal table to calculate p-values and confidence intervals.

How to do it

<u>Stata</u>

```
logit inlf kidslt6 age educ huswage city exper
```

```
Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -412.23248
Iteration 2:   log likelihood = -407.67284
Iteration 3:   log likelihood = -407.60257
Iteration 4:   log likelihood = -407.60255

Logistic regression                               Number of obs   =        753
                                                  LR chi2(6)      =     214.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -407.60255                       Pseudo R2       =     0.2083
```

inlf	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
kidslt6	-1.450909	.1988898	-7.30	0.000	-1.840725	-1.061092
age	-.09771	.0134316	-7.27	0.000	-.1240355	-.0713846
educ	.2120982	.0423591	5.01	0.000	.1290759	.2951206
huswage	-.0409741	.0220901	-1.85	0.064	-.0842699	.0023216
city	.0244788	.1919434	0.13	0.899	-.3517233	.4006809
exper	.1212059	.0132837	9.12	0.000	.0951703	.1472416
_cons	1.25433	.7380909	1.70	0.089	-.1923017	2.700961

Compared to OLS of LPM, coefficients are same sign but larger in magnitude. z-statistics and p-values are very similar.

The **or** option produces "adjusted" odds ratios instead of beta coefficients. But z-statistics are still based on the beta coefficients:

```
logit, or
```

inlf	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
kidslt6	.2343573	.0466113	-7.30	0.000	.1587023	.3460778
age	.9069118	.0121813	-7.27	0.000	.8833485	.9311037
educ	1.236269	.0523673	5.01	0.000	1.137776	1.343288
huswage	.959854	.0212032	-1.85	0.064	.9191831	1.002324
city	1.024781	.1966999	0.13	0.899	.7034747	1.492841
exper	1.128857	.0149954	9.12	0.000	1.099846	1.158634
_cons	3.505488	2.587369	1.70	0.089	.8250579	14.89404

Identical results are produced by

```
logistic inlf kidslt6 age educ huswage city exper
```

SAS

```
PROC LOGISTIC DATA=my.mroz DESC;
MODEL inlf=kidslt6 age educ huswage city exper;
RUN;
```

The DESC option is short for “descending”. Without it, the model predicts the probability of a 0 rather than a 1, and all the signs are reversed.

The LOGISTIC Procedure

Model Information

Data Set	MY.MROZ	
Response Variable	inlf	inlf
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	753
Number of Observations Used	753

Response Profile

Ordered Value	inlf	Total Frequency
1	1	428
2	0	325

Probability modeled is inlf=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1031.746	829.205
SC	1036.370	861.574
-2 Log L	1029.746	815.205

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	214.5413	6	<.0001
Score	189.4650	6	<.0001
Wald	147.0978	6	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2543	0.7381	2.8880	0.0892
kidslt6	1	-1.4509	0.1989	53.2175	<.0001
age	1	-0.0977	0.0134	52.9205	<.0001
educ	1	0.2121	0.0424	25.0715	<.0001
huswage	1	-0.0410	0.0221	3.4405	0.0636
city	1	0.0245	0.1919	0.0163	0.8985
exper	1	0.1212	0.0133	83.2543	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
kidslt6	0.234	0.159	0.346
age	0.907	0.883	0.931
educ	1.236	1.138	1.343
huswage	0.960	0.919	1.002
city	1.025	0.703	1.493
exper	1.129	1.100	1.159

Association of Predicted Probabilities and Observed Responses

Percent Concordant	79.3	Somers' D	0.589
Percent Discordant	20.5	Gamma	0.590
Percent Tied	0.2	Tau-a	0.289
Pairs	139100	c	0.794