

BIOS 600: Principles of Statistical Inference

Logistic Regression

Fall 2012

Reading

- ▶ Pagano and Gauvreau, Chapter 20
- ▶ For more information on logistic regression, take BIOS 665! Prof. Koch who teaches it is one of the most famous researchers in categorical data analysis worldwide.

Goal

Logistic regression is an important topic, and you are strongly encouraged to take a categorical data analysis course. The goal of our coverage of logistic regression is to give you the skills you need to understand results in a subject-area journal that are obtained from fitting a logistic regression model. The goal is not to give you all the skills you need to fit such models yourself on a regular basis (take BIOS 665 for that!).

Models for Binary Outcomes

Suppose we have a binary outcome (e.g., $Y = 1$ if diseased and $Y = 0$ if not) and predictors on a variety of scales.

If the predictors are discrete and the binary outcomes are independent, we can use the Bernoulli distribution for individual 0-1 data or the binomial distribution for grouped data that are counts of successes in each group.

Models for Binary Outcomes

Contingency tables for continuous predictors (and more than a few categorical predictors) can quickly become unwieldy, so we need a new analytic method to model $\pi = \text{Pr}(Y = 1)$.

One strategy might be to fit a linear regression model to the probabilities, e.g. model

$$\pi_i = \beta_0 + \beta_1 x_i.$$

The problem is that as a probability, π_i must be in the interval $[0, 1]$, but there is nothing in the model that enforces this constraint, so that you could be estimating probabilities that are negative or that are greater than 1 – not a good thing!

Models for Binary Outcomes

An alternative that is sometimes used is to fit the model

$$\pi_i = \exp(\beta_0 + \beta_1 x_i),$$

which is equivalent to

$$\ln(\pi_i) = \beta_0 + \beta_1 x_i.$$

While exponentiating the linear predictor $(\beta_0 + \beta_1 x_i)$ does ensure our estimated values of π_i are not negative, they can still be greater than 1. This is not ideal either.

Logistic Regression Model

An attractive solution is to model the log (natural log, or \ln) odds using

$$\ln \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i,$$

which is equivalent to

$$\frac{\pi_i}{1 - \pi_i} = \exp(\beta_0 + \beta_1 x_i),$$

which defines a multiplicative model for the odds. For example, if we change the j th predictor by one unit while holding the other variables constant, we multiply the odds by $\exp(\beta_j)$ because $\exp(\beta_j(z + 1)) = \exp(\beta_j z) \exp(\beta_j)$.

Logistic Regression Model

This is also equivalent to

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

The expression on the right is called a *logistic function* and cannot yield a value that is negative or a value that is > 1 . Fitting a model of this form is known as *logistic regression*.

Other transformations (also called links) can be used to ensure the probabilities lie in $[0, 1]$, including the probit (popular in econometrics, so HPM students may see it a good bit) and complementary log-log. In public health, the logistic is by far the most common model used.

Logistic Regression

$$\ln \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i$$

Negative logits represent probabilities less than one-half, and positive logits represent probabilities above one-half.

Logits can also be defined in terms of the binomial mean, $\mu_i = n_i \pi_i$, as the log of the ratio of expected successes (μ_i) to expected failures ($n_i(1 - \pi_i) = n_i - \mu_i$), as the n_i cancels out when you calculate the odds.

Contraceptive Use Data

Consider data from the Fiji Fertility Survey. Contraceptive use is considered as a function of education (lower school only versus more), age, and desire for additional children.

TABLE 3.1: Current Use of Contraception Among Married Women
by Age, Education and Desire for More Children
Fiji Fertility Survey, 1975

Age	Education	Desires More Children?	Contraceptive Use		Total
			No	Yes	
<25	Lower	Yes	53	6	59
		No	10	4	14
	Upper	Yes	212	52	264
		No	50	10	60
25-29	Lower	Yes	60	14	74
		No	19	10	29
	Upper	Yes	155	54	209
		No	65	27	92
30-39	Lower	Yes	112	33	145
		No	77	80	157
	Upper	Yes	118	46	164
		No	68	78	146
40-49	Lower	Yes	35	6	41
		No	46	48	94
	Upper	Yes	8	8	16
		No	12	31	43
Total			1100	507	1607

Contraceptive Use Data

In the contraceptive use data, we have 507 users of contraception among 1607 total women, so the probability of contraceptive use is estimated as $\frac{507}{1607} = 0.316$. The odds are $\frac{507}{1607-507} = \frac{507}{1100} = 0.461$ to one, so non-users outnumber users roughly two to one. The logit is $\log(0.461) = -0.775$.

Interpreting Parameters in Logistic Regression

Typically we interpret functions of parameters in logistic regression rather than the parameters themselves. For the model

$$\ln \left\{ \frac{\pi_i}{1 - \pi_i} \right\} = \beta_0 + \beta_1 x_i,$$

we note that the probability that $Y = 1$ when $X = 0$ is

$$\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}.$$

Interpreting Parameters in Logistic Regression

Suppose that X is a binary (0/1) variable, e.g. $X = 1$ for females and $X = 0$ for males. In this case, the coefficient β_1 has a special interpretation: we interpret $\exp(\beta_1)$ as the odds ratio of the response for the two possible levels of X . For X on other scales, $\exp(\beta_1)$ is interpreted as the odds ratio of the response comparing two values of X one unit apart.

Why? The log odds of response for $X = 1$ is given by $\beta_0 + \beta_1$, and the log odds of response for $X = 0$ is β_0 . So the odds ratio of response comparing $X = 1$ to $X = 0$ is given by
$$\frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

In a *multiple logistic regression* model with more than one predictor, this OR is interpreted conditionally on the values of the other predictors staying fixed at any given values.

Hypothesis Tests in Logistic Regression

Generally, we wish to know whether the OR=1 or equivalently whether the log OR (a β coefficient)=0. To test $H_0 : \beta_j = 0$, we can compare the ratio of a parameter estimate to its standard error using

$$z = \frac{\hat{\beta}_j - 0}{\sqrt{\widehat{\text{var}}(\hat{\beta}_j)}},$$

comparing this z-statistic to the standard normal distribution.

Confidence intervals for the effects on the logit scale,

$\hat{\beta}_j \pm 1.96\sqrt{\widehat{\text{var}}(\hat{\beta}_j)}$, are typically translated into confidence intervals for OR's by exponentiating the lower and upper confidence limits as $\left(\exp\left(\hat{\beta}_j - 1.96\sqrt{\widehat{\text{var}}(\hat{\beta}_j)}\right), \exp\left(\hat{\beta}_j + 1.96\sqrt{\widehat{\text{var}}(\hat{\beta}_j)}\right)\right)$.

Contraceptive Use Data

Recall data from the Fiji Fertility Survey. Contraceptive use is considered as a function of education (lower school only versus more), age, and desire for additional children.

TABLE 3.1: Current Use of Contraception Among Married Women
by Age, Education and Desire for More Children
Fiji Fertility Survey, 1975

Age	Education	Desires More Children?	Contraceptive Use		Total
			No	Yes	
<25	Lower	Yes	53	6	59
		No	10	4	14
	Upper	Yes	212	52	264
		No	50	10	60
25-29	Lower	Yes	60	14	74
		No	19	10	29
	Upper	Yes	155	54	209
		No	65	27	92
30-39	Lower	Yes	112	33	145
		No	77	80	157
	Upper	Yes	118	46	164
		No	68	78	146
40-49	Lower	Yes	35	6	41
		No	46	48	94
	Upper	Yes	8	8	16
		No	12	31	43
Total			1100	507	1607

Example: Logistic Regression for Contraception Data

Let $Y_i = 1$ if a woman uses contraception and 0 otherwise. Let $kids_i = 1$ if the woman wants more kids and 0 otherwise, define the age variables to take value 1 if in the specified range and 0 otherwise, and let $educ_i = 1$ if the woman has more than a lower school education and 0 otherwise. We fit the following model to the Fiji contraceptive use data.

$$\text{logit}(Pr(Y_i = 1)) = \beta_0 + \beta_1 kids_i + \beta_2 educ_i + \beta_3 age2529_i + \beta_4 age3039_i + \beta_5 age4049_i$$

We fit the model in Stata.

Example: Contraception Data

The **logistic** command provides odds ratios $\exp(\beta)$ rather than the coefficients on the log scale. The **logit** command provides the coefficients on the log scale instead.

```
. logistic contra kids educ age2529 age3039 age4049 [freq=count]
```

```
Logistic regression                               Number of obs   =       1607
                                                LR chi2(5)      =       135.86
                                                Prob > chi2     =       0.0000
Log likelihood =  -933.9192                    Pseudo R2      =       0.0678
```

contra	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
kids	.4347628	.0510718	-7.09	0.000	.3453515	.5473226
educ	1.384023	.1716681	2.62	0.009	1.085336	1.76491
age2529	1.476068	.2595666	2.21	0.027	1.045737	2.083484
age3039	2.48088	.4084051	5.52	0.000	1.796719	3.42556
age4049	3.28458	.7043125	5.55	0.000	2.15753	5.00038
_cons	.3219965	.0602413	-6.06	0.000	.2231529	.4646219

How do you interpret each estimated OR?

Example: Logistic Regression for Contraception Data

- ▶ Women who want more children have 0.43 (0.35, 0.55) times the odds of current use of contraception as their counterparts of the same age and educational level who do not want more children.

Example: Logistic Regression for Contraception Data

- ▶ Women who want more children have 0.43 (0.35, 0.55) times the odds of current use of contraception as their counterparts of the same age and educational level who do not want more children.
- ▶ Women who have more than a lower school education have 1.38 (1.09, 1.76) times the odds of current use of contraception as their counterparts of the same age and same desire for additional children who have only a lower school education.

Example: Logistic Regression for Contraception Data

- ▶ Women who are 25-29 years old have 1.48 (1.05, 2.08) times the odds of current use of contraception as their counterparts of the same educational level and same desire for more children who are < 25 .

Example: Logistic Regression for Contraception Data

- ▶ Women who are 25-29 years old have 1.48 (1.05, 2.08) times the odds of current use of contraception as their counterparts of the same educational level and same desire for more children who are < 25 .
- ▶ Women who are 30-39 years old have 2.48 (1.80, 3.43) times the odds of current use of contraception as their counterparts of the same educational level and same desire for more children who are < 25 .

Example: Logistic Regression for Contraception Data

- ▶ Women who are 25-29 years old have 1.48 (1.05, 2.08) times the odds of current use of contraception as their counterparts of the same educational level and same desire for more children who are < 25 .
- ▶ Women who are 30-39 years old have 2.48 (1.80, 3.43) times the odds of current use of contraception as their counterparts of the same educational level and same desire for more children who are < 25 .
- ▶ Women who are 40-49 years old have 3.28 (2.16, 5.00) times the odds of current use of contraception as their counterparts of the same educational level and same desire for more children who are < 25 .

Example: Logistic Regression for Contraception Data

We may wish to test the hypothesis of any association between age and contraceptive use. The Stata output gives us separate tests of $H_0 : \beta_j = 0$ for each predictor, so we get three separate p-values, comparing each older age group to those women under 25. To test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ versus the alternative that at least one coefficient is not zero, we can use a **test** statement in Stata.

```
. test age2529 age3039 age4049

( 1)  [contra]age2529 = 0
( 2)  [contra]age3039 = 0
( 3)  [contra]age4049 = 0

             chi2( 3) =      42.40
Prob > chi2 =      0.0000
```

Example: Tap Water Disinfection By-Products and Preterm Birth

Horton et al. (2011) studied the association between total organic halides (TOX) in tap water (a by-product of drinking water disinfection by chloramination) and preterm birth <32 weeks gestation (sometimes called 'very preterm' birth). Variables were coded using indicator variables as follows:

- ▶ PTB: 1 if birth occurred <32 weeks' completed gestation; 0 otherwise
- ▶ TOX50: 1 if TOX was in the 50th-75th %ile of concentration levels; 0 otherwise
- ▶ TOX75: 1 if TOX was in the 75th-90th %ile of concentration levels; 0 otherwise
- ▶ TOX90: 1 if TOX was >90th %ile of concentration levels; 0 otherwise
- ▶ TOX exposures < the 50th %ile will form the reference group

Example: Tap Water Disinfection By-Products and Preterm Birth

The model was

$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 TOX50_i + \beta_2 TOX75_i + \beta_3 TOX90_i$, where $\pi_i = Pr(PTB_i = 1)$. Results are provided below.

TOX concentration	OR (95% CI)
<50th %ile	1
50th-75th %ile	1.29 (0.75, 2.22)
75th-90th %ile	2.43 (1.22, 4.84)
>90th %ile	4.17 (1.14, 15.32)

How do we interpret the results of the study?

Pregnancy, Infection, and Nutrition (PIN) Study

The PIN Study was carried out at UNC Hospitals and the Wake County Health Department to examine risk factors for preterm birth (birth <37 completed weeks of gestation), with a focus on nutrition, infection, stress, and health disparities.



The model was $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Black}_i + \beta_2 \text{Bleed}_i + \beta_3 \text{AMA}_i$, where $\pi_i = \text{Pr}(PTB_i = 1)$.

- ▶ PTB: 1 if birth occurred before 37 weeks, 0 otherwise
- ▶ Black: 1 if mother self-identifies as black and 0 otherwise
- ▶ Bleed: 1 if mother experienced spotting or bleeding during pregnancy and 0 otherwise
- ▶ AMA: 1 if mother is advanced maternal age (≥ 35 years) and 0 otherwise

Pregnancy, Infection, and Nutrition (PIN) Study

Here we illustrate use of the **logit** command, which gives estimates on the log scale.

```
. logit ptb black bleed matage
```

```
Iteration 0:  log likelihood = -722.37614
Iteration 1:  log likelihood = -709.0186
Iteration 2:  log likelihood = -708.58934
Iteration 3:  log likelihood = -708.58903
```

Logistic regression

```
Number of obs   =      1805
LR chi2(3)      =      27.57
Prob > chi2     =      0.0000
Pseudo R2      =      0.0191
```

Log likelihood = **-708.58903**

ptb	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
black	.6937575	.1530368	4.53	0.000	.3938109	.9937042
bleed	.3996351	.1455227	2.75	0.006	.1144158	.6848545
matage	.2480388	.1992835	1.24	0.213	-.1425496	.6386273
_cons	-2.166263	.1006875	-21.51	0.000	-2.363606	-1.968919

Pregnancy, Infection, and Nutrition (PIN) Study

```
. logit ptb black bleed matage, or
```

```
Iteration 0:  log likelihood = -722.37614
Iteration 1:  log likelihood = -709.0186
Iteration 2:  log likelihood = -708.58934
Iteration 3:  log likelihood = -708.58903
```

```
Logistic regression
```

```
Log likelihood = -708.58903
```

```
Number of obs   =      1805
LR chi2(3)      =      27.57
Prob > chi2     =      0.0000
Pseudo R2      =      0.0191
```

ptb	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
black	2.001221	.3062605	4.53	0.000	1.48262	2.701222
bleed	1.49128	.2170152	2.75	0.006	1.121218	1.983483
matage	1.28151	.2553837	1.24	0.213	.8671445	1.893879

How do we interpret the results of the study?