

PUBH 501

Biostatistics

STATA: LOGISTIC REGRESSION

Overview

- Logistic regression
- Confounding
- Interaction
- 2x2 tables

Data from web source

Today we're going to use a stata web file. Find available options here:

<https://www.stata-press.com/data/r16/r.html>

Need to be connected to the internet

STATA Press Catalog Datasets Res

- Click on a filename to download it to a local folder on your machine.
- Alternatively, you can first establish an Internet connection, and then, in

`. webuse filename, clear`

to use the file. You could then save the file with Stata's **save** command

Entries starting with the letter ...

A B C D E F G H I J K L M N O

anova

- apple.dta <http://www.stata-press.com/data/r16/apple.dta>
- systolic.dta <http://www.stata-press.com/data/r16/systolic.dta>
- manuf.dta <http://www.stata-press.com/data/r16/manuf.dta>
- byssin.dta <http://www.stata-press.com/data/r16/byssin.dta>
- census2.dta <http://www.stata-press.com/data/r16/census2.dta>
- machine.dta <http://www.stata-press.com/data/r16/machine.dta>
- sewage.dta <http://www.stata-press.com/data/r16/sewage.dta>

Data from web source

- Using Stata dataset: load data using the following code

```
webuse lbw.dta
```

```
save "path\...\lbw.dta", replace
```

- We have used this dataset before. Data on birth weight & maternal factors.
- Outcome of interest is binary low birth weight (low)
- What maternal factors are associated with low birth weight?

Logistic regression

Logistic regression

- Binary outcome
 - Must be coded 0 / 1 NOT 1 / 2 or any other combination of two numbers
- Control for multiple independent variables of varying types
- Examine association of multiple independent variables with a given outcome
- Output is in odds ratios

But first

- Summary statistics
- Data editing (need to change any variables?)
- Bivariable association between outcome and independent variables
- Consider confounding

Relabel

- tab low
- label def YNFMT 0 "no" 1 "yes"
- label val low YNFMT
- tab low

Birthweight <2500g	Freq.	Percent	Cum.
0	130	68.78	68.78
1	59	31.22	100.00
Total	189	100.00	

Birthweight <2500g	Freq.	Percent	Cum.
no	130	68.78	68.78
yes	59	31.22	100.00
Total	189	100.00	

Recode to binary

- tab ptl
- recode ptl (2/3=1), gen(ptlYN)
- label var ptlYN "Any history of premature labor"
- label val ptlYN YNFMT
- tab ptlYN

Premature labor history (count)	Freq.	Percent	Cum.
0	159	84.13	84.13
1	24	12.70	96.83
2	5	2.65	99.47
3	1	0.53	100.00
Total	189	100.00	

Any history of premature labor	Freq.	Percent	Cum.
no	159	84.13	84.13
yes	30	15.87	100.00
Total	189	100.00	

-logistic- command

- Logistic regression is run using the `-logistic-` command
- `Logistic low smoke i.race`
- Where the first variable is the outcome, or the dependent variable
- The next variables are the independent variables in any order

Factor notation

- Can use continuous, categorical, and binary independent variables in regression
- With categorical variables, you can tell Stata they are categorical, otherwise it treats them as continuous
- Do this with factor notation
 - Add “i.” to the beginning of the variable name in the regression command
- Binary variables don’t need to be factorized
- This way you don’t have to make your own dummy variables

. logistic low smoke ptlYN i.race lwt

Logistic regression

Number of obs = 189

LR chi2(5) = 27.62

Prob > chi2 = 0.0000

Log likelihood = -103.52789

Pseudo R2 = 0.1177

low		Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke		2.403979	.9420098	2.24	0.025	1.115279	5.181765
ptlYN		3.398325	1.484427	2.80	0.005	1.443612	7.999803
race							
black		3.588652	1.867187	2.46	0.014	1.29434	9.949799
other		2.45803	1.041095	2.12	0.034	1.071672	5.637835
lwt							
_cons		.6814221	.6235199	-0.42	0.675	.1133818	4.09533

Note: _cons estimates baseline odds.

Summary

- Smoking, history of premature labor, and race are all significantly associated with low birth weight
- Women who smoke during pregnancy have higher odds of having low birth weight babies than children who do not smoke, when controlling for other variables of interest (OR 2.40, 95% CI 1.12-5.18)
- Compared to white women, black women have 3.6 times (95% CI 1.29-9.95) and women of another race have 2.5 times (95% CI 1.07-5.64) the odds of low birth weight babies, when controlling for smoking, history or premature labor, and maternal age.

Confounding

Confounding

- Is there confounding between smoking status and race?
- Does race distort the relationship between smoking and low birth weight?

Suspected confounding

- Suspect race may be a confounder
 - Is race related to smoking status?
 - Is it related to low birth weight?
 - Does adding race to the logistic model change the smoking OR by >10%?

2x2 tables

Calculate OR and RR from 2x2 tables

- Use the cs and cc commands to get 2x2 tables
- For cohort studies
`cs disease exposure`
- For case control and cross sectional
`cc case_status exposure_status`
- For either cs or cc, you can enter the numbers yourself without data (using the “i” suffix, for immediate)
 - Be careful in what order you enter the numbers, this matters

Calculate OR and RR from 2x2 tables

```
. csi 30 25 19 10
```

	Exposed	Unexposed	Total
Cases	30	25	55
Noncases	19	10	29
Total	49	35	84
Risk	.6122449	.7142857	.6547619
	Point estimate		[95% conf. interval]
Risk difference	-.1020408	-.3045519	.1004703
Risk ratio	.8571429	.6312732	1.163829
Prev. frac. ex.	.1428571	-.1638287	.3687268
Prev. frac. pop	.0833333		

chi2(1) = 0.94 Pr>chi2 = 0.3322

```
. cci 30 25 19 10
```

	Exposed	Unexposed	Total	Proportion exposed
Cases	30	25	55	0.5455
Controls	19	10	29	0.6552
Total	49	35	84	0.5833
	Point estimate		[95% conf. interval]	
Odds ratio	.6315789		.2204534	1.754965 (exact)
Prev. frac. ex.	.3684211		-.7549649	.7795466 (exact)
Prev. frac. pop	.2413793			

chi2(1) = 0.94 Pr>chi2 = 0.3322

Interaction

Including an interaction in Stata

- Start by examining the relationship between two variables and the outcome
- Use `##` to include an interaction term
`logistic low ptlYN##ui`
- Output includes “main effects” of the original variable, plus the interaction term
- A significance level of $p < 0.1$ is generally considered a significant interaction term

Example

logistic low pt1YN##ui

Logistic regression

Number of obs = 189

$$\text{LR } \chi^2(3) = 17.60$$

```
Prob > chi2      =      0.0005
```

Log likelihood = -108.53744

Pseudo R2 = 0.0750

	low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.ptlYN	5.484375	2.700365	3.46	0.001	2.08938	14.39583
1.ui	3.0375	1.523664	2.21	0.027	1.136419	8.118842
ptlYN#ui						
1 1	.2532447	.2407215	-1.44	0.149	.0393035	1.631733
_cons	.2962963	.0596353	-6.04	0.000	.1997127	.439589

Example

- We're interested in the effect of previous premature labor (ptlYN) on low birth weight (low)
- Can calculate by hand the OR of ptlYN at different levels of uterine irritability (ui)
 - OR for ptlYN at $ui=0$ is 5.48
 - OR for ptlYN at $ui=1$ is $5.48 * 0.25 = 1.37$

Example

- Can also get the same results using the margins command
- `margins, over(ptlYN ui) expression(exp(xb()))`
- Margins displays the predicted odds at given values. Here we will see odds at each level of our interaction

Example

```
. margins, over(ptlYN ui) expression(exp(xb()))
```

```
Predictive margins                                Number of obs      =           189
Model VCE      : OIM
```

```
Expression    : exp(xb())
over          : ptlYN ui
```

		Delta-method					
		Margin	Std. Err.	z	P> z	[95% Conf. Interval]	
ptlYN#ui							
0 0		.2962963	.0596353	4.97	0.000	.1794133	.4131793
0 1		.9	.4135215	2.18	0.030	.0895128	1.710487
1 0		1.625	.7302076	2.23	0.026	.1938194	3.056181
1 1		1.25	.8385255	1.49	0.136	-.3934798	2.89348

Example

- Can calculate the odds ratio from the predicted odds in the margins command
- OR for ptLYN at ui=0 is $1.625 / 0.296 = 5.48$
- OR for ptLYN at ui=1 is $1.25 / 0.9 = 1.38$
- If we need to get a confidence interval, we could use the Stata command `—lincom—`
 - We won't do that now, but know you can use that in the future when you need a CI and p-value for your calculated ORs

Example conclusions

- there is a difference in the relationship of pti with low based on ui
- where there is UI absent, previous premature birth is associated with increased odds of low birth weight. OR 5.48
- Where UI is present, there is no association between previous premature birth and low birth weight. OR 1.37