SESSION 11 LOGISTIC REGRESSION: MORE DETAILS

Review

Data: Binary Y's; Parameters of interest: PROPORTIONS (P's)

Logistic regression = Logit regression = Log odds regression

Logit of $P(Y=1)|X_1...X_k = \beta_0 + \beta_1.X_1 \ldots + \beta_k.X_k$

Odds = antiLog[Logit] = exp[Logit]

$= \exp[\text{Logit}] = \exp[\beta_0 + \beta_1.X_1 \ldots + \beta_k.X_k]$

Odds Ratio corresponding to δX_j $= \exp[\beta_j.\delta X_j]$
("all other X's Equal")

Fitting of β's by Method of Maximum Likelihood

Output from Logistic Regression via INSIGHT

HIV study (NEJM)

HIV = CESAREAN TPERIODS ADV_MDIS LBW

Response Distribution: Binomial
Link Function: Logit

Nominal Variable Information

Level	TPERIODS (Trimesters of Treatment)
1	0.0 — 0
2	1.5 — 1 or 2
3	3.0 — all 3

Parameter Information

Parameter		Variable	TPERIODS
(β_0)	1	INTERCEPT	
(β_1)	2	CESAREAN	
(β_2)	3	TPERIODS	0.0 <- *note that one of these is*
(β_3)	4		1.5 <- *"redundant" and its beta*
(––)	5		3.0 <- *will be set to 0*
(β_4)	6	ADV_MDIS	
(β_5)	7	LBW	

If you do not want the last one to be the "reference" best to "make your own"

Summary of Fit

Mean of Response	0.16	Deviance	6573.86	Pearson Chi-Sq	7773.98
SCALE	1.00	Deviance/DF	0.84	Pearson Chi-Sq/DF	0.99
		Scaled Dev	6573.86	Scaled Chi-Sq	7773.98

NOTES:

Mean of Response = mean(Y): 1241/7840 = 16% became HIV+

Deviance: = -2{
log[Likelihood of current model]
minus
log[Likelihood of "saturated" model]
}
"saturated" model: as many parameters as obsn's
= 6573.86

Deviance for logistic regression plays same role as residual sum of squares does for "regular" or "Gaussian-error, Identity Link" regression

NOTES on Summary of Fit (continued...):

SCALE = 1.00

If we have a good model, the magnitudes of the deviations are predicable from the Binomial, since the binomial variance for the count in a particular cell or covariate pattern is

# of subjects in cell x fitted P x (1- fitted P)

So the ratio of observed to predicted residual variance should be approximately 1. This ratio is referred to as the SCALE, and is usually set to 1 by default.

If there is considerably "greater than Binomial" variation ("extra-binomial variation" as it is known in the trade), it indicates that there may be non-independence of responses (e.g. if units are several offspring of same mother and treatments assigned to mother while units in utero, or if units are several patients of same physician). Unless you have such "correlated" responses, you should leave the scale at 1.

NOTES on Summary of Fit (continued...):

Pearson Chi-Sq = 7773.983 is $\Sigma(O-E)^2/E$,

or if you prefer, $\Sigma(Y-Y\text{-hat})^2/Y\text{-hat}$,

with the Σ over all observations.

A low value, relative to the degrees of freedom, indicates a better fit. This is a goodness of fit test rather than the usual chi-square test for testing a certain NULL hypothesis. Unfortunately, when we enter the data as 0's and 1's, the software treats each observation as a separate "cell", and you remember from your earlier statistics courses that the chi-square table is not that accurate for the $\Sigma(O-E)^2/E$ statistic if the E's are small (say less than 5). Here, the E's are fitted proportions, with values between 0 and 1! So do not take the chi-square statistic too seriously if it is based on individual Y's and Y_hats (the large DF will warn you!). If however, the data are aggregated, so that Y is no longer 0/1 but a sizable numerator (and accompanying denominator), the chi-square table is a reasonably accurate reference for the so-called "chi-square" statistic.

NOTES on Summary of Fit (continued...):

For examples of data in this "numerator/denominator" format, see (in 626) the low birthweight, asthma and Down's syndrome data. In INSIGHT, you enter the numerator as "Y" and when you check "Binomial" in the Method dialog box, you enter the denominator in the box designated Binomial. If running PROC LOGISTIC from the Program editor, you enter the numerator & denominator in the model statement as

MODEL numerator_variable/denominator_variable = ... ;

Pearson Chi-Sq/DF = 0.99 ... as the label implies.

One reason to show this is that the average value of a statistic having a chi-squared distribution with ν degrees of freedom is ν, in other words, the average value of a chi-squared random variable divided by ν, is 1.

however, as explained above, this Chi-sq/df guide works best when data are already grouped (in cells)

Analysis of Deviance

Source	DF	Deviance	Deviance/DF	Scaled Dev	Prob > Scaled Dev
Model	5	275.55	55.11	275.55	0.0001
Error	7834	6573.86	0.84	6573.86	
C Total	7839	6849.41	.	.	

NOTES:

* Statistical Inferences are now via Likelihood

* Larger ("Full") vs Smaller ("Reduced") model
  (use # of terms rather than # of variables)

  Number of terms (not counting intercept):

MODEL "Reduced"	MODEL Full	Test Statistic	diff in df
0	k (= 5 here)	Chi-Sq(model)	k

This is an "Overall test" of

H_0: $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ ARE ALL ZERO

vs H_{alt}: AT LEAST ONE of $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ IS NOT ZERO

The "Model Deviance" is a Difference of two deviances:

Recall, using L as shorthand for "Likelihood", ...

Deviance is like error sum of squares,
so will be larger with smaller model,
so..

Deviance(smaller model)
minus
Deviance(larger model)

= Model Deviance

Deviance: = -2{log[L(smaller)] - log[L("saturated")]
(smaller model)

Deviance: = -2{log[L(larger)] - log[L("saturated")]
(larger model)

----

difference = -2{log[L(smaller)] - log[L(larger)]

{ part with* log[L("saturated")] *cancels}

Type III (Wald) Tests

Source	DF	Chi-Sq	Pr > Chi-Sq
CESAREAN	1.00	40.24	0.0001
TPERIODS	2.00	110.84	0.0001
ADV_MDIS	1.00	35.51	0.0001
LBW	1.00	60.19	0.0001

NOTES:

Again, Type III refers to variable "ADDED LAST"

If 1 DF, the test statistics is the square of (beta_hat / its SE), and it is referred to a Chi_sq Table with 1 df. The reason it is Z [or Z^2 = Chi-sq(1)] rather than t is that there is no separate estimation of σ^2 when Y's are binary...

With Binary Y's , $\sigma^2(Y) = P(1-P)$, where P = Proportion of Y's that are 1, i.e., the variance is a known function of the mean, and so does not have to be estimated separately. In Gaussian error models, the separate estimation of σ^2 invokes the Student's t distribution.

If a categorical variable has c levels, represented by c-1 indicator variables, the test statistic is more complicated, and is referred to a Chi-Square Table with c-1 df.

Type III (LR) Tests

Source	DF	Chi-Sq	Pr > Chi-Sq
CESAREAN	1.00	48.89	0.0001
TPERIODS	2.00	138.99	0.0001
ADV_MDIS	1.00	33.58	0.0001
LBW	1.00	56.98	0.0001

What is "LR"

Remember -2{log[L(smaller)] - log[L(larger)]

A difference of the logs of two quantities is the log of their ratio.. can rewrite test statistic as

-2 log [L(smaller)] / log[L(larger)]

= -2 log ["Likelihood Ratio"]

WALD vs LR ??

	Source	WALD	LR
	CESAREAN	40.24	48.89
COMPARE the	TPERIODS	110.84	138.99
Chi-Sq statistics	ADV_MDIS	35.51	33.58
	LBW	60.19	56.98

The LR test statistic is more accurate, and preferred

(takes more computation, but that is hardly an issue nowadays)

Parameter Estimates

Variable	Levels *(cat.)*	DF	Estimate (β_hat)	SE	Chi-Sq	Pr > Chi-Sq	*OR-hat exp[β_hat] by hand!*
INTERCEPT		1	-2.79	.11	627.4	.0001	
CESAREAN		1	-0.85	.13	40.2	.0001	*0.43*
TPERIODS	0.0	1	1.18	.11	106.7	.0001	*3.25*
	1.5	1	0.82	.14	31.6	.0001	*2.27*
	3.0	0	0.00	.		.	
ADV_MDIS		1	0.53	.09	35.5	.0001	*1.70*
LBW		1	0.58	.07	60.1	.0001	*1.79*

Since all terms are binary, exp[β_hat] provides the estimate of the ODDS RATIO, contrasting the odds of HIV+ among infants with and without the factor in question (or in case of TPERIODS, relative to the (reference) group treated in all 3 trimesters)

Type I (LR) Tests

Source	DF	Chi-Sq	Pr > Chi-Sq
CESAREAN	1	46.16	0.0001

^^^^^^

test of $\beta_{CESAREAN}$) = 0

TPERIODS	2	130.96	0.0001

^^^^^^

test of **both** $\beta_{TPERIODS \mid CESAREAN}$) = 0

ADV_MDIS	1	41.45	0.0001

^^^^^^

test of $\beta_{ADV\_MDIS \mid CESAREAN\ TPERIODS}$) = 0

LBW	1	56.98	0.0001

^^^^^^

test of $\beta_{LBW \mid CESAREAN\ TPERIODS\ ADV\_MDIS}$) = 0

In Type I Tests , ORDER MATTERS!! Each Type I Chi_square statistic tests the contribution of the TERM, GIVEN THAT THE TERMS BEFORE IT IN THE LIST ARE ALREADY INCLUDED

TESTS OF GOODNESS OF FIT

With Binomial outcome data, it is possible to assess if "remaining" variation is compatible with pure binomial variation about the means (expected values) specified by model

This is because of the relationship between the Binomial variance and Binomial mean

"Expected" numerator = nP --> σ^2(numerator) = nP(1-P

If Deviance/DF ratio is close to 1, it may mean that other variables can't explain much more of the remaining variation (any better than chance).

Pearson Chi-square Goodness of Fit Test

The Pearson Chi_square is best calculated using the numerators for the different covariate patterns. Neither it, nor the Error Deviance statistic, is very accurate if there is only one observation in each cell or --even if there are several observations per cell but the data analysis is set up as Y=0 and Y=1 (as in our example above).

When there are only a small number of covariate patterns, each with sizable expected numbers of events and nonevents, it is helpful to redo the analysis using the cell as the unit of analysis.

See next page (24 non-empty cells or "covariate patterns")

The last 3 columns are

- **Residual from (fitted) proportion... from FIT(Y X)**
- **Predicted (fitted) proportion... from FIT(Y X)**
- **"Expected" Number Positive, calculated by user as a derived variable as the product of N_PAIRS and P_NHIVPO**

Setup using "cell" or "covariate pattern" as observation

9		Int		Nom		Int		Int		Int		Int		Int		Int		Int
24		CESAREAN		TPERIODS		ADV_MDIS		LBW		N_PAIRS		NHIVPOS		R_NHIVPO		P_NHIVPO		EXPECTED
■	1		1		0.0		0		0		372		30		0.002		0.079		29.3
■	2		0		0.0		0		0		3850		652		0.002		0.167		642.4
■	3		1		0.0		1		0		28		5		0.051		0.127		3.6
■	4		0		0.0		1		0		303		74		-0.011		0.255		77.2
■	5		1		0.0		0		1		110		17		0.022		0.132		14.6
■	6		0		0.0		0		1		767		196		-0.008		0.264		202.2
■	7		1		0.0		1		1		27		4		-0.059		0.207		5.6
■	8		0		0.0		1		1		114		40		-0.028		0.379		43.2
■	9		1		1.5		0		0		41		0		-0.056		0.056		2.3
■	10		0		1.5		0		0		441		49		-0.011		0.122		53.9
■	11		1		1.5		1		0		23		3		0.038		0.092		2.1
■	12		0		1.5		1		0		186		33		-0.015		0.192		35.8
■	13		1		1.5		0		1		7		0		-0.096		0.096		0.7
■	14		0		1.5		0		1		83		22		0.066		0.199		16.6
■	15		1		1.5		1		1		10		3		0.146		0.154		1.5
■	16		0		1.5		1		1		54		19		0.053		0.298		16.1
■	17		1		3.0		0		0		124		2		-0.010		0.026		3.2
■	18		0		3.0		0		0		878		49		-0.002		0.058		51.0
■	19		1		3.0		1		0		34		1		-0.014		0.043		1.5
■	20		0		3.0		1		0		208		24		0.020		0.095		19.8
■	21		1		3.0		0		1		25		0		-0.045		0.045		1.1
■	22		0		3.0		0		1		109		11		0.002		0.099		10.8
■	23		1		3.0		1		1		8		1		0.051		0.074		0.6
■	24		0		3.0		1		1		38		6		-0.000		0.158		6.0

Model

NHIVPOS/N_PAIRS = CESAREAN T12 T3 ADV_MDIS LBW

Response Distribution: Binomial
Link Function: Logit

One specifies the denominator "N_PAIRS" of the NHIVPOS/N_PAIRS in the window where specify binomial. A box called "binomial" is provided to indicate which variable name represents the denominator.

Summary of Fit

Mean of Response	0.15	Deviance	18.39	Pearson Chi-Sq	14.84
SCALE	1.00	Deviance/DF	1.02	Pearson Chi-Sq/DF	0.82
	.	Scaled Dev	18.39	Scaled Chi-Sq	14.84

Analysis of Deviance

Source	DF	Deviance	Deviance/DF	Scaled Dev	Pr>Scaled Dev
Model	5	275.55	55.11	275.55	0.0001
Error	18	18.39	1.02	18.39	
C Total	23	293.95			

Pearson Chi_sq based on 18 df: 24 cells to start with, but model involves 6 parameters, so 18 remaining DF.

Hosmer-Lemeshow Goodness of Fit Test

When covariates are continuous, there may be as many covariate patterns as there are individuals. In this situation, Hosmer-Lemeshow recommend grouping individuals by their predicted probabilities and then calculating the chi-square statistic using the observed and expected numbers in each category. For example, if the predicted probabilities ranged from 0.2 to 0.6, one might form say 10 equal-sized groups, with those in the 1st category having the smallest predicted probabilities, and so on. The Expected number of events for a category is the sum of the predicted probabilities for the individuals in the category.

A LARGE Chi-square statistic i.e. a large $\Sigma(O-E)^2/E$, is an indication of LACK of FIT (O's far from E's).

This test is a bit like asking how accurate are (weather) forecasters who use probabilities in their forecasts. To test the accuracy, one might group together all of the days on which the probabilities were say between 0.00 and 0.05, those between 0.05 and 0.10, etc...., enough in each group to give a sizable expected number. One can then calculate the Expected and observed numbers and the corresponding $\Sigma(O-E)^2/E$.