

Binary Logistic Regression

Gail Ryser, Ph.D.
Testing, Research-Support, and Evaluation Center

1

Presentation Overview

- Basic concepts and terminology
- Overview of the logistic regression model
- Odds and odds ratios
- SPSS examples
- Building and comparing models

2

Sample Research Questions

1. A retrospective study is conducted to examine multiple risk factors for breast cancer (Y=case, control; Xs=risk factors).
2. A researcher wishes to study how the probability of passing a course varies with the number of absences and high school GPA (Y=pass, fail; X=absences, HS GPA)

3

Generalized Linear Models

- Generalized linear models are generalizations or extensions of the general linear model (used for multiple regression, ANOVA, ANCOVA...).
- Two components in model:
 - random component-identifies the probability distribution that the response variable (Y) follows.
 - systematic component-specifies the explanatory or independent variables.
- When random component is not normally distributed, a *link function* provides the relationship between the systematic component (explanatory variables) and the expected value (mean) of the random component.

4

Link Function

- When Y is normally distributed, link function is considered identity because we can model the mean (or expected) directly, therefore no adjustment is made to the random component (Y).
- When Y is not normally distributed, must model a function of Y because we cannot model the mean directly.
- For logistic regression, Y follows the binomial distribution and we model the expectation (or mean) using \ln odds or logit.

5

Logistic Regression

- DV is dichotomous, multinomial, or ordinal.
- IV(s) or predictors are continuous or categorical.
- If DV is dichotomous, code it using 0 and 1. If DV is multinomial or ordinal, code the reference point as 0.
- Expectation of the distribution of Y (binomial for dichotomous outcomes) is equal to the proportion of 1s and is denoted as p . $1 - p$ is the proportion of 0s.

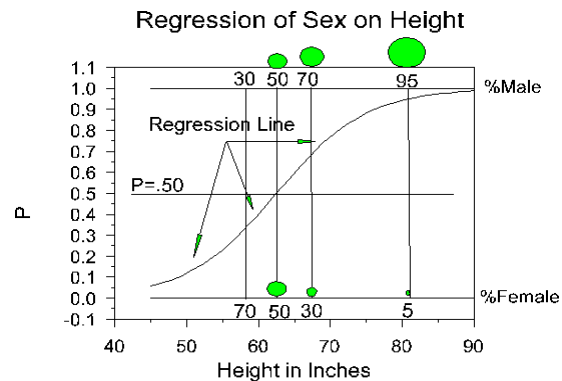
6

Frequency Table of Height by Sex

Height Range	<i>n</i>	Male (1)	Female (0)	Mean of 1 (proportion)
60-64	24	6	18	.25
65-69	23	8	15	.35
70-74	28	18	10	.64
75-79	25	20	5	.80
80-84	15	14	1	.93

7

Logistic Curve



8

Linear Probability Model

$$y = \beta_0 + \beta_1 X_1$$

$$\pi = \beta_0 + \beta_1 X_1$$

- π = probability of successes
- DV=gender (0=female; 1=male)
- IV=height

Problems:

- Probabilities are bounded, but linear functions can take on any value.
- The relationship between probabilities and X is nonlinear.

9

Logistic Regression Model

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

10

Odds

$$\text{Odds} = \frac{\pi(x)}{1 - \pi(x)}$$

11

More on Odds

1. The odds of an event = $p / (1 - p)$, where p is the probability of the event.
2. An event with m to n odds would have probability $m / (m + n)$.
3. Probability (or risk) of a male being recommended for remedial reading is $35/100=0.35$. The odd of a male being recommended is $35/65=0.54$ or $.54:1$. This means that for every male *not* recommended for remedial reading, 0.54 males will be recommended.

	Not Recommended	Recommended	Total N	Probability of Recommended	Odds of Recommended
Females	90	10	100	0.10	0.11
Males	65	35	100	0.35	0.54
Total	155	45	200	0.23	0.29

12

Odds

$$\text{Odds} = e^{\beta_0 + \beta_1 X_1}$$

13

Logit

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$$

14

Odds Ratio

$$Odds_A = \frac{p_A}{1 - p_A}$$

$$Odds_B = \frac{p_B}{1 - p_B}$$

$$Odds\ Ratio = \frac{Odds_A}{Odds_B}$$

15

More on Odds Ratios

1. Odds ratio is measure of effect size; varies from 0 to positive infinity.
 - An odds ratio of 1 indicates that the odds of the event is the same in both groups.
 - An odds ratio greater than 1 indicates that the odds of the event is greater in the reference group.
 - An odds ratio less than 1 indicates that the odds of the event is smaller in the reference group.
2. As odds of the reference group (numerator) approaches zero, odds ratio approaches zero. As odds of the non-reference group (denominator) approaches zero, odds ratio approaches positive infinity.

16

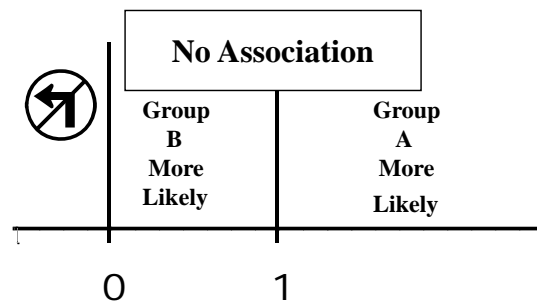
More on Odds Ratios, continued

- In our example below: males=reference group; event=recommended for reading remediation. Odds ratio=0.54/0.11=4.91.
- Interpretation: the odds of being assigned to reading are 4.91 greater for males relative to females.
- Change to relative risk: $RR = OR / [(1 - P_0) + (P_0 \times OR)]$, where P_0 = the proportion of nonexposed individuals (i.e., not the reference group), that experience the outcome. In our reading example, relative risk is $4.91 / (1 - .1) + (.1 \times 4.91) = 3.53$. CAUTION: RR *cannot* be used in case-control retrospective studies.

	Not Recommended	Recommended	Total N	Probability of Recommended	Odds of Recommended
Females	90	10	100	0.10	0.11
Males	65	35	100	0.35	0.54
Total	155	45	200	0.23	0.29

17

Properties of the Odds Ratio



18

Maximum Likelihood Estimation

- Likelihood Function = $L(\theta)$
 - Likelihood function is the joint probability of obtaining the observed Y values (in the case of logistic regression, the log odds) given θ .
 - θ is(are) the parameter(s) in the distribution.
- Maximum likelihood estimation (MLE) is used find the values of the parameters that maximize the probability of obtaining the observed set of data.
- MLE maximizes the log of the likelihood function because it is mathematically easier to do so.

19

Important Terminology

$$*D = -2 \ln(\text{likelihood of fitted model})$$

$$*G = D(\text{model without variable}) \\ - D(\text{model with variable})$$

$$*G = -2 \ln \left[\frac{(\text{likelihood without variable})}{(\text{likelihood with variable})} \right]$$

20

Important Terminology, continued

$$\text{Wald Statistic} = \left(\frac{\beta_1}{SE(\beta_1)} \right)^2$$

$$\text{CI} = \beta_1 \pm z_{1 - \alpha/2} SE(\beta_1)$$

21

Research Problem Example 1

- Is the purpose of the research, gender, and ethics position related to attitudes towards animal research in college students? College students presented with five purposes for research: cosmetics, theory, meat, veterinary, and medical.
- Chemicals introduced into cats' brains via a cannula and the cats are given various psychological tests. Following completion of testing, the cats' brains are subjected to histological analysis.
 - X_s =gender, purpose
 - Y =continue research (1=Yes, 0=No)

22

SPSS Example #1: 1 Categorical Explanatory Variable

- Open LRanimalresearch.sav
- Analyze>Regression>Binary Logistic
- Select gender (0=female; 1=male) as covariate. Select decision (0=stop research; 1= continue research) as dependent.
- Click <Options>, check iteration history and CI for exp(B).
- Click on: OK.

23

Interpreting the SPSS Output

Variables in the Equation								95% C.I. for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1 ^a	gender	1.217	.245	24.757	1	.000	3.376	2.090	5.452
	Constant	-.847	.154	30.152	1	.000	.429		

a. Variable(s) entered on step 1: gender.

Log Odds:

$$g(\text{gender}) = [\text{Ln} (P=\text{continue research})] = -.847+1.217(\text{gender})$$

Odds:

For gender = 0 (or female)

$$\text{Ln odds} = -.847+1.217(0) = e^{-.847} = .429 \quad (p = \text{odds}/1+\text{odds} = .30)$$

For gender = 1 (or male)

$$\text{Ln odds} = -.847+1.217(1) = e^{.37} = 1.448 \quad (p=.59)$$

Exp(B) [odds ratios]:

$$e^{1.217} = 1.448/.429 = 3.376 \quad (\text{males compared to females})$$

24

2 X 2 Contingency Table

decision * gender Crosstabulation

Count		gender		Total
		Female	Male	
decision	stop	140	47	187
	continue	60	68	128
Total		200	115	315

Odds of females continuing research: $60/140=0.4286$

Odds of males continuing research: $68/47=1.4468$

Odds ratio: $1.4468/0.4286=3.376$

25

Categorical Explanatory Variables

- Purpose of the research: 5 levels cosmetics, theory, meat, veterinary, and medical. When coding categorical variables, one needs k-1 design variables, thus there will be 4 coefficients.

Purpose	D ₁	D ₂	D ₃	D ₄
Cosmetics	1	0	0	0
Theory	0	1	0	0
Meat	0	0	1	0
Veterinary	0	0	0	1
Medical	0	0	0	0

26

SPSS Example #2: 2 Categorical Explanatory Variables

- Click Dialog Recall button and select logistic regression.
- Select purpose as an additional covariate.
- Click on: Categorical and select purpose as Categorical Covariate. Contrast default is Indicator and reference category is Last, do not change.
- Click on: Continue and OK.

27

Interpreting the SPSS Output

		Variables in the Equation					95% C.I. for EXP(B)		
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1 ^a	gender	1.316	.254	26.903	1	.000	3.730	2.268	6.133
	purpose			9.822	4	.044			
	purpose(1)	-.796	.384	4.286	1	.038	.451	.212	.959
	purpose(2)	-1.168	.392	8.890	1	.003	.311	.144	.670
	purpose(3)	-.804	.382	4.429	1	.035	.448	.212	.946
	purpose(4)	-.560	.377	2.213	1	.137	.571	.273	1.195
	Constant	-.229	.272	.712	1	.399	.795		

a. Variable(s) entered on step 1: gender, purpose.

Logistic regression equation:

$\ln(P=\text{continue research}) = -.229 + 1.316(\text{gender}) + .796(\text{cosmetic}) - 1.168(\text{theory}) - .804(\text{meat}) - .560(\text{veterinary})$

Exp(B) [odds ratios]:

3.730 (males compared to females)
 $.451 = 1/.451 = 2.217$ (cosmetic compared to medical)
 $.311 = 1/.311 = 3.215$ (theory compared to medical)
 $.448 = 1/.448 = 2.321$ (meat compared to medical)
 $.571 = 1/.571 = 1.751$ (veterinary compared to medical)

28

Research Problem Example 2

- A bank loan officer studies how the probability of defaulting on a loan varies with two continuous explanatory variables.
 - X_s =debt to income ratio and number of years employed by same employer
 - Y =default on loan (1=Yes, 0=No)

29

SPSS Example #3

Logistic Regression with 1 Continuous Explanatory Variable

- Open LRbankloan.sav
- Analyze>Regression>Binary Logistic
- Select previously defaulted as the dependent variable (notice that Yes (default)=1, so we are modeling the probability of a default).
- Select debt to income ratio [debtinc] as covariate.
- Click <Options>, check iteration history and CI for exp(B).
- Click on: OK.

30

Interpreting the SPSS Output

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a debttinc	.132	.014	85.377	1	.000	1.141	1.109	1.173
Constant	-2.531	.195	168.524	1	.000	.080		

a. Variable(s) entered on step 1: debttinc.

Log Odds:

$$g(\text{debttinc}) = [\text{Ln}(P = \text{Default})] = -2.531 + .132(\text{debttinc})$$

Odds:

For debttinc = 0

$$\text{Ln odds} = -2.531 + .132(0) = e^{-2.531} = .079$$

$$\text{Ln odds} = -2.531 + .132(1) = e^{-2.531 + .132} = .090$$

Exp(B) [odds ratios]:

$$e^{.132} = .090 / .079 = 1.141$$

$$g(x+c) - g(x) = c\beta_i; \text{ OR}(c) = \exp(c\beta_i)$$

$$e^{10 * .132} = 3.74$$

31

Estimated Logistic Probability

- To estimate the probability of a person with a particular debt to income ratio (say 15) defaulting on his or her bank loan (can add these to the SPSS file by selecting <Save> and under Predicted Values, check Probabilities.:

$$\pi(x) = \frac{e^{-2.531 + .132(15)}}{1 + e^{-2.531 + .132(15)}} = 0.366$$

$$\text{CI for logit: } g(x) \pm z_{1-\alpha/2} \text{SE}[g(x)] \quad 1.141 \pm 1.96 * .014 = (1.114, 1.168)$$

32

SPSS Example #4: 1 Explanatory Variable and 1 Confounder Variable

$$g(x, a) = \beta_0 + \beta_1 x + \beta_2 a$$

- Click Dialog Recall button and choose Logistic Regression.
- Select group as a single covariate, where 0=no class and 1=class.
- Click Dialog Dialog Recall button and add debtinc as a confounder variable.
- Click on: OK.

33

Interpreting the SPSS Output

Variables in the Equation								
	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a class	.774	.175	19.447	1	.000	2.168	1.537	3.058
Constant	-1.430	.130	120.424	1	.000	.239		

a. Variable(s) entered on step 1: class.

Variables in the Equation								
	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a class	.550	.189	8.419	1	.004	1.733	1.195	2.512
debtinc	.126	.014	77.125	1	.000	1.135	1.103	1.167
Constant	-2.746	.214	164.327	1	.000	.064		

a. Variable(s) entered on step 1: class, debtinc.

Log Odds:

$$g(\text{group}) = [\text{Ln}(P = \text{Default})] = -1.430 + .774(\text{class})$$

Odds:

$$\text{Ln odds} = -1.130 + .774(0) = e^{-1.430} = .239$$

$$\text{Ln odds} = -1.130 + .774(1) = e^{-1.430 + .774} = .519$$

Odds Ratio:

$$e^{.774} = .519 / .239 = 1.733$$

34

Odds Ratio

- The odds ratio becomes adjusted by debt to income ratio in the following manner:
 - $OR = e^{.550} = 1.733$
- $\text{Ln}(P = \text{Default given class} = 0 \text{ and adjusting for debt to income ratio}) = 2.746 + .550(0) + .126(9.18) = e^{-1.59} = 0.204$
- $\text{Ln}(P = \text{Default given class} = 1 \text{ and adjusting for debt to income ratio}) = -2.746 + .550(1) + .126(11.53) = e^{-1.29} = 0.354$
 - $OR = 0.204 / 0.354 = .577 = 1 / .577 = 1.733$

35

Building Models: Variable Selection

1. Conduct univariate analysis of nominal, ordinal, and continuous variables with few integer values using contingency table of outcome ($y=0, 1$) and k levels of independent variable; smoothed scatterplot of continuous variables.
2. Include variables that have p -value < 0.25 and other variables that are of known clinical or theoretical importance.
3. Verify the importance of each variable by a) examining the Wald statistic for each variable and b) comparing the estimated coefficient with the coefficient from the model containing only that variable.

36

SPSS Example #5: Building Models

- Click Dialog Recall button and choose Logistic Regression.
- Select newed, debtinc, and employ as variables of interest.
- Click <Categorical> and select newed as categorical covariate. Leave as indicator; continue.
- Click <Options> and select Hosmer-Lemeshow goodness-of-fit and Iteration History; continue.
- Click on: OK.

37

Interpreting the SPSS Output

- Goodness of fit indices
 - -2 Log Likelihood ratio
 - Hosmer-Lemeshow

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	625.633 ^a	.225	.330

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Step	Chi-square	df	Sig.
1	8.681	8	.370

Iteration History^{a,b,c}

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 1	0	805.337	-.954
2		804.365	-1.037
3		804.364	-1.039
4		804.364	-1.039

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 804.364

c. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

Step 1	Step	Chi-square	df	Sig.
	Block	178.731	4	.000
	Model	178.731	4	.000

38



Building Models, Continued

4. Check to ensure continuous variables are linear to the logit.
5. Check interactions among the variables in the model.

39



Citations for Further Info

- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*, 2nd ed. New York: John Wiley & Sons.
- Fleiss, J. L., Levin, B., & Cho Paik, M. (2003). *Statistical methods for rates and proportions*, 3rd ed. New York: John Wiley & Sons.
- Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage Publications.
- <http://www2.chass.ncsu.edu/garson/PA765/logistic.htm>

40