Multivariate Logistic Regression

Rosanna Irene Comoretto

Department of Public Health and Pediatrics,
University of Torino

Multivariate Logistic Regression

Outline

- Estimate from multivariate logistic regression models
- Comparison of the results from simple and multivariate logistic regression models to assess confounding

Example 1: Breast Feeding and child's sex

- Data from a random sample of 192 Nepali children [12, 36) months old
- **Question:** what is the relationship between breast feeding and sex of a child?
- Data:
  - Breast fed: 70%
  - Sex: 48% female (1 = male, 0 = female)

Example 1: Breast Feeding and child's sex

- The resulting unadjusted association

$$\ln\left(\frac{p}{1-p}\right) = 1.12 + 0.02 \times \text{sex}$$

- $\hat{\beta} = 0.02$: the ln(odds ratio) of being breast fed for males to females is 0.02
- $\hat{\alpha} = 1.12$: the ln(odds) of being breast fed for female children is 1.12

Example 1: Breast Feeding and child's sex

- The results of a multiple regression of breast feeding status on sex and age of the child (months)

$$\ln\left(\frac{p}{1-p}\right) = 7.2 + 0.27 \times \text{sex} - 0.24 \times \text{age}$$

- Including uncertainty:
  - 95%CI for $\hat{\beta}_1$: $(-0.50; 1.04)$, p-value $= 0.48$
  - 95%CI for $\hat{\beta}_2$: $(-0.31; -0.17)$, p-value < 0.001

Example 1: Breast Feeding, sex and age

- The slope estimate for **sex** is $\hat{\beta}_1 = 0.27$
  - an estimated ln(odds ratio) of breast feeding for male children to female children, who are *of the same age*
  - it is called *age adjusted association between breast feading and sex*
- The resulting odds ration estimate is $e^{0.27} = 1.30$
  - male children in che sample have 30% greater odds of being breastfed than females *of the same age*
- The 95% for the age adjusted odds ratio for males compared to females is $(e^{-0.50}, e^{1.04}) \rightarrow (0.61, 2.82)$

Example 1: Breast Feeding, sex and age

- The slope estimate for **age** is $\hat{\beta}_2 = -0.24$
  - an estimated ln(odds ratio) of breast feeding for children who differ by one month in age (older to younger) but are *of the same sex*
  - it is called *sex adjusted association between breast feading and age*
- The resulting odds ration estimate is $e^{-0.24} = 0.79$
  - a one month difference in age is associated with a 21% reduction in the odds of being breastfed (older to younger) among children of the *same sex of the same age*
- The 95% for the true sex adjusted odds ratio for age is $(e^{-0.31}, e^{-0.17}) \rightarrow (0.73, 0.84)$

Example 1: Presentation of finding

- In research articles, frequently a single table of unadjusted and adjusted associations is reported (for non-randomized studies)

Table 1: Logistic Regression Results for Predictors of Breast Feeding
Odds Ratio (95% CI)

Predictor	Unadjusted	Adjusted
Sex		
female	1	1
male	1.02 (0.55, 1.90)	1.30 (0.61, 2.82)
age (months)	0.79 (0.73 - 0.84)	0.79 (0.73, 0.84)
Baseline Odds (exponetiated intercept)		1,333

Example 1: Additional Predictors

- Some other additional predictors of interest include:
  - maternal parity
    - No previous children 17%
    - 1 previous child 16%
    - 2 previous children 14%
    - 3 previous children 15%
    - $>$ 3 previous children 14%
  - Maternal age: mean $=$ 27.7 years, range 17-43 years

Example 1: Presentation of finding

- The results of several models are presented

Table 1: Logistic Regression Results for Predictors of Breast Feeding
Odds Ratio (95% CI)

Predictor	Unadjusted	Model 2	Model 3	Model 4
Sex				
female	1	1	1	1
male	1.02 (0.55, 1.90)	1.30 (0.61, 2.82)	1.23 (0.54, 2.77)	1.22 (0.54, 1.77)
age (months)	0.79 (0.73 - 0.84)	0.79 (0.73, 0.84)	0.77 (0.72, 0.83)	0.77 (0.71, 0.83)
Maternal Parity	p=0.40		p=0.12	p=0.11
No previous children	1		1	1
1 previous child	0.38 (.12, 1.22)		0.23 (0.05, 1.01)	0.24 (0.05, 1.14)
2 previous children	0.50 (0.15, 1.69)		0.36 (0.08, 1.54)	0.39 (0.08, 1.83)
3 previous children	0.34 (0.11, 1.10)		0.18 (0.04, 1.05)	0.21 (0.04, 1.04)
>=4 previous children	0.61 (0.18, 2.08)		0.61 (0.18, 2.08)	0.75 (0.14, 4.2)
Mother's Age (years)	0.99 (0.94, 1.04)			0.98 (0.89, 1.08)
Baseline Odds (exponetiated intercept)		1,333	4,932	7,071

Example 1: Presentation of finding

- The results of several models are presented

Table 1: Logistic Regression Results for Predictors of Breast Feeding
Odds Ratio (95% CI)

Predictor	Model 4
female	1
male	1.22 (0.54, 1.77)
age (months)	0.77 (0.71, 0.83)
No previous children	1
1 previous child	0.24 (0.05, 1.14)
2 previous children	0.39 (0.08, 1.83)
3 previous children	0.21 (0.04, 1.04)
>=4 previous children	0.75 (0.14, 4.2)
Mother's Age (years)	0.98 (0.89, 1.08)
Baseline Odds	7,071

Example 2: Predictors of Obesity

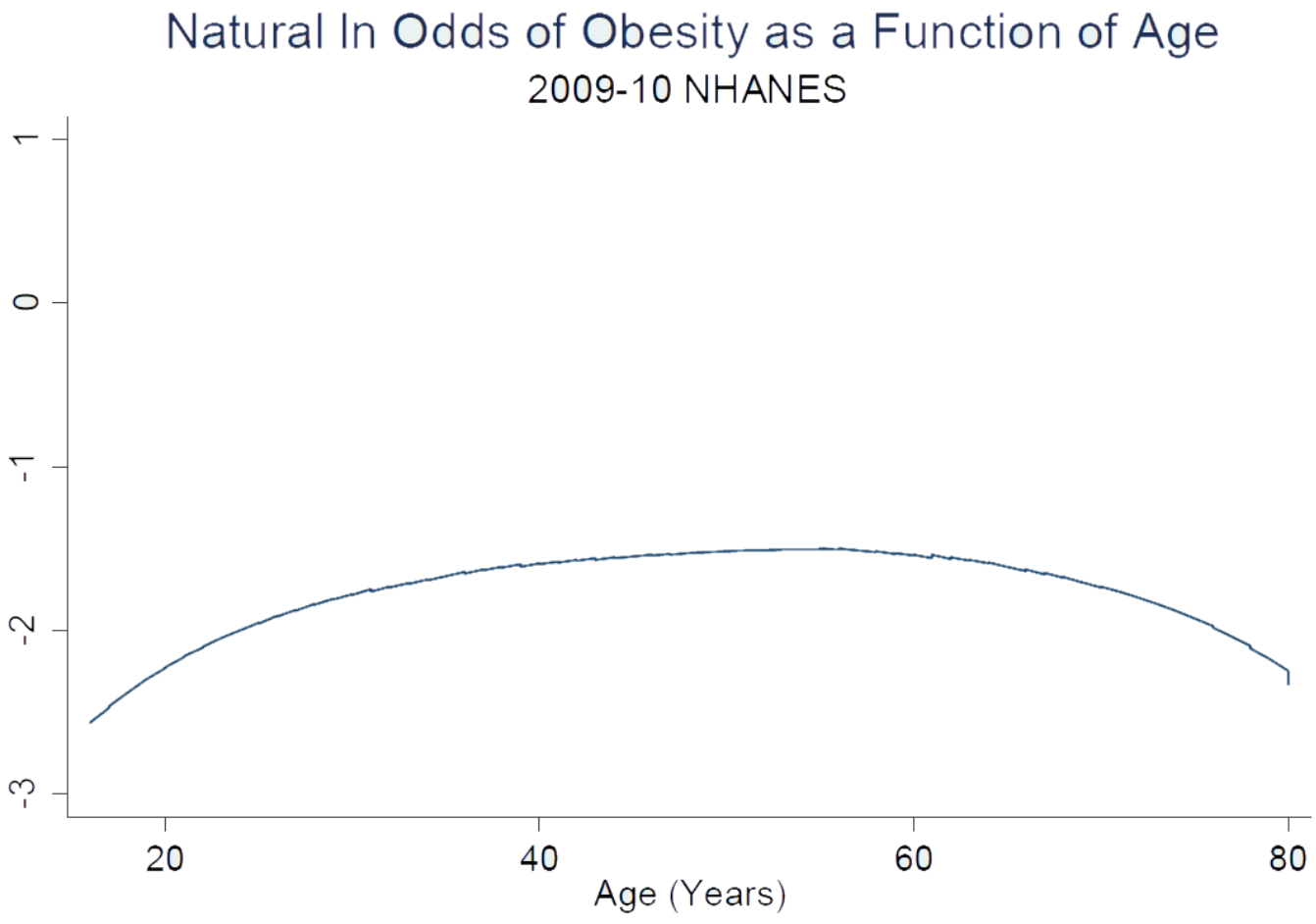
- Data from 2009-10 NHANES
- Sample of over 6,400 US residents, 16-80 years old
- HDL levels
  - mean: 52.4 mg/dL
  - sd: 16 mg/dL
  - range: 11-14
  - obesity: 15% of the sample (by BMI)
- Other potential predictors include sex, age (yrs) and marital status

Example 2: Predictors of Obesity

- Data from 2009-10 NHANES
- Sex distribution:
  - 49% F, 51% M
- Age (years): 46.3 years, range: 16 to 80
- Marital status
  - Married 52%
  - Widowed 9%
  - Divorced 11%
  - Separated 3%
  - Never Married 18%
  - Living together 7%

Example 2: Predictors of Obesity

- Obesity/age relationship as from a lowess plot (it shows unadjusted association)



Example 2: Logistic regression

Table 1: Logistic Regression Results for Predictors of Obesity
Odds Ratio (95% CI)

Predictor	Unadjusted	Model 2	Model 3
HDL (mg/dL)	0.967 (0.961, 0.973)	0.956 (0.951, 0.962)	0.958 (0.952, 0.964)
Males	1.75 (1.52,2.01)	2.63 (2.25, 3.07)	2.61 (2.22, 3.08)
Age Category	p<0.001	p < 0.001	p< 0.001
< 30 years	1	1	1
30-46 years	1.79 (1.46, 2.19)	1.76 (1.42, 2.17)	1.62 (1.25, 2.10)
46-62 years	1.82 (1.49, 2.24)	1.95 (1.57, 2.43)	1.79 (1.37, 2.36)
>= 62 years	1.47 (1.19, 1.81)	1.66 (1.34, 2.07)	1.53 (1.15, 2.05)
Marital Status	p=0.69		o
Married	1		1
Widowed	1.10 (0.85, 1.41)		1.05 (0.79, 1.41)
Divorced	1.13 (0.90, 1.42)		1.14 (0.89, 1.44)
Separated	1.18 (0.81, 1.73)		1.16 (0.78, 1.72)
Never Married	0.99 (0.81, 1.20)		1.19 (0.94, 1.50)
Living together	0.91 (0.69, 1.20)		0.92 (0.68, 1.24)
Baseline Odds (exponentiated intercept)		0.61	0.58

Example 2: Logistic regression

Table 1: Logistic Regression Results for Predictors o
Odds Ratio (95% CI)

Predictor	Model 2
HDL (mg/dL)	0.956 (0.951, 0.962)
Males	2.63 (2.25, 3.07)
Age Category	p < 0.001
< 30 years	1
30-46 years	1.76 (1.42, 2.17)
46-62 years	1.95 (1.57, 2.43)
>= 62 years	1.66 (1.34, 2.07)
Marital Status	
Married	
Widowed	
Divorced	
Separated	
Never Married	
Living together	
Baseline Odds	0.61

Basics of Model Selection and Estimating Outcomes

Outline

- Understand the "linearity assumption" as it applies to multiple logistic regression
- Explain different strategies for picking a *final* (multivariate) regression model among candidates

-??? Use the results of multivariate logistic regression models to compare groups who differ by more than one predictor X, and estimate proportions/probabilities for groups given their Xs values

Method of Estimation

- The method used to estimate the regression coefficients in multivariate logistic regression is called the **method of maximum likelihood**
  - ame as that used with simple logistic
- The estimates of the slopes $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ and intercept $\hat{\alpha}$ are the values that make the observed data *most likely* among all choices of values for $\hat{\alpha}$ and $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$
- This method is computationally intensive and is best handled by computer

Linearity assumption

- The **linearity assumption** assumes that the adjusted relationship being estimated beween the $\ln(\text{odds of } Y = 1)$ for a binary outcome Y and each predictor X_j is linear in nature
  - this is an issue for continuous predictors, not for binary or multi-categorical predictors

Choosing a *Final Model*

- When faced with potentially many possible predictors, how to come up choosing the *best* model?
- Model building and selection is a combination of science, statistics, and the research goal(s)

Choosing a *Final Model*

- If goal is to maximize precision of adjusted estimates:
  - Keep only those predictors that are statistically significant in the final model
- If goal is to present results comparable to results of similar analyses presented by other researchers (on similar or different populations)
  - Present at least one model that includes the same predictor set as the other research
- If goal is prediction ... >- this is a slightly more complicated story

Example 1: Prediction with Regression Results

- Recall the models for looking at predictors of breast feeding status in Nepali children, 12-36 months

Table 1: Logistic Regression Results for Predictors of Breast Feeding
Odds Ratio (95% CI)

Predictor	Unadjusted	Adjusted
Sex		
female	1	1
male	1.02 (0.55, 1.90)	1.30 (0.61, 2.82)
age (months)	0.79 (0.73 - 0.84)	0.79 (0.73, 0.84)
Baseline Odds (exponetiated intercept)		1,333

Example 1: Prediction with Regression Results

_ Adjusted model

$$\ln\left(\frac{p}{1-p}\right) = 7.2 + 0.27 \times \text{sex} - 0.24 \times \text{age}$$

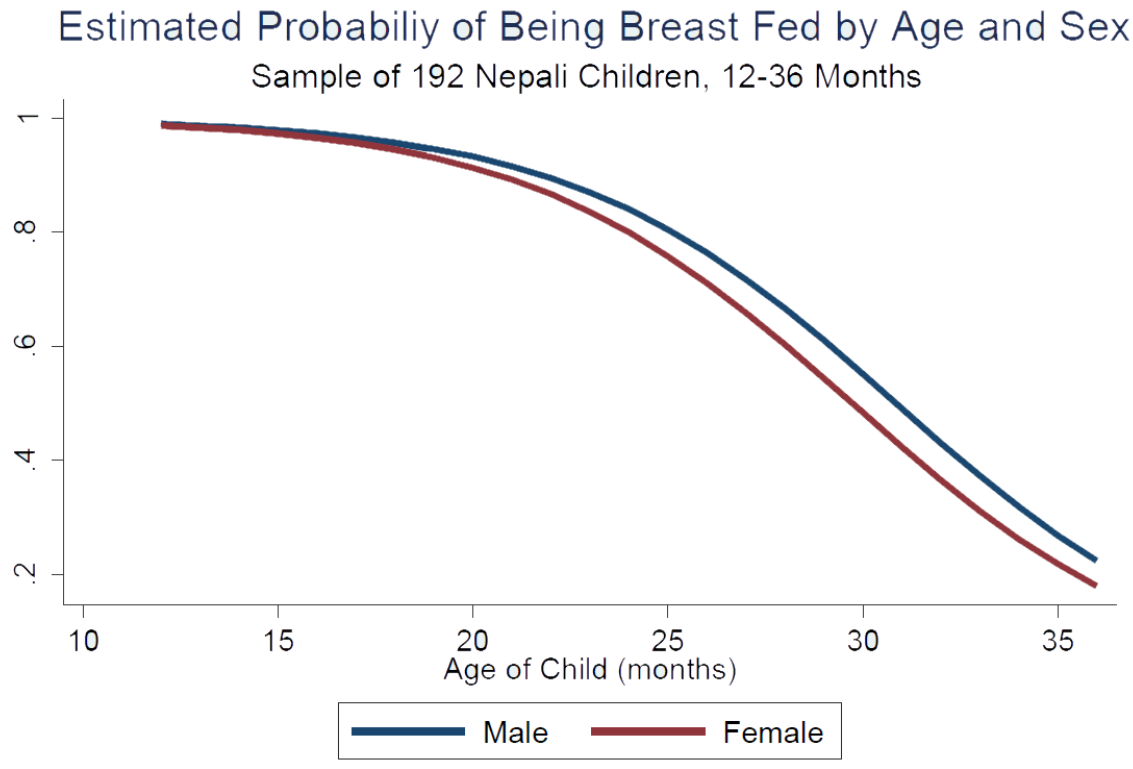
- **Question:** estimate the probability (proportion of) that female children 22 months old are breast fed

Example 1: Prediction with Regression Results

Predictor	Adjusted
Sex	
female	1
male	1.30 (0.61, 2.82)
age (months)	0.79 (0.73, 0.84)
Baseline Odds (exponetiated intercept)	1,333

Example 1: Prediction with Regression Results

- Possible way to present the associated estimated probabilities from the logistic regression results



Example 1: Prediction with Regression Results

_ Adjusted model

$$\ln\left(\frac{p}{1-p}\right) = 7.2 + 0.27 \times \text{sex} - 0.24 \times \text{age}$$

- Estimate the odds ratio of being breast fed for female (sex = 0) children 22 months compared to male (sex = 1) children , 19 months old
  - F, 22 months: $\ln(\text{odds}) = 7.2 - 0.24 \times 22 + 0.27 \times 0 = 1.92$
  - M, 19 months: $\ln(\text{odds}) = 7.2 - 0.24 \times 19 + 0.27 \times 1 = 2.91$

Example 2: Predictors of Obesity: NHANES

Table 1: Logistic Regression Results for Predictors of Obesity
Odds Ratio (95% CI)

Predictor	Unadjusted	Model 2	Model 3
HDL (mg/dL)	0.967 (0.961, 0.973)	0.956 (0.951, 0.962)	0.958 (0.952, 0.964)
Males	1.75 (1.52,2.01)	2.63 (2.25, 3.07)	2.61 (2.22, 3.08)
Age Category	p<0.001	p < 0.001	p< 0.001
< 30 years	1	1	1
30-46 years	1.79 (1.46, 2.19)	1.76 (1.42, 2.17)	1.62 (1.25, 2.10)
46-62 years	1.82 (1.49, 2.24)	1.95 (1.57, 2.43)	1.79 (1.37, 2.36)
>= 62 years	1.47 (1.19, 1.81)	1.66 (1.34, 2.07)	1.53 (1.15, 2.05)
Marital Status	p=0.69		o
Married	1		1
Widowed	1.10 (0.85, 1.41)		1.05 (0.79, 1.41)
Divorced	1.13 (0.90, 1.42)		1.14 (0.89, 1.44)
Separated	1.18 (0.81, 1.73)		1.16 (0.78, 1.72)
Never Married	0.99 (0.81, 1.20)		1.19 (0.94, 1.50)
Living together	0.91 (0.69, 1.20)		0.92 (0.68, 1.24)
Baseline Odds (exponentiated intercept)		0.61	0.58

Example 2: Predictors of Obesity: NHANES

- $\ln(\text{odds of being obese}) =$
  $-0.5 - 0.045X_1 + 0.56X_2 + 0.67X_3 + 0.56X_4 + 0.97X_5$

Table 1: Logistic Regression Results for Predictors o
Odds Ratio (95% CI)

Predictor	Model 2
HDL (mg/dL)	0.956 (0.951, 0.962)
Males	2.63 (2.25, 3.07)
Age Category	p < 0.001
< 30 years	1
30-46 years	1.76 (1.42, 2.17)
46-62 years	1.95 (1.57, 2.43)
>= 62 years	1.66 (1.34, 2.07)
Marital Status	
Married	
Widowed	
Divorced	
Separated	
Never Married	

Example 2: Prediction of Obesity

- Estimate the proportion (probability) of 50 year old males with HDL of 80 mg/dL who are obese

$$\begin{aligned}&\ln(\text{odds of being obese})\\&= -0.5 - 0.045 \times 80 + 0.56 \times 1 + 0.67 \times 1 + 0.56 \times 0 + 0.97 \times 0\\&= -2.46\end{aligned}$$

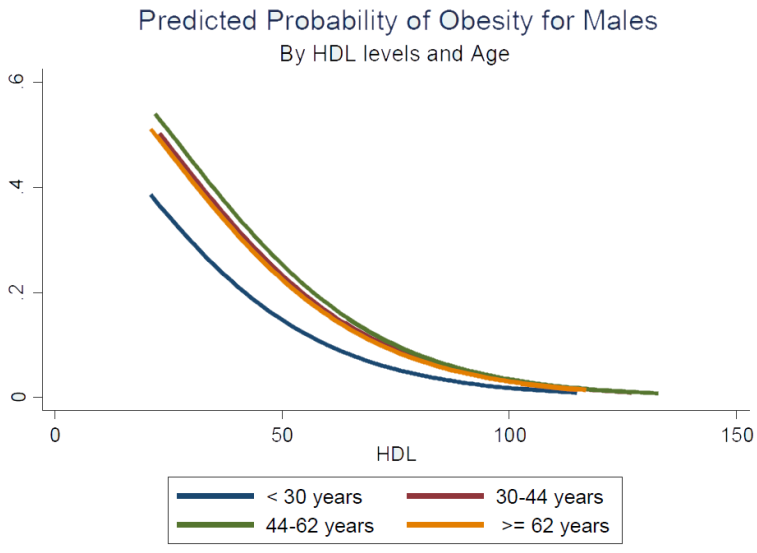
- Thus

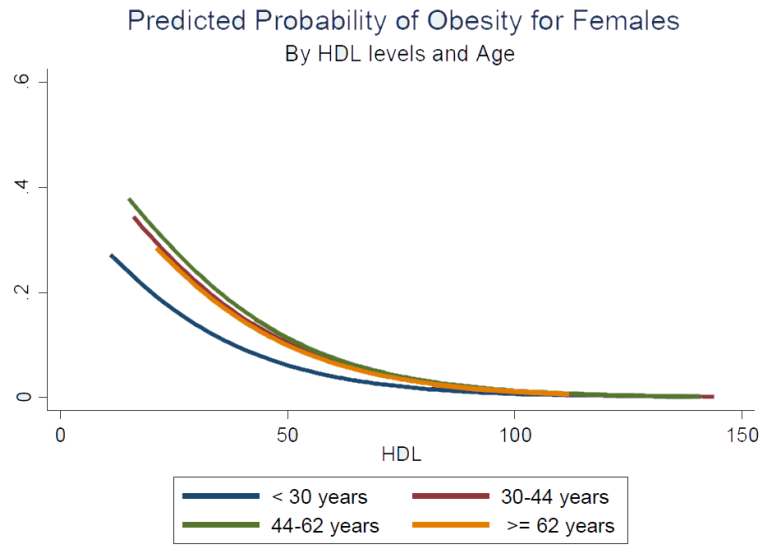
$$\text{odds} = \mathrm{e}^{-2.466} = 0.09$$

from which:

$$p = \frac{0.09}{1 + 0.09} = 8.3\%$$

Example 2: Prediction of Obesity





Summary

- Multivariate logistic regression results can be used to estimate probabilities (proportions) of binary outcomes for a given subset in a population given their predictor values
- Multivariate logistic results can be used to estimate odds ratios between groups who differ by more than one characteristic (predictor)
- Confidence intervals can be estimated for each of the above

Tying it All Together: Examples of Logistic Regression and Some Loose Ends