Lecture 18: Multiple Logistic Regression

Mulugeta Gebregziabher, Ph.D.

BMTRY 701/755: Biostatistical Methods II Spring 2007

Department of Biostatistics, Bioinformatics and Epidemiology

Medical University of South Carolina

Topics to be covered

- Review
  1. Purpose of empirical models: Association vs Prediction
  2. Design of observational studies: cross-sectional, prospective, case-control
  3. Randomization, Stratification and Matching
- Multiple logistic regression
  1. The model
  2. Estimation and Interpretation of Parameters
  3. Confounding and Interaction
  4. Effects of omitted variables
  5. Model Fitting Strategies
  6. Goodness of Fit and Model Diagnostics
- Matching (group and individual)
- Conditional vs Unconditional analysis
- Methods III: Advanced Regression Methods

Review: Purpose of empirical models

Empirical models: are models that are fitted to provide succinct descriptions of relationships observed in data. They can be of different forms, here we focus on regression models that have wide applicability

- They are data-driven models that provide a range of possible relationships between variables often specified by mathematical convenience and a preference for simplicity.
- If the model fits well, inferences are possible about the nature of relationships between variables in the ranges where they are observed (NO extrapolation)
- Examples: **Association studies** in Epidemiology and **Prediction studies** in clinical or policy making research

Association Studies

- Interest centers on what variables (variables of interest and adjustment variables) are in the model and the size and sign of their coefficients

- Predicted value for each observation or model fit is not of interest per se

Example 1. After adjusting for appropriate covariates, is broccoli intake associated with colorectal adenomatous polyps?

$\text{logit}(\Pr(\text{polyps})) = \beta_0 + \beta_1 \text{energyintake} + ... + \beta_k \text{Broccoliintake}$

Example 2. After adjusting for age, is heart disease (HD) associated with hypertension?

$\text{logit}(\Pr(\text{HD})) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{hypertension}$

Prediction Studies

- Interest centers on being able to accurately estimate or predict the response for a given combination of predictors

- Focus is not much about which predictor variable allow to do this or what their coefficients are (Model fit is important)

Example 1. A multiple logistic regression model for screening diabetes (Tabaei and Herman (2002) in Diabetes Care, 25, 1999-2003)

$$\text{logit}(\Pr(\text{Diabetes})) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Plasmaglucose} + \beta_3 \text{Postprandialtime} + \beta_4 \text{Female} + \beta_5 \text{BMI}$$

Estimates: $\hat{\beta}_0 = -10.038, \hat{\beta}_1 = 0.033, \hat{\beta}_2 = 0.031, \hat{\beta}_3 = 0.250, \hat{\beta}_4 = 0.562, \hat{\beta}_5 = 0.035$

They used a cutoff of 20% to predict a previously undiagnosed diabetes with sensitivity=65% and specificity=96%

Review: Designs for observational studies

We discuss three important designs that have a lot of use of logistic regression in their analysis.

Define X to denote an exposure or treatment and D to be an outcome indicator (disease, death, etc).

Example: For a binary X and D,

CROSS-SECTIONAL DESIGN: randomly select n from a population of N records

		D	
X	D=1	D=0	total
X=1	n_{11}	n_{10}	$n_{1.}$
X=0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	$\mathbf{nfixed}$

Review: Designs for observational studies

PROSPECTIVE DESIGN: randomly select $n_{1.}$ from N_1 with $X = 1$ and $n_{0.}$ from N_0 with $X = 0$

	D		
X	D=1	D=0	total
X=1	n_{11}	n_{10}	$n_{1.}$ fixed
X=0	n_{01}	n_{00}	$n_{0.}$ fixed
Total	$n_{.1}$	$n_{.0}$	$\mathbf{n}$

CASE-CONTROL DESIGN: randomly select $n_{.1}$ from N_1 cases and $n_{.0}$ from N_0 controls

	D		
X	D=1	D=0	total
X=1	n_{11}	n_{10}	$n_{1.}$
X=0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$ fixed	$n_{.0}$ fixed	$\mathbf{n}$

Review: Example

Consider a hypothetical study of the association between maternal age and birth weight using data from 1000 hospital delivery records.

We can use either of the three designs discussed above.

Let X=I(maternal age<=20 yrs) and D=I(birth weight <=2500 g), Where I is an indicator function

CROSS-SECTIONAL DESIGN: randomly select 200 from the 1000 records

	D		
X	D=1	D=0	total
X=1	10	40	50
X=0	15	135	150
Total	25	175	200

Review: Example

PROSPECTIVE DESIGN:randomly select a 100 pregnant women age $<= 20$ and 100 age > 20

	D		
X	D=1	D=0	total
X=1	20	80	100
X=0	10	90	100
Total	30	170	200

CASE-CONTROL DESIGN: Randomly select 100 infants with birth weight $<= 2500g$ and 100 with birth weight $> 2500g$

	D		
X	D=1	D=0	total
X=1	40	23	63
X=0	60	77	137
Total	100	100	200

Very Important Observation

We can measure the association between X and D using Ratio of Proportions

$$PR = \frac{Pr(D = 1/X = 1)}{Pr(D = 1|X = 0)}$$

Or using ratio of Odds

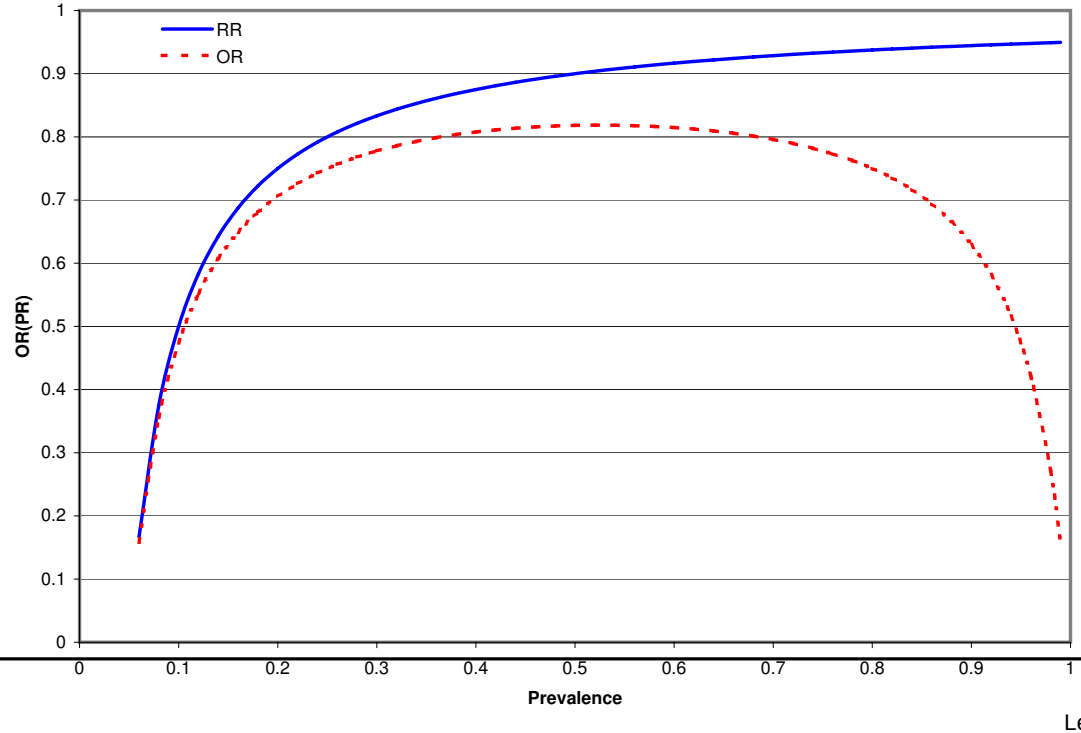
$$OR = \frac{Pr(D = 1/X = 1)/Pr(D = 0/X = 1)}{Pr(D = 1|X = 0)/Pr(D = 0/X = 0)} = \frac{n_{11} * n_{00}}{n_{01} * n_{10}}$$

Measures of Association

Design	Pr(D=1/X=1)	Pr(D=1/X=0)	Pr(X=1/D=1)	Pr(X=1/D=0)	PR	OR
Cross-sectional	10/50=0.2	15/150=0.1	10/25=0.4	40/175=0.23	2.0	2.25
Prospective	20/100=0.2	10/100=0.1	NA	NA	2.0	2.25
Case-control	NA	NA	40/100=0.4	23/100=0.23	NA	2.25

Very Important Observation

- The difference between the OR and PR grows with $Pr(D = 1/X)$
- The bottom line is, in cross-sectional studies, DO NOT use OR when the disease is common ($Pr(D = 1/X) > 10\%$)



Randomization, Stratification and Matching

Usually investigators are interested to find out the net effect of a certain risk factor controlling for confounding and effect modifying factors.

For example to control for age, race and gender differences (if they are not the main factors under consideration)

These can be done by using : Randomization, Stratification and/or Matching

- Randomization: an intervention at the design stage to balance the groups under comparison on factors that are potential confounders of the relationship between X and D
- Stratification: can be done at the design stage or at the analysis stage. It is also used to control for potential confounders of the relationship between X and D
- Matching: it could be **1:m matching or group matching**. It is done at the design stage. It is used to control for potential confounders of the relationship between X and D

Multiple Logistic Regression

- **Problem:** It is likely that the outcome variable will be determined not by a single predictor variable, but by many variables.

- **Goal:** To consider the simultaneous influence of several variables on the response. This will help to reveal the relationships that may have been hidden during the univariate analysis.

- Suppose we have p variables (nominal, ordinal or continuous) that are measured on n individuals, the **Data Layout:** for the n subjects is

Subject i	X_1	X_2	.	.	.	X_p
1	X_{11}	X_{21}	.	.	.	X_{p1}
2	X_{12}	X_{22}	.	.	.	X_{p2}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	X_{1n}	X_{2n}	.	.	.	X_{pn}

The Model

For $E(D|X) = \mu_{D|X}$, where D is the disease indicator and X is the exposure

$$\text{logit}(\mu_{D|X}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

In Matrix notation, this can be re written as: $\text{logit}(\mu_{D|X}) = \beta' \mathbf{X}$

Where $\beta' = [\beta_0, ..., \beta_p]$ and $X' = [X_1, ..., X_p]$

This model can be used for different purposes:

- to estimate an adjusted effect of X_i and D controlling for confounding factors ($X_{j \neq i}$) (Eg. Effect of condom use on STD adjusting for number of partners)
- to assess or investigate interaction or effect modification (Eg. Effect of seat belt use on fatality for speeders and non-speeders)
- to obtain the best prediction model (Eg. Gail et al. JNCI 81:1879-86, 1989: present a prediction model for a White Woman's risk of breast cancer based on Age at menarche, number of previous biopsies, age at first live birth, number of first-degree relatives with breast cancer)

Output from a typical regression package

A computer output from a typical regression package will contain:

1. $\widehat{\beta}_0$ which is the effect when all X's are zero
2. $\widehat{\beta}_i$ which is the effect of X_i controlling all other X_js to be same
3. an overall test of $H_0 : \beta_1 = ... = \beta_p$ vs H_1 : Some β_i's are not equal to zero.

   (a) Can be tested using LR test which has Chi-square distribution with p degrees of freedom.

   (b) **Note that rejecting the global null hypothesis means some/all the predictors considered do aid in predicting D or outcome. On the other hand failing to reject it does not imply none of the covariates are important. There can be effect of some covariates masked by others.**

4. a Wald test to assess the significance of each covariate in the model

Example: A two variable model for typical output

```
Analysis of condom use and STD;
proc format;
value condom   0='worn' 1='not worn';
value partners 0='<5'   1='>=5';
value std      0='no'   1='yes';
run;

data std;
input condom partners std repeat;
datalines;
1 0 0 10
1 0 1 5
1 1 0 30
1 1 1 50
0 0 0 52
0 0 1 30
0 1 0 8
0 1 1 15
;
run;
```

Example: Two variable model and typical SAS output

$$\text{logit}(\Pr(\text{std} = 1)) = \beta_0 + \beta_1 \text{Condom} + \beta_2 \text{Partners}$$

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.8913	2	0.0006
Score	14.6987	2	0.0006
Wald	14.3097	2	0.0008

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-0.5522	0.2190	6.3590	0.0117	0.576
condom	1	-0.1281	0.3801	0.1136	0.7361	0.880
partners	1	1.1889	0.3800	9.7910	0.0018	3.284

Estimation and Interpretation of Parameters

- Estimation is done using Maximum Likelihood Methods with Newton Raphson iterative algorithm (there is closed form solution for p=1, binary)

- Interpretation of logit$(\Pr(D = 1)) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$

  1. β_0 is the log-odds when $X_1 = ... = X_p = 0$

  2. β_1 is the log-odds ratio comparing levels of X_1, LIKE $X = 1$ vs $X = 0$ or for a unit change in X_1 given $X_2, ..., X_p$ are held constant

- In our example: logit$(\Pr(\text{std} = 1)) = -0.5522 + -0.1281\text{Condom} + 1.1889\text{Partners}$

- **Homework** Write the interpretation of the coefficient of Condom use and number of partners

Confounding and Interaction

The first step in multiple logistic regression is to test any **apriori** hypothesis of interaction effect followed by confounding effect. These are the two ways an extraneous variable may affect the relationship between outcome and exposure

Interaction : exists when the relationship between two variables is different for different levels of a third variable. It is also called **effect modification**. **For example in the Nurses Health Study: is the association between breast cancer and oral contraceptive use different in women of age 30-39 and women of age 40-55?**

Confounding: exists when the estimated relationship of interest changes when we add a third variable. **For example in the STD and condom use study, is the association between STD and condom use over-estimated because of the relationship between STD and number of partners?**

In general, the basic questions to consider (Breslow and Day, Vol I, 1980) are:

- the degree of association between risk for disease and the factors under study
- the extent to which the observed association may result from bias, confounding and/or chance

Effect Modification: Example1

Consider the Nurses Health Study: the association between breast cancer and oral contraceptive use in women of age 30-39 and women of age 40-55

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	105.6737	3	<.0001
Score	95.0990	3	<.0001
Wald	87.4061	3	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Exp(Est)
Intercept	1	-5.8592	0.1693	1198.2508	<.0001	0.003
OC	1	-0.1326	0.2068	0.4109	0.5215	0.876
Age	1	0.9291	0.1783	27.1581	<.0001	2.532
OC*Age	1	0.0895	0.2301	0.1513	0.6973	1.094

Effect Modification: Example1

logit$(\Pr(\text{Bcancer} = 1)) = -5.8592 + -0.1326\text{OC} + 0.9291\text{Age} + 0.0895\text{OC} * \text{Age}$

- For Age=30-39: OR(OC vs non OC) = exp(-0.1326)=0.88
- For Age=40-55: OR(OC vs non OC) = exp(-0.1326+0.0895)=0.96

- Test of $H_0 = \beta_{11} = \beta_{12}$ has p-value=0.6973
- P-VALUE for INTERACTION IS NOT SIGNIFICANT
- Usually TEST interaction at >5% level of significance....WHY 10%? Unless the power and sample size at design stage account for interaction.

Effect Modification: Example2

Consider the example on the effect of wearing seat belt on accident fatality

	Driver			
Seat Belt	Dead	Alive	total	fatality rate
Not worn	20	30	50	20/40=40%
Worn	10	40	50	10/50=20%
Total	30	70	100	

How about if we stratified by impact speed? Effect on inference?

		Driver			
Speed	**Seat Belt**	Dead	Alive	total	fatality rate
$\leq$ 40Mph	Not worn	2	18	20	2/20=10%
$\leq$ 40Mph	Worn	3	27	30	3/30=10%
> 40Mph	Not worn	18	12	30	18/30=60%
> 40Mph	Worn	7	13	20	7/20=35%

Stratification: Example data in SAS

```
* Analysis of traffic safety data;
proc format;
value seatbelt 0='worn' 1='not worn';
value speed 0='<45Mph' 1='>=45Mph';
value deadalive 0='alive' 1='dead';
run;

data saftey;
input seatbelt speed deadalive repeat;
datalines;
1 0 0 18
1 0 1 2
1 1 0 12
1 1 1 18
0 0 0 27
0 0 1 3
0 1 0 13
0 1 1 7
;
run;
```

Stratified Logistic regression in SAS: Pooled

Using the above data structure we can have a crude measure of the effect of seatbelt on fatality

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq	Exp(Est)
Intercept	1	-1.3863	0.3536	15.3745	<.0001	0.250
seatbelt	1	0.9808	0.4564	4.6177	0.0316	2.667

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
seatbelt	2.667	1.090	6.524

Stratified Logistic regression in SAS: stratified

```
---------------------- speed>45mph ------------------------

               Odds Ratio Estimates

                  Point          95% Wald
   Effect      Estimate      Confidence Limits
   seatbelt       2.786       0.861       9.009

---------------------- speed<45mph ------------------------

               Odds Ratio Estimates

                  Point          95% Wald
   Effect      Estimate      Confidence Limits
   seatbelt       1.000       0.152       6.593
```

Confounding

Confounding: the outcome-exposure association is over or under estimated because of the relation of an underlying variable with exposure and outcome

Confounder (C): is therefore a variable which, because of its relationship to outcome (D) and exposure (X), either obscures or enhances the D and X association.

Criteria for a confounding factor: Rothman and Greenland, 1998

- a confounding factor must be a risk factor for disease
- a confounding factor must be associated with X in the population at risk from which the cases are derived. Typically for a rare disease check this condition by looking for an association in the control group
- a confounding factor must not be affected by the exposure or the disease. In particular, it can not be an intermediate step in the causal path between the exposure and the disease.

Confounding: Important facts

1. C must be related to X in the data

2. C must be independently causally related to D in the population

3. if C lies in the causal pathway between X and D, then it is not confounder

   (Eg. Menstrual estrogen$\rightarrow$ Endometrial hyperplasia$\rightarrow$ endometrial cancer) Endometrial hyperplasia is a benign condition that occurs when the lining of the uterus or endometrium grows too much.

4. if C is caused by both X and D, then it is not confounder

   (Eg. Vaginal bleeding is caused by both menopausal estrogens and endometrial cancer)

SAS code for Logistic regression on STD and condom use

Lets observe how the effect of condom use (X) changes when adjusted for a possible confounder number of partners (C) :

```
proc logistic data=std desc;
model std=condom/expb;
weight repeat;
run;

proc logistic data=std desc;
model std=condom partners/expb;
weight repeat;
run;
```

SAS output for Logistic regression

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
condom	1.833	1.046	3.214

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
condom	0.880	0.418	1.853
partners	3.284	1.559	6.915

Determining Confounding

The change in the Odds Ratio of condom use is (1.833-0.880)/1.833 = 53%. Which shows a huge confounding by number of partners!

In general the **rule of thumb** is to see for 20% or more change in the Odds Ratio.

Other guidelines to choose a confounder for adjustment are to report several planned analysis of:

- Unadjusted effect
- Adjusted for primary set of covariates (known risk factors)
- Adjusted for primary and secondary set of covariates (known and suspected risk factors)

HOMEWORK: Check if Age confounds the relationship between OC use and Breast cancer in the Nurses health study.

Effect of Omitted Variables-I (hypothetical data)

Samuels (1981)... Biometrika 68:577-88 and Gail et al (1984)...Biometrika 71:431-44

Consider two models: Model 1=logit$(\Pr(D=1)) = \beta_0 + \beta^* X$

Model 2=logit$(\Pr(D=1)) = \beta_0 + \beta X + \eta Z$, where Z is the omitted variable

1. Z is associated with X both marginally and among cases and controls AND Z is associated with D both marginally and among $X=1$ and $X=0$

	Z=1			**Z=0**			**total**		
X	D=1	D=0	total	D=1	D=0	total	D=1	D=0	total
X=1	50	10	60	2	10	12	52	20	72
X=0	10	2	12	10	50	60	20	52	72
Total	60	12		12	60		72	72	

2. Model 1: logit$(p) = -0.956 + 1.911X (SE = 0.372)$
3. Model 2: logit$(p) = -1.609 + 3.219Z + 0.000X (SE = 0.600)$
4. If Z is a confounder (clinically) then $\widehat{\beta}^*$ is a confounded estimate of β AND test of $\beta^* = 0$ is wrong.

Effect of Omitted Variables-II

1. Z is associated with X both marginally and among cases and controls AND Z is NOT associated with D marginally, but is associated among $X = 1$ and $X = 0$ (Eg. a matching variable, i.e we have made D=0 and D=1 groups same on Z)

	Z=1			**Z=0**			**total**		
X	D=1	D=0	total	D=1	D=0	total	D=1	D=0	total
X=1	30	10	40	50	30	80	80	40	120
X=0	30	50	80	10	30	40	40	80	120
Total	60	60		60	60		120	120	

2. Model 1: logit$(p) = -0.693 + 1.386X (SE = 0.274)$
3. Model 2: logit$(p) = -1.099 + 0.588Z + 1.609X (SE = 0.306)$
4. Z is a confounder (clinically) and hence $\widehat{\beta}^*$ is a confounded estimate of β AND test of $\beta^* = 0$ is wrong.

Effect of Omitted Variables-III

1. Z is NOT associated with X marginally, but is associated among cases and controls AND Z is associated with D both marginally and among $X = 1$ and $X = 0$

	Z=1			Z=0			total		
X	D=1	D=0	total	D=1	D=0	total	D=1	D=0	total
X=1	30	30	60	50	10	60	80	40	120
X=0	10	50	60	30	30	60	40	80	120
Total	40	80		80	40		120	120	

2. Model 1: logit$(p) = -0.693 + 1.386X (SE = 0.274)$
3. Model 2: logit$(p) = 0.000 + -1.609Z + 1.609X (SE = 0.305)$
4. Z is a confounder (clinically) and hence $\widehat{\beta}^*$ is a confounded estimate of β AND test of $\beta^* = 0$ is wrong.

Effect of Omitted Variables-IV

1. Z is associated with X both marginally and among cases and controls AND Z is associated with D marginally but not among $X = 1$ and $X = 0$

	Z=1			**Z=0**			**total**		
X	D=1	D=0	total	D=1	D=0	total	D=1	D=0	total
X=1	50	20	70	5	2	7	55	22	77
X=0	2	5	7	20	50	70	22	55	77
Total	52	25		25	52		77	77	

2. Model 1: logit$(p) = -0.916 + 1.833X (SE = 0.357)$
3. Model 2: logit$(p) = -0.916 + 0.000Z + 1.833X (SE = 0.620)$
4. Z is NOT a confounder and hence $\widehat{\beta}^*$ is a TRUE estimate of β BUT SE could get inflated.

Effect of Omitted Variables-V

1. Z is associated with X marginally but not among cases and controls AND Z is associated with D both marginally and among $X = 1$ and $X = 0$

	Z=1			**Z=0**			**total**		
X	D=1	D=0	total	D=1	D=0	total	D=1	D=0	total
X=1	50	2	52	5	20	25	55	22	77
X=0	20	5	7	2	50	52	22	55	77
Total	70	7		7	70		77	77	

2. Model 1: logit$(p) = -0.916 + 1.833X (SE = 0.357)$
3. Model 2: logit$(p) = -0.916 + 4.605Z + 1.833X (SE = 0.620)$
4. Z is NOT a confounder and hence $\widehat{\beta}^*$ is a TRUE estimate of β BUT SE could get inflated. If Z is clinically a potential confounder, adjust and report the implication on the precision of the estimate.

Effect of Omitted Variables-VI

1. Z is NOT associated with X either marginally or among cases and controls AND Z is NOT associated with D both marginally and among $X = 1$ and $X = 0$

	Z=1			Z=0			total		
X	D=1	D=0	total	D=1	D=0	total	D=1	D=0	total
X=1	30	10	40	3	1	4	33	11	44
X=0	30	50	80	3	5	8	33	55	88
Total	60	60		6	6		66	66	

2. Model 1: logit$(p) = 0.511 + 1.609X (SE = 0.412)$
3. Model 2: logit$(p) = -0.511 + -0.000Z + 1.609X (SE = 0.412)$
4. Z is NOT a confounder and hence $\widehat{\beta}^*$ is a TRUE estimate of β. Even if Z is clinically a potential confounder, it does not matter.

Strategies for Model Building-I

Two main purposes for fitting regression models are:

1. to estimate the underlying relationship of D to a set of risk factors/independent variables (Association)

2. to predict D using a set of risk factors/independent variables (Prediction)

In model building:

- We are interested in finding the BEST model within the scientific context of the problem.
- finding “best” involves a joint effort — statistical knowledge, careful thought, experience and common sense

- aim to get a parsimonious model—easy to understand and interpret, the more variables the bigger the standard errors=more instable the model
- So we need to have **variable selection plan AND methods of assessing the adequacy of a model**

Strategies for Model Building-II

In Association studies

1. all variables that are clinically thought to be confounders should be studied by including them in the model
2. if control of a variable ALTERS to an important degree the estimate of the OR or its standard error, the variable should be included in the model regardless of whether its coefficient is significant or not.
3. If it does not ALTER the coefficient, make decision based on–believability of results, statistical significance of the variable
4. interaction and confounding are very important

In Prediction studies,

- predictive models are built primarily based on statistical decision making
- if a variable increases the over all log-likelihood significantly, it will be included
- interaction is very important, but confounding is not

Strategies for Model Building-IIIa

- start with careful univariate analysis and screen candidate variables at pvalue=0.25 (Hosmer and Lemeshow) and build a multivariable model.

- Why larger p-value than the traditional?
  So that you do not miss those variables weakly associated with outcome but can become important when taken together

  1. contingency table analysis for discrete and simple univariate logistic or t-test for continuous variables

  2. look for zero cells which can lead to OR=infinity or zero in logistic regression.

  3. The strategies for dealing with zero cells include:
     (a) collapsing the categories in some sensible fashion
     (b) eliminating the zero category
     (c) for ordinal variables model as continuous

Strategies for Model Building-IIIb

OR use stepwise methods (mechanical selection methods)–importance of a variable is determined by “significance”. But, unless the problem is new, it is not recommended

1. Forward selection – Start with a small model and keep adding. It can lead to “too many” significant findings. So, it is better to start from a smaller α

2. Backward elimination – Start with a full model and drop variables. Has better control of type-I error rate. BUT very difficult for large number of variables

3. Best-subset selection –gives “Best” subset of models (one variable, two variable,...)

Strategies for Model Building-IV

- after fitting a multivariable model then verify the importance of each variable via the Wald test and the difference in the coefficient with the univariate model containing only that variable.
- continue until you get the **main effects model**. Then check if the logit is linear for the continuous variables (plot the logit against the covariate).
- after the main effects model, look for any interaction (statistical test, clinical sense). Typically, an interaction term that is significant would alter both the point and interval estimates. Then you get **final model**
- Note that:
  1. you can not interpret the exclusion of a variable from the model as lack of relationship
  2. it is not safe to interpret inclusion in the model as indication of a direct relationship
  3. Pay attention to the number of variables considered viz the sample size (10 for each variable)

Data Example for stepwise, forward and backward methods

SIZE: 116 observations (29 cases, 87 controls), 9 variables

LIST OF VARIABLES:

Variable	Abbreviation
Stratum Number	STRATUM
Observation Type (1 = Case, 2,3,4 = Controls)	OBS
Age of the Mother in Years	AGE
Low Birth Weight (0 = Birth Weight ge 2500g, l = Birth Weight < 2500g)	LOW
Weight in Pounds at the Last Menstrual Period	LWT
Smoking Status During Pregnancy (1 = Yes, 0 = No)	SMOKE
History of Hypertension (1 = Yes, 0 = No)	HT
Presence of Uterine Irritability (1 = Yes, 0 = No)	UI
History of Premature Labor (0 = None, 1 = Yes)	PTD

SAS code for stepwise, forward and backward methods

```
title 'Forward Selection on Low birth Weight Data';
   proc logistic data=library.lowbwt13;
      model low=age lwt smoke ptd ht ui/ selection=backward slstay=0.2 ctab
   run;


title 'Backward Elimination on Low birth Weight Data';
   proc logistic data=library.lowbwt13;
      model low=age lwt smoke ptd ht ui/ selection=backward fast slstay=0.2
   run;


title 'Stepwise Regression on Low birth Weight Data';
   proc logistic data=library.lowbwt13 desc outest=betas covout;
      model low=age lwt smoke ptd ht ui/ selection=stepwise slentry=0.3
      slstay=0.35 details lackfit;
      output out=pred p=phat lower=lcl upper=ucl predprob=(individual cross
   run;
```

Summary of the stepwise method

- SLENTRY=0.3 implies a significance level of 0.3 is required to allow a variable into the model
- SLSTAY=0.35 implies a significance level of 0.35 is required for a variable to stay in the model.
- A detailed account of the variable selection process is requested by specifying the DETAILS option.
- In stepwise selection, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model.
- the FAST option operates only on backward elimination steps. In this example, the stepwise process only adds variables, so the FAST option would not be useful

Model fitting strategies: Example

Step 0. Intercept entered:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0986	0.2144	26.2511	<.0001

The LOGISTIC Procedure

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
18.3672	6	0.0054

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
AGE	1	0.0000	1.0000
LWT	1	2.4562	0.1171
SMOKE	1	5.2581	0.0218
PTD	1	14.8178	0.0001
HT	1	0.0697	0.7918
UI	1	5.2242	0.0223

Model fitting strategies: Example

Step 1. Effect PTD entered:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4917	0.2609	32.6944	<.0001
PTD	1	1.9436	0.5494	12.5157	0.0004

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
AGE	1	0.1297	0.7187
LWT	1	1.0145	0.3138
SMOKE	1	1.8865	0.1696
HT	1	0.0094	0.9229
UI	1	1.2965	0.2548

Step 2. Effect SMOKE entered:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7458	0.3358	27.0256	<.0001
SMOKE	1	0.6458	0.4745	1.8529	0.1734
PTD	1	1.7389	0.5685	9.3560	0.0022

Model fitting strategies: Example

```
 Analysis of Effects Eligible for Entry
                          Score
  Effect      DF    Chi-Square    Pr > ChiSq
  AGE          1        0.0793        0.7783
  LWT          1        0.9037        0.3418
  HT           1        0.0042        0.9483
  UI           1        1.2472        0.2641

Step  3. Effect UI entered:
  Analysis of Maximum Likelihood Estimates
                                Standard          Wald
 Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

 Intercept     1     -1.8489      0.3551       27.1042        <.0001
 SMOKE         1      0.6418      0.4772        1.8088        0.1787
 PTD           1      1.5463      0.5926        6.8084        0.0091
 UI            1      0.6139      0.5536        1.2296        0.2675

Step  4. No (additional) effects met the 0.3 significance level
         for entry into the model.
```

Summary of the stepwise method

- First, the intercept-only model is fitted and individual score statistics for the potential variables are evaluated.
- In Step 1, variable PTD is selected into the model since it is the most significant variable among those to be chosen($p=0.0001$). Then, the model that contains an intercept and PTD is fitted. PTD remains significant and is not removed.
- In Step 2, variable SMOKE is added to the model with ($p=0.1696 < 0.3$). The model then contains an intercept and variables PTD and SMOKE. PTD remains significant ($p=0.0022$) and SMOKE is also significant at 0.35 level. Therefore, neither PTD or SMOKE is removed from the model.
- In step 3, variable UI is added to the model ($p=0.2641 < 0.3$). The model then contains an intercept and variables PTD, SMOKE, and UI. None of these variables are removed from the model since all are significant at the 0.35 level.
- Finally, none of the remaining variables outside the model meet the entry criterion, and the stepwise selection is terminated.