Lecture 19: 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

- *Model Fitting Strategies*
- Goodness of Fit and Model Diagnostics
- matching (group and individual)
- Conditional vs Unconditional analysis
- Methods III: Advanced Regression Methods

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,...)

4. Stepwise regression – combination of the forward and backward methods

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

Note: since data is individually matched, the correct analysis is using conditional logistic

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
                                         slentry=0.2 ctable;
   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 ctable;
   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.

Goodness of Fit and Model Diagnostics-I

- assuming that we are preliminarily satisfied with the final model (model contains main and interaction effects in their correct functional form)

- the objective is to look at how closely model fitted responses approximate observed responses

- measure statistics are usually based on observed ($d_1, ...d_n$) and fitted ($\hat{d}_1, ...\hat{d}_n$) value differences
  1. Overall measures of fit: Deviance, Pearson Chi-square, Hosmer-Lemeshow test

  2. Detecting influential observations: Residual analysis, Plot of residuals, influence statistics

Goodness of Fit and Model Diagnostics-II

- To assess model performance, we must evaluate the fit of the model under different covariate patterns (which is a set of values for the covariates in the model)

  Example: for low-birth-weight final model, 8 patterns: 000 001 010 011 100 101 110 111

  Example: if age was included, the covariate pattern could be as large as n

- SAS computes predicted values and residuals for each each individual and you need to aggregate your data by covariate pattern. You can do this by using scale=none and aggregate=(smoke ui ptd) in the model options.

Pearson Chisquare and Deviance Statistic

Summary statistics for goodness of fit and diagnostics are based on differences between observed and fitted values.

J = the number of distinct covariate patterns (<=n)
m_j = the number of subjects with the jth covariate pattern
d_j = the number of $d = 1$ in the jth pattern
$\hat{\pi}_j$ = estimated Pr(D=1|X) from logistic regression for the jth pattern

Pearson residual = $r_j = \frac{d_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$

Deviance residual = $l_j = \pm\sqrt{2[d_j \log\frac{d_j}{m_j\hat{\pi}_j} + (m_j - d_j)\log\left(\frac{(m_j-d_j)}{m_j(1-\hat{\pi}_j)}\right)]}$

Pearson chisquare = $\sum_{j=1}^{J} r_j^2$ is chisquare with df=J - number of parms in model

Deviance Statistic = $\sum_{j=1}^{J} l_j^2$ is chisquare with df=J - number of parms in model

Note: if $J < n$ then both have chisquare distributions as $m_i \to \infty$
if $J \approx n$ then use the Hosmer-Lemeshow test

Hosmer-Lemeshow Test

- Collapse the observed/expected table into G (usually 10) groups based on the percentiles of the estimated probabilities (lower 10th percentile,...,to upper 90th percentile)

- OR Collapse the observed/expected table into 10 groups based on fixed values of the estimated probabilities (0-9percent,...,90-100 percent)

  $\hat{C} = \sum_{g=1}^{10} \frac{(O_g - n_g \bar{\pi}_g)^2}{n_g \bar{\pi}_g (1 - \bar{\pi}_g)}$

  where n_g number of subjects in each group, O_g number of observed $d = 1$ in each group

- $\hat{C}$ follows a chi-square distribution with df=$10 - 2$
- The Hosmer and Lemeshow goodness-of-fit test for the final selected model is requested by specifying the LACKFIT option in SAS.

Hosmer-Lemeshow Test–Example

Partition for the Hosmer and Lemeshow Test

		LOW = 0		LOW = 1	
Group	Total	Observed	Expected	Observed	Expected
1	13	4	4.58	9	8.42
2	11	8	6.28	3	4.72
3	28	21	21.55	7	6.45
4	8	4	6.20	4	1.80
5	56	50	48.38	6	7.62

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.1253	3	0.1628

There is no evidence to reject H0, so the fit is Good???.

Residual Statistic Options in SAS

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	13.4227	4	3.3557	0.0094
Pearson	12.0334	4	3.0084	0.0171

Number of unique profiles: 8

Note: When J«n these two tests are more appropriate tests of goodness of fit

- H=name...specifies the diagonal element of the hat matrix for detecting extreme points in the design space.

- RESCHI=name...specifies the Pearson (Chi) residual for identifying observations that are poorly accounted for by the model.

- RESDEV=name...specifies the deviance residual for identifying poorly fitted observations.

Diagnostic Tests-I

The Next important step in model building is to perform an analysis of residuals and diagnostic statistics to study the influence of observations.

To do this we need the h statistic in addition to the residuals (r_j and l_j).

The h statistic for each covariate pattern is obtained from the hat matrix,

$$H = V^{1/2}X(X'VX)^{-1}X'V^{1/2}$$

where V is the JxJ diagonal matrix with elements $v_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)]$.

The diagnostic measures we consider here are based on deletion of subjects with the jth covariate pattern

Diagnostic Tests-II

- Influence on parameter estimates is studied by the DFBETA

  $\triangle\hat{\beta}_j = \frac{r_j^2 h_j}{(1-h_j)^2}$

- Influence on the overall significance is studied by change in the chisquare OR

  $\triangle\chi^2{}_j = \frac{r_j^2}{(1-h_j)}$

- change in Deviance

  $\triangle D_j = \frac{l_j^2}{(1-h_j)}$

- Change in chisquare and Deviance help to identify poorly fit covariate patterns
  In logistic regression, diagnostics are interpreted by visual assessment.
- plot of $\triangle\chi^2{}_j$ versus $\hat{\pi}_j$

- plot of $\triangle D_j$ versus $\hat{\pi}_j$ AND plot of $\triangle\hat{\beta}_j$ versus $\hat{\pi}_j$

Model Diagnostics-Example

- if you identify an observation or set of observation which have an influence on one or more of the three diagnostic statistics, then investigate the observations one by one to see what is going on

- you might also want to see if removing the bad behaving observation(s) brings any change in the goodness of fit statistics (Hosmer-Lemeshow C statistic)

In SAS, you can output diagnostic statistics using key words,

- C=name...specifies the confidence interval displacement diagnostic that measures the influence of individual observations on the regression estimates.
- CBAR=name...specifies the another confidence interval displacement diagnostic, which measures the overall change in the global regression estimates due to deleting an individual observation.

Diagnostic Statistic Options in SAS

- DFBETAS= $\_ALL\_$ and DFBETAS=var-list...specifies the standardized differences in the regression estimates for assessing the effects of individual observations on the estimated regression parameters in the fitted model. You can specify a list of the explanatory variables in the MODEL statement, or you can specify just the keyword _ALL_. In the former specification, the first variable contains the standardized differences in the intercept estimate, the second variable contains the standardized differences in the parameter estimate for the first explanatory variable in the MODEL statement, and so on. In the latter specification, the DFBETAS statistics are named DFBETA_xxx, where xxx is the name of the regression parameter. For example, if the model contains two variables X1 and X2, the specification DFBETAS=_ALL_ produces three DFBETAS statistics: DFBETA_Intercept, DFBETA_X1, and DFBETA_X2. If an explanatory variable is not included in the final model, the corresponding output variable named in DFBETAS=var-list contains missing values.
- DIFCHISQ=name...specifies the change in the chi-square goodness-of-fit statistic attributable to deleting the individual observation.
- DIFDEV=name...specifies the change in the deviance attributable to deleting the individual observation.

Model Diagnostics-Example

Diagnostics on the final model for the low birth weight data given above

```
 title 'Stepwise Regression on Low birth Weight Data';
 proc logistic data=library.lowbwt13;
  model low=smoke ptd ui/lackfit influence iplots scale=none aggregate=(smo
  output out=pred p=phat h=hat reschi=chires resdev=devres c=csmoke cbar=cb
  dfbetas=_all_ difchisq=difchi difdev= difdev lower=lcl upper=ucl
  run;
```

				Regression Diagnostics				
	Covariates					Hat		
Case				Pearson	Deviance	Matrix	Intercept	SMOK
Number	SMOKE	PTD	UI	Residual	Residual	Diagonal	DfBeta	DfBet
1	0	0	0	-2.5205	-1.9975	0.0148	-0.3115	0.185
2	0	0	0	0.3968	0.5407	0.0148	0.0490	-0.029
3	1	0	0	0.5469	0.7234	0.0269	0.0168	0.062
4	0	0	0	0.3968	0.5407	0.0148	0.0490	-0.029
5	1	1	1	-0.6209	-0.8076	0.0700	0.0518	-0.042
6	0	0	0	0.3968	0.5407	0.0148	0.0490	-0.029
7	1	0	0	0.5469	0.7234	0.0269	0.0168	0.062
.								
.								

Model Diagnostics example

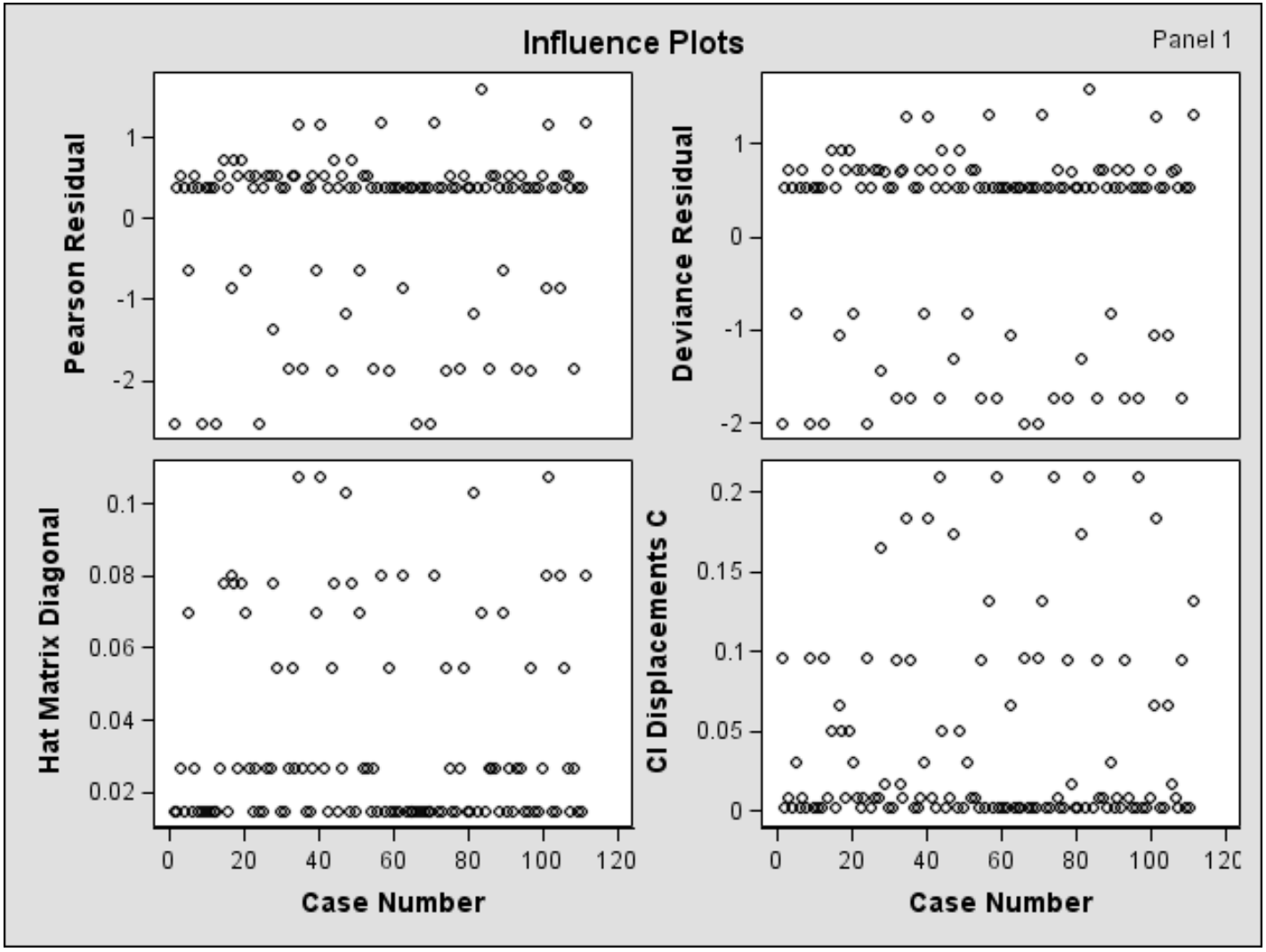
- diagnostics on the final model for the low birth weight data given above

```
title 'Stepwise Regression on Low birth Weight Data';
 proc logistic data=library.lowbwt13;
  model low=smoke ptd ui/lackfit influence iplots scale=none aggregate
  output out=pred p=phat h=hat reschi=chires resdev=devres c=csmoke cb
  dfbetas=_all_ difchisq=difchi difdev= difdev lower=lcl upper=ucl
  run;
```

Case Number	UI DfBeta	Confidence Interval Displacement C	Confidence Interval Displacement CBar	Delta Deviance	Delta Chi-Squar
1	0.0945	0.0970	0.0956	4.0857	6.4483
2	-0.0149	0.00240	0.00237	0.2947	0.1598
3	-0.0252	0.00850	0.00827	0.5316	0.3073
4	-0.0149	0.00240	0.00237	0.2947	0.1598
5	-0.0880	0.0312	0.0290	0.6812	0.4146
6	-0.0149	0.00240	0.00237	0.2947	0.1598
7	-0.0252	0.00850	0.00827	0.5316	0.3073

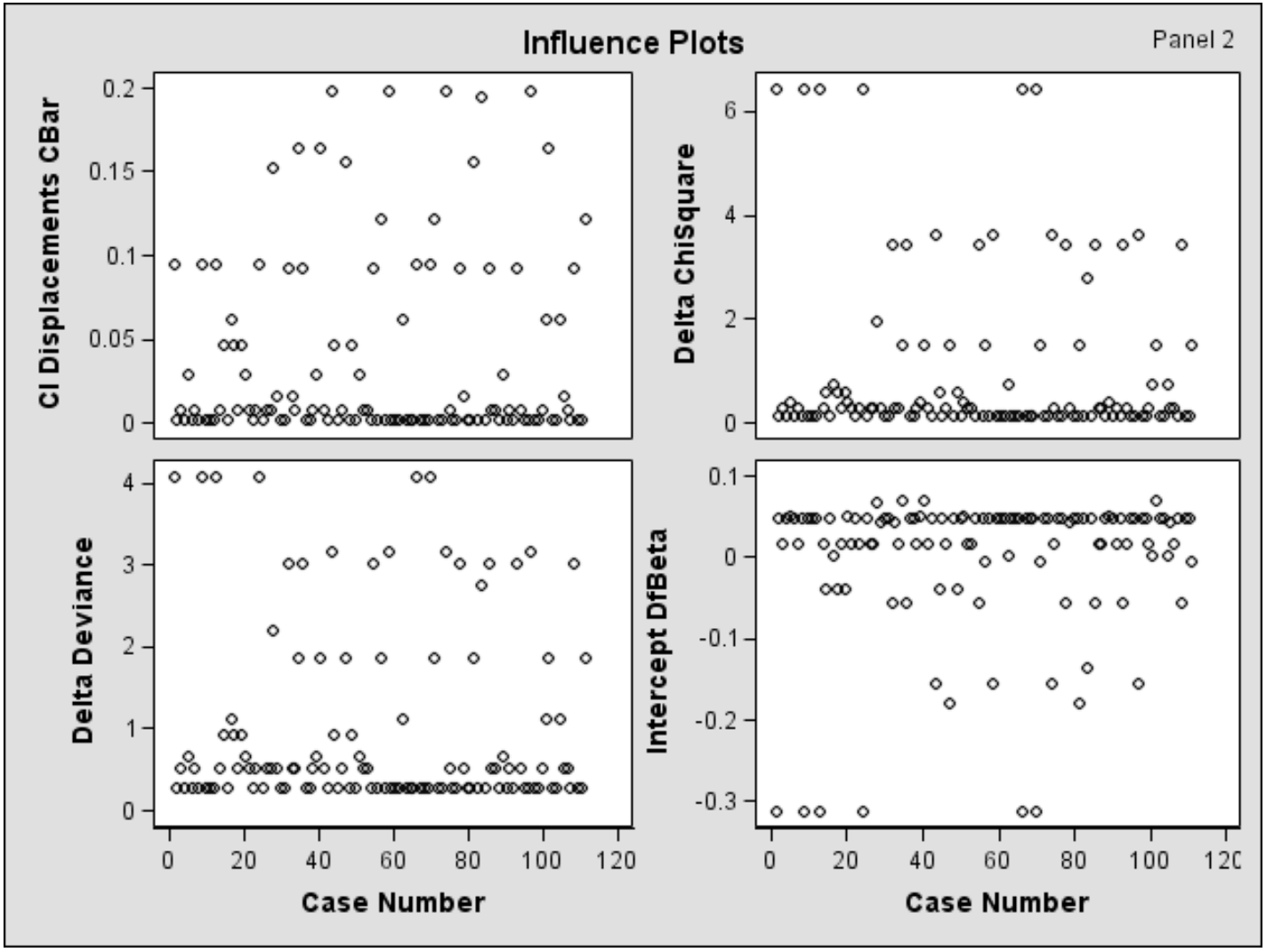
Model Diagnostics example

Diagnostics on the final model for the low birth weight data given above



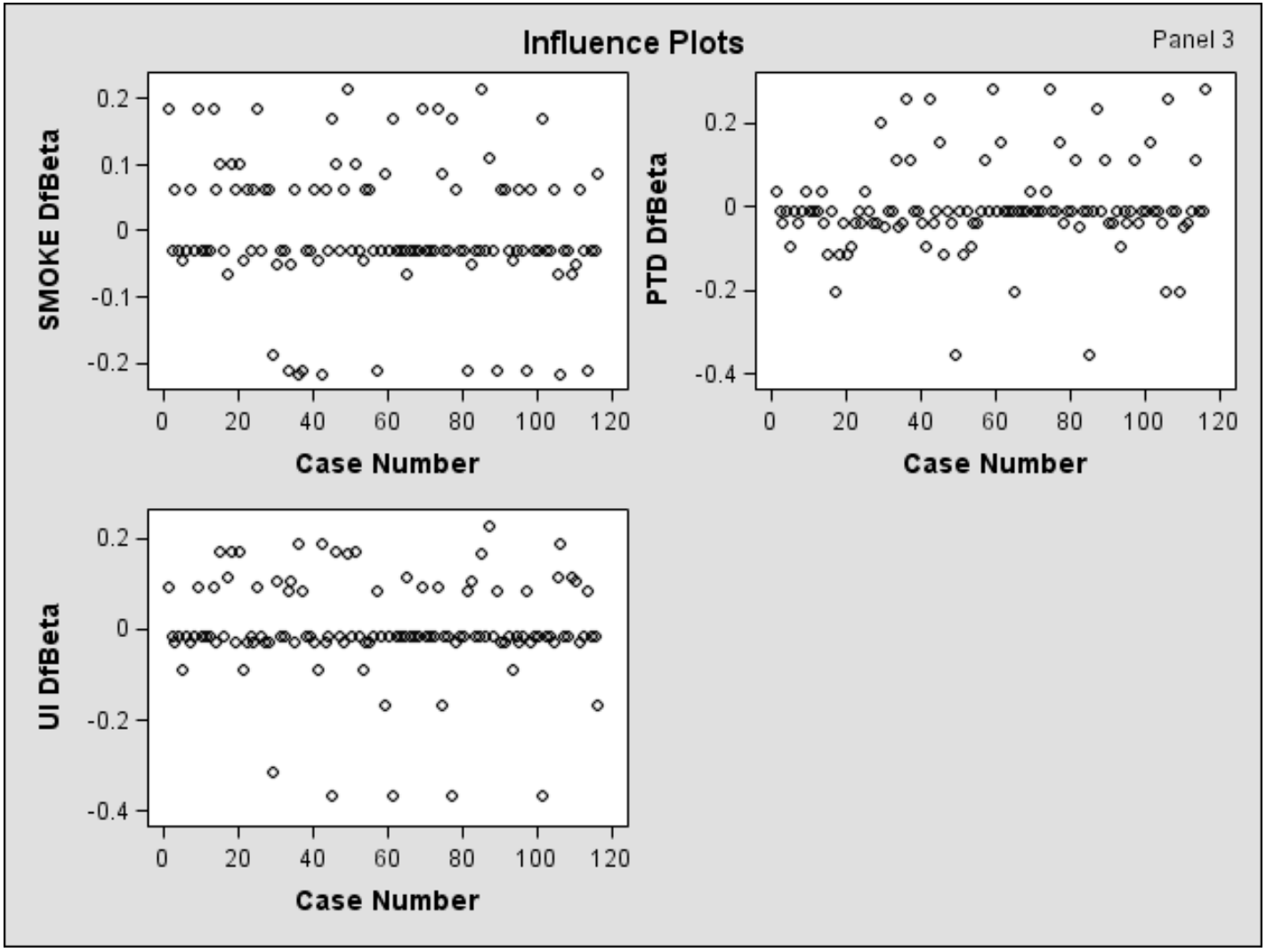
Model Diagnostics example

Diagnostics on the final model for the low birth weight data given above



Model Diagnostics example

Diagnostics on the final model for the low birth weight data given above



Matching

The main objective of matching is to make the comparison groups same on everything except the variable of interest

First lets consider paired binary data (eg. data that comes from pre and post treatment, two eyes, twins)

Consider a hypothetical data of 595 subjects pre and post intervention and evaluated for their outcome

	outcome		
x	D=1	D=0	total
pre trt	166	429	595
post trt	276	319	595
Total	442	748	1190

Can we test change in outcome (H0: Pr(D=1/pre trt)=Pr(D=1/post trt)) using a Chi-square test? NO, because the Chi-square test assumes the rows are INDEPENDENT samples, but we have the same people pre and post intervention.

Paired Match Data

As in paired t test, it is required to analyze the data differently as follows:

	post	
pre	D=0	D=1
D=0	n_{00}	n_{01}
D=1	n_{10}	n_{11}

Paired Match Data

- The concordant pairs (n_{00} and n_{11}) do not contribute any information about the effect of X or the intervention.

- So, we use the information in the discordant pairs (n_{01} and n_{10}) to measure treatment effect

- The appropriate test for (H0: Pr(D=1/pre trt)=Pr(D=1/post trt)) is the McNemar's test
- Under H0, we expect equal change from 0 to 1 and from 1 to 0, i.e $E(n_{10}) = E(n_{01})$. So, under the null,
  $n_{10}|(n_{01} + n_{10})$ is Binomial$(n_{01} + n_{10}, 1/2)$

$$Z = \frac{n_{10} - E(n_{10})}{\sqrt{((n_{01}+n_{10})1/2(1-1/2))}} \sim N(0,1)$$

Z^2 is approximately distributed χ^2(1)

Paired Binary data analysis

The MLE for the odds ratio comparing pre and post trt groups is

$$\text{OR}=\frac{n_{01}}{n_{10}}$$

For the example data we considered above,

	post	
pre	D=0	D=1
D=0	251	178
D=1	68	98

Hypothesis: H0: Pr(D=1/pre trt)=Pr(D=1/post trt))

$$Z = \frac{n_{10}-E(n_{10})}{\sqrt{((n_{01}+n_{10})1/2(1-1/2))}} \sim N(0,1)$$

$= \frac{178-(178+68)/2}{\sqrt{(178+68)/4}}$=7.01 leads to p-value<0.001

OR=178/68=2.62

SAS results (Corrected!)

```
*McNemar's test in Proc FREQ;

Data pairedbinary;
input pre post repeat;
datalines;
0 0 251
0 1 178
1 0 68
1 1 98
;
run;
proc freq data=pairedbinary order=data;
table pre*post/agree cl; * we can use exact mcnem;
weight repeat;
run;
     McNemar's Test

 Statistic (S)    49.1870
 DF                     1
 Pr > S            <.0001
```

In STATA, this is easily done by mcci 251 178 68 98

Matching in case-control studies

- It can be “individually matched” where one case is matched to one or more controls or “group matched” where two or more cases are matched to one or more controls
- we focus on the commonly used design, one case with one to five controls are matched
- The most commonly used design is 1:1 matched
- once we match on certain factors, we are forfeiting estimating their effect
- So, they are nuisance parameters in the model

Conditional vs Unconditional Logistic Likelihood

The model for a matched data with $k = 1, ..., K$ strata is

$$\text{logit}(\pi_k(X) = \alpha_k + \beta_1 X_1 + ... + \beta_p X_p$$

Where $\pi_k(X) = \Pr(D_{ik} = 1|X)$, α_k is log-odds in the kth stratum

- unless the number of subjects in each stratum is large, fitting these models using the unconditional ML does not work well
- if we use to do fully stratified analysis, we end up with $p + K$ parameters to estimate using $n = n_1 + ... + n_K$ samples. For 1:1 matching using 2n pairs.
- in individually matched there is only one case in each stratum and hence we need some way of getting rid of the nuisance parameters
- Conditional likelihood - condition on a sufficient statistic for the nuisance parameter
- the sufficient statistic for α_k is the total number of cases observed in stratum k
- so the conditional likelihood for the k the stratum is obtained as the probability of the observed data conditional on the stratum total and the number of cases observed

Logistic regression for Matched data

Consider the simplest case, the 1:1 matched design with $k = 1, ..., K$ strata and p covariates

$$\text{logit}(\pi_k(X)) = \alpha_k + \beta' X$$

Where $\pi_k(X) = \Pr(D_{ik} = 1|X)$, α_k is log-odds in the kth stratum

- There are two subjects in each stratum
- assume X_{0k} be the data vector for the control and X_{1k} be the data vector for the case

$$\begin{aligned} L_k(\beta) &= \Pr(D_{1k} = 1|X_{1k}, ncases = 1, n_k = 2) \\ &= \frac{\Pr(D_{1k}=1|X_{1k})}{\Pr(D_{1k}=1|X_{1k})+\Pr(D_{0k}=1|X_{0k})} \\ &= \frac{exp(\alpha_k+\beta' X_{1k})}{exp(\alpha_k+\beta' X_{1k})+exp(\alpha_k+\beta' X_{0k})} \\ &= \frac{exp(\beta' X_{1k})}{exp(\beta' X_{1k})+exp(\beta' X_{0k})} \end{aligned}$$

$$L(\beta) = \prod_{k=1}^{K} L_k$$

- This for binary univariate X results in the same OR reported above.

Data Example for 1:1 matched analysis

SIZE: 112 observations (56 cases, 56 controls), 8 variables

LIST OF VARIABLES:

Variable	Abbreviation
Stratum Number	PAIR
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

Note: since data is individually matched, the correct analysis is using conditional logistic

Conditional logistic in SAS–low birthweight data

```
* full stratified analysis;
proc logistic data=library.lowbwt11 desc;
class pair;
      model low=pair smoke/expb;
* McNemar's test;
proc freq data=library.lowbwt11 desc;
table low*smoke/agree cl;
run;
*conditional logistic;
proc logistic data=library.lowbwt11 desc;
      model low=smoke ptd /expb;
      strata pair;
   run;

title 'Stepwise Regression on Low birth Weight Data';
   proc logistic data=library.lowbwt11 desc outest=betas covout;
      model low=age lwt smoke ptd ht ui/ selection=stepwise slentry=0.3
                     slstay=0.35 details lackfit;
      strata pair;
      output out=pred p=phat lower=lcl upper=ucl dfbeta=\_all\_ h=hat;
   run;
```

Conditional logistic in SAS–low birthweight data

Strata Summary

Response Pattern	LOW 1	LOW 0	Number of Strata	Frequency
1	1	1	56	112

Newton-Raphson Ridge Optimization

Without Parameter Scaling

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	77.632	72.839
SC	77.632	75.557
-2 Log L	77.632	70.839

Conditional logistic in SAS–low birthweight data

The LOGISTIC Procedure
Conditional Analysis

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.7939	1	0.0091
Score	6.5333	1	0.0106
Wald	6.0036	1	0.0143

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
SMOKE	1	1.0116	0.4129	6.0036	0.0143	2.750

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
SMOKE	2.750	1.224	6.177

Conditional logistic in SAS–low birthweight data

The LOGISTIC Procedure
Full Stratified Analysis
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8310	0.3139	7.0067	0.0081
PAIR 1	1	-0.1806	1.5841	0.0130	0.9092
PAIR 2	1	0.8310	1.4238	0.3406	0.5595
.					
.					
PAIR 55	1	-0.1806	1.5841	0.0130	0.9092
SMOKE	1	2.0232	0.5839	12.0071	0.0005

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
PAIR 1 vs 56	1.000	0.012	84.111
PAIR 2 vs 56	2.750	0.040	187.621
.			
.			
PAIR 55 vs 56	1.000	0.012	84.111
SMOKE	7.562	2.408	23.750

Mcnemar's test in SAS–low birthweight data

-------Please fix this ---------------

The FREQ Procedure

Statistics for Table of LOW by SMOKE

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	2.8846	1.3193	6.3069
Cohort (Col1 Risk)	1.5385	1.1099	2.1324
Cohort (Col2 Risk)	0.5333	0.3298	0.8624

McNemar's Test

Statistic (S)	2.3810
DF	1
Pr > S	0.1228

Conditional logistic in SAS–the full model from variable selection

The LOGISTIC Procedure
Conditional Analysis
Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.6464	3	0.0008
Score	13.9668	3	0.0030
Wald	10.6506	3	0.0138

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
SMOKE	1	1.1867	0.4745	6.2533	0.0124	3.276
PTD	1	1.4183	0.6262	5.1300	0.0235	4.130
UI	1	1.0046	0.6417	2.4506	0.1175	2.731

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
SMOKE	3.276	1.293	8.304
PTD	4.130	1.210	14.093
UI	2.731	0.776	9.606

Data Example for 1:3 matched case-control data

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

Note: since data is individually matched, the correct analysis is using conditional logistic

Homework

1. estimate the association between smoking and birth weight in the 1:3 matched low birth weight data using fully stratified analysis
2. estimate the association between smoking and birth weight in the 1:3 matched low birth weight data using conditional logistic
3. check if there is interaction between smoking and UI in model 2
4. does UI confound the relationship between smoking and low birth weight?
5. predict low birth weight in the 1:3 matched low birth weight data using conditional logistic (find the best predictive model. Do not use mechanical variable selection techniques
6. write the interpretation of the regression coefficeints of the model in (2)
7. report the Wald test for the coefficient of smoking of the model in (2)
8. predict low birth weight in the 1:3 matched low birth weight data using conditional logistic (find the best predictive model) using stepwise
9. Assess the goodness of fit of the model in (8) using Hosmer-Lemeshow test, Deviance and Pearson Chisquare
10. Find any influential observations in the model in (8) using DFBETAS

Methods III: New course in Fall 2007

1. Risk evaluation: Measures of Disease Occurrence, Measures of Association, Attributable Risk, Asymptotic Theory (delta methods, Slutsky's theorem, Cramer Rao lower bound)
2. Logistic regression for different sampling models: Cross-sectional, Cohort, Case-control (matched and unmatched).
3. Multinomial logistic regression: Model specification, Estimation of Parameters, Interpretation of Parameter Estimates, Model Diagnostics
4. Ordinal logistic regression: Model specification, Estimation of Parameters, Interpretation of Parameter estimates, Model Diagnostics
5. Logistic regression for correlated data: Generalized Estimating Equations, Covariance Structure, Model Diagnostics
6. Exact Methods for Logistic Regression
7. Analysis of Count Data (Poisson and Log-Binomial regression): Model specification, Estimation of Parameters, Interpretation of Parameter estimates
8. Likelihood Techniques: Full Likelihood, Marginal Likelihood, Quasi Likelihood, Profile Likelihood
9. Missing Data Methods: Nature of Missing Data, Adhoc Missing data techniques, Multiple Imputation