
Lecture 13 – More on Matched Models

■ Outline

- Interactions with the matching variables
- Additive models
- Diagnostics
- Efficiency

Underlying model :

$$\text{logit}(P_{ji}) = \log(P_{ji} / (1 - P_{ji})) = \alpha_j + \beta_1 X_{ji1} + \beta_2 X_{ji2} + \dots + \beta_k X_{jik}$$

for the j th matched set, i th person in that set

Unconditional model : Need to estimate $\alpha_1, \alpha_2, \dots, \alpha_J$

Conditional model : “Remove” $\alpha_1, \alpha_2, \dots, \alpha_J$ from the model

As J gets large, the need to fit a conditional model becomes more acute

Results for β 's potentially biased

Leisure world data

- Neither ignoring the matching nor fitting individual α_j for each matched set is a good idea

Leisure world 1 to 1 data

	GALL $\hat{\psi}$	ESTROGEN $\hat{\psi}$
Conditional (matched)	2.00	9.11
Unconditional - breaking the matches	2.34	8.95
Unconditional - fitting estimates for the matched pairs	4.01	82.93

Leisure world 1 to 4 data

	GALL $\hat{\psi}$	ESTROGEN $\hat{\psi}$
Conditional (matched)	3.58	8.29
Unconditional - breaking the matches	3.18	7.55
Unconditional - fitting estimates for the matched pairs	6.03	14.85

Interactions with the matching variables

What do you lose in the conditional approach ?

- the ability to look at main effects of the variables used for matching

Can we still fit interaction terms involving the matching variables? Yes!

In this case we do not want to use “Strata”, but rather “Age” since it is one of the matching variables

stratum	outcome	age	gall	gall * age	est	est * age	Conjugated estrogen dose
1	1	74	0	0	1	74	4
1	0	75	0	0	0	0	0
1	0	74	0	0	0	0	0
1	0	74	0	0	0	0	0
1	0	75	0	0	1	75	1
2	1	67	0	0	1	67	6
2	0	67	0	0	1	67	6
2	0	67	0	0	0	0	0
2	0	67	0	0	1	67	3
2	0	68	0	0	1	68	3
3	1	76	0	0	1	76	1
3	0	76	0	0	1	76	2
3	0	76	0	0	1	76	0
3	0	76	0	0	1	76	3
3	0	77	0	0	1	77	0

Note that the first three sets have no variation in gall so would not contribute

to estimation of that odds ratio; third set has no variation in estrogen use either

Interactions with the matching variables

Significant interactions indicate effect modification by the matching variable:

For example, estrogen may have effects only on one part of the age spectrum

Could create age group variable, but will only try age as a continuous variable here

Est * Age and Gall * Age

```
. clogit case gall estrogen , group(set)
Conditional (fixed-effects) logistic regression   Number of obs   =       315
                                                LR chi2(2)      =       45.05
                                                Prob > chi2     =       0.0000
Log likelihood = -78.871308                    Pseudo R2       =       0.2221
```

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gall	1.274654	.4108678	3.10	0.002	.4693683 2.079941
estrogen	2.114785	.4397942	4.81	0.000	1.252804 2.976766

```
. est store A
```

Now add the interaction terms to the model

```
. gen estage= estrogen*age
. clogit case gall estrogen estage , group(set)
Conditional (fixed-effects) logistic regression   Number of obs   =       315
                                                LR chi2(3)      =       45.06
                                                Prob > chi2     =       0.0000
Log likelihood = -78.864691                    Pseudo R2       =       0.2222
```

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gall	1.276811	.4111677	3.11	0.002	.470937 2.082685
estrogen	1.594645	4.535164	0.35	0.725	-7.294114 10.4834
estage	.007207	.0626158	0.12	0.908	-.1155177 .1299318

Interactions with the matching variables

```
. lrtest . A
Likelihood-ratio test                    LR chi2(1) =      0.01
(Assumption: A nested in .)             Prob > chi2 =    0.9084

. gen gallage=gall*age

. clogit case gall estrogen gallage , group(set)
Conditional (fixed-effects) logistic regression   Number of obs   =      315
                                                    LR chi2(3)      =      45.25
                                                    Prob > chi2     =      0.0000
                                                    Pseudo R2      =      0.2232
Log likelihood = -78.768358
```

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gall	-.8909208	4.825236	-0.18	0.854	-10.34821 8.566369
estrogen	2.130138	.4447269	4.79	0.000	1.258489 3.001787
gallage	.0305221	.0678171	0.45	0.653	-.1023969 .1634412

```
. lrtest . A
Likelihood-ratio test                    LR chi2(1) =      0.21
(Assumption: A nested in .)             Prob > chi2 =    0.6500
```

No evidence of effect modification by age

Underlying model for the interaction with age:

$$\text{logit}(P_{ji}) = \alpha_j + \beta_1 \text{estrogen}_{ji} + \beta_2 \text{gall}_{ji} + \beta_3 \text{age}_{ji} * \text{estrogen}_{ji}$$

for the j th matched set, i th person in that set

Condition out the $\alpha_1, \alpha_2, \dots, \alpha_J$ from the model

Matching variable age appears in the interaction term, but the main effect is removed by matching

Additive models

Additive models

Can also consider a linear relative risk model

General conditional likelihood for matched set j

$$\frac{r_{j0}}{r_{j0} + \sum_{\text{controls } i=1 \text{ to } m} r_{ji}}$$

where r is the risk function

Multiplicative (exponential risk) $r = e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$

Additive (linear risk) $r = 1 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$

Have to impose some constraints so that the risk is positive

Can be more biologically plausible than a standard multiplicative model

Compare additive and multiplicative models for leisure 1 to 4 data

This is what we found before with no interaction in the multiplicative model

Additive models

Estimated OR	No estrogen	Estrogen
No gall bladder	1.0	8.29
Gall bladder	3.58	29.65

And with the interaction term

Estimated OR	No estrogen	Estrogen
No gall bladder	1.0	14.88
Gall bladder	18.07	34.52

- As is often the case, the interaction term allows $OR(A=1 \& B=1) < OR(A=1) \times OR(B=1)$
- Suggests an additive model without interaction might be as good
- Stata cannot fit the additive model without writing a special routine; some specialized packages (Egret; Epicure) can do these model fits

Additive models

Fitting the linear (additive) model in Egret gives a two parameter model whose fit is almost as good as the three parameter multiplicative model

Estimated OR	No estrogen	Estrogen
No gall bladder	1.0	14.95
Gall bladder	19.23	$1.0+13.95+18.23 = 33.18$

```
-----RESULTS-----[CLR/a]--
TERM                ODDS RATIO      95% CONFIDENCE BOUNDS
gallbladdr          19.23           -9.103           47.57
estuse              14.95           -2.949           32.84
```

Note the Wald confidence bounds are nonsensical

The two parameter additive model fits the data most simply, however, despite the poor statistical properties associated with additive models

- Additive models can be attractive scientifically, but in practice are difficult to fit and have poor statistical properties

Diagnostics

Diagnostics

Based on the “score” contribution to the conditional log likelihood

Cain and Lange (1982) developed for Cox regression

Adapted by Barlow and Prentice (1988) to conditional logistic regression
and Cox regression with time-dependent covariates

One covariate (\mathbf{X}) $\Delta\hat{\beta} = \text{Var}(\hat{\beta}) * (Y_i - p_i) * (X_i - \bar{X}_{\text{weighted}}) = \text{Var}(\hat{\beta}) * \text{score}_{ji}$

Actual outcome Y_i

Conditional expected probability $p_{ji} = \frac{e^{x_{ji}\hat{\beta}_1}}{\sum_k e^{x_{jk}\hat{\beta}_1}}$

Observed covariate X_i

Expected covariate $\bar{X}_{\text{weighted}} = \sum_k p_{jk} x_{jk}$

Multivariate version slightly more complicated

Diagnostics

Robust (empirical) Variance

Based on the delta beta diagnostic from the multivariate model

$$\text{Robust variance} = \Delta \hat{\beta}^T \Delta \hat{\beta}$$

Stata modifies this slightly since it adjusts for the df (minor change)

Stata does not give the delta beta's but it does give the robust variance estimates

```
. clogit case gall estrogen , group(set) robust
Conditional (fixed-effects) logistic regression   Number of obs   =       315
                                                    Wald chi2(2)    =       36.09
                                                    Prob > chi2     =       0.0000
Log pseudolikelihood = -78.871308                Pseudo R2       =       0.2221
```

(Std. Err. adjusted for clustering on set)

case	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
gall	1.274654	.4465026	2.85	0.004	.3995255	2.149783
estrogen	2.114785	.397369	5.32	0.000	1.335956	2.893614

```
. clogit case gall estrogen , group(set)
Conditional (fixed-effects) logistic regression   Number of obs   =       315
                                                    LR chi2(2)     =       45.05
                                                    Prob > chi2     =       0.0000
Log likelihood = -78.871308                Pseudo R2       =       0.2221
```

case	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
gall	1.274654	.4108678	3.10	0.002	.4693683	2.079941
estrogen	2.114785	.4397942	4.81	0.000	1.252804	2.976766

Efficiency

What if you match when the matching variable is unimportant ?

**⇒ Loss of efficiency compared to unconditional logistic regression
except when the true odds ratio is one**

**Degree of loss depends on the true odds ratio and how common
the binary exposure is**

Figure 7.1 in Breslow & Day Volume 1 shows the expected efficiency

For $\beta = 1$, the loss can be as great as 10%

**For $\beta = 2$, the loss can be as much as 40% (when the prevalence of
exposure in controls is 30%)**

- **Matching on known risk factors is important**
- **Matching on well measured characteristics is also important**