BST 210
Applied Regression Analysis

Lecture 19
Plan for Today

- Logistic regression for

  - Retrospective designs

  - Matched designs
    (Conditional Logistic Regression)

Case-Control versus Cohort Studies

Prospective Study:

- watches for outcomes, such as the development of a disease, during the study period and relates this to other factors such as suspected risk or protection factor(s)
- usually involves taking a cohort of subjects and watching them over a long period
- outcome of interest should be common; otherwise, the number of outcomes observed will be too small to be statistically meaningful (indistinguishable from those that may have arisen by chance)
- all efforts should be made to avoid sources of bias such as the loss of individuals to follow up during the study
- Prospective studies usually have fewer potential sources of bias and confounding than retrospective studies

Case-Control versus Cohort Studies

Restrospective Study:

- looks backwards and examines exposures to suspected risk or protection factors in relation to an outcome that is established at the start of the study

- Many valuable case-control studies, such as Lane and Claypon's 1926 investigation of risk factors for breast cancer, were retrospective investigations

- Confounding and bias are more common in retrospective studies than in prospective studies; special care must be taken to try to avoid (retrospective studies are thus sometimes criticized)

- In retrospective studies the odds ratio provides an estimate of relative risk (sampling fraction issue).

Case-Control versus Cohort Studies

- Cohort studies are usually but not exclusively, prospective

- Case-Control studies are usually but not exclusively, retrospective

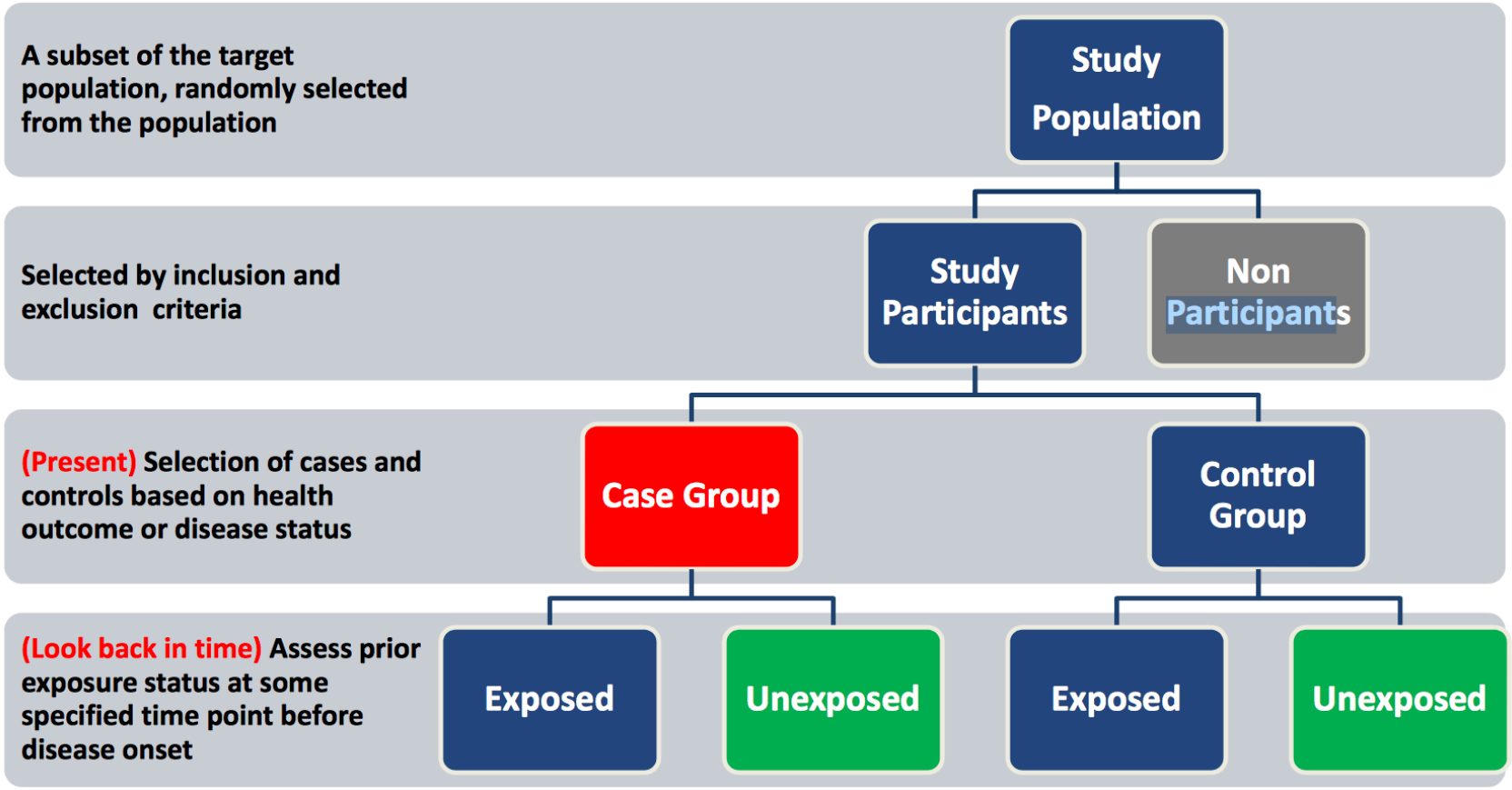
Cohort Studies

- **outcome is measured after exposure (we are deciding on exposure, then waiting for outcomes)**
- yields true incidence rates, relative risks, odds ratios (no *sampling fraction* to worry about)
- may uncover unanticipated associations with outcome
- best for common outcomes
- expensive
- requires large numbers
- takes a long time to complete
- prone to attrition bias or bias of change in methods over time

Case-Control Studies

- **outcome is measured before exposure (we are deciding on outcome, then looking back for exposures– 2x2 table)**
- controls are selected on the basis of not having the outcome
- often conducted before a cohort or an experimental study to identify the possible etiology of the disease.
- must incorporate *sampling fraction* in calculation of incidence rates, relative risks -- not possible to estimate incidence of disease unless study is population based and all cases in a defined population are obtained (odds ratios are always valid in cohort or case-control)
- good for rare outcomes
- relatively inexpensive
- smaller numbers required
- quicker to complete
- prone to selection bias and recall/retrospective bias

Case-Control Study Design



Recall: Multiple Logistic Regression Model

- We know that in logistic regression, to predict a binary outcome Y with covariates $X_1, \ldots, X_p$, we use the model:

$$\text{logit}(p) = \log[p / (1 - p)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Here we assume that the relationship between logit(p) and the covariates $x_1, \ldots, x_p$ is linear

- Usually we are thinking that the probability p is the probability of the event occurring *prospectively*

- What if the events of interest actually happened in the *past*? How does this change our statistical approach?

Case-Control Studies

- We then work under the assumptions of a case-control design

- Suppose we have risk factors x_1, …, x_K in a case-control study with the underlying logistic regression model given by:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki}.$$

- The complicating issue in case-control studies is that the *sampling fraction of cases* might be different (probably higher) than the *sampling fraction of controls*

Case-Control Studies

How do we address this?

- Let τ_1 = proportion of cases sampled in the case-control study, and τ_0 = proportion of controls sampled (different from τ_1, probably lower)

- Assume that the sampling fractions of cases and controls are independent of other risk factors

- We run a logistic regression model for this data with sampling fractions τ_1 and τ_0

Case-Control Studies

- The true logistic regression model can then be shown to be:

$$\ln[p_i / (1 - p_i)] = \alpha + \ln(\tau_1 / \tau_0) + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki}$$

- We can let $\boldsymbol{\alpha}^* = \alpha + \ln(\frac{\tau_1}{\tau_0})$, and proceed as usual

- We can validly estimate the odds ratio for the k^{th} risk factor by $\exp(\beta_k)$, but the meaning of α changes due to our 'artificially constructed' data in a case-control study

- However, we cannot estimate absolute probabilities of disease (the p_i), since the sampling fractions τ_1 and τ_0 are usually not known in a case-control study

Example: Passive Smoking

- A case-control study examining the association between passive smoking and risk of (any) cancer
- 509 cancer cases and 489 controls with a similar distribution of age and gender were enrolled
- Passive smoking was defined as cigarette smoking by a spouse of ≥ 1 cigarette per day for ≥ 6 months
- One possible confounding variable is active smoking by the subjects themselves

Example: Passive Smoking

Non-Smokers		**Passive Smoker**		
		Yes	No	Total
Case Control Status	Case	120	111	231
	Control	80	155	235
	Total	200	266	466

Example: Passive Smoking

Smokers		**Passive Smoker**		
		Yes	No	Total
Case Control Status	Case	161	117	278
	Control	130	124	254
	Total	291	241	532

Example: Passive Smoking

$$\ln[p_i / (1 - p_i)] = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i},$$

where

$x_{1i} = 1$ if subject i is a passive smoker,

$= 0$ otherwise

$x_{2i} = 1$ if subject i is an active smoker,

$= 0$ otherwise.

The results are shown in the next slide.

Logistic Regression Results

Variable	β	s.e.(β)	*OR* = exp(β)	P-value
Intercept	–0.226	0.198	---	0.037
Passive smoking	0.487	0.128	1.628	< 0.001
Active smoking	0.051	0.129	1.052	0.692

Mantel-Haenszel Method

- The M-H test compares the ORs from several 2x2 tables

```
. cc cancer passive_smk [fweight = freq], by
  (active_smk)
```

```
      active_smk |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   2.094595        1.41754   3.097166     19.05579 (exact)
               1 |   1.312558       .9184614   1.875813     28.59023 (exact)
-----------------+-------------------------------------------------
           Crude |   1.637406       1.265013   2.119599              (exact)
    M-H combined |   1.625329       1.263955   2.090024
-------------------------------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     3.27  Pr>chi2 = 0.0706

                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     14.42
                                                 Pr>chi2 =    0.0001
```

Logistic Regression and Mantel-Haenszel

- The M-H method first provides statistical tests of whether the ORs are equal (homogeneous) or unequal (heterogeneous) across strata (active smoker). Second, it provides an estimate of the OR of the exposure variable (passive smoking), adjusted for the strata variable (active smoker).

- The adjusted odds ratio for passive smoking obtained from the logistic regression model (1.628) is very similar to the estimate of the adjusted odds ratio from the Mantel-Haenszel method (1.625, the M-H combined estimate)

- Both adjust for active smoking status

Logistic Regression and Mantel-Haenszel

- Surprisingly, the odds ratio for active smoking is only 1.05, 95% confidence interval (0.82, 1.35) and $p = 0.69$, after controlling for passive smoking

- How could this be?

- This could be due to the large number of cancers that are not smoking-related

- We also have marginal evidence (via the M-H test of homogeneity of ORs) of effect modification ($p = 0.07$)

Prediction Probabilities in Case-Control Setting

- For a case-control study, we *cannot* use the logistic regression model for prediction of probabilities of the outcome because of the differential case-control sampling

- We usually cannot estimate the term $\log(\tau_1/\tau_0)$ in the true model because we don't know the true sampling fractions

- Let's try anyway, in order to demonstrate this point →

Prediction Probabilities in Case-Control Setting

- Take the subjects exposed to neither active nor passive smoking, and consider the predicted (from the model) probabilities of (any) cancer:

- $$\hat{p}_i = \frac{\exp(\alpha)}{1+\exp(\alpha)}$$

$$= \frac{\exp(-0.226)}{1+\exp(-0.226)} = 0.444,$$

which is too high due to the case - control sampling.

Example: Nurses' Health Study (NHS)

- A subgroup of women participating in the Nurses' Health Study (NHS) provided a blood sample in 1989-1990
- The blood was stored in freezers and used in nested case-control studies
- One such case-control study looked at serum estradiol (a hormone) and breast cancer risk

Example: NHS Sub-study

- 164 breast cancer cases (diagnosed between the time of the blood draw and the year 2000)
- 346 controls (1 or 2 controls for each case, with no breast cancer at the time of diagnosis of the case) were selected
- In many case-control studies, you might match the cases and controls on the basis of likely confounding variables that you know you want to control for
- Then need to account for this matching in your analysis (e.g., match on age within 2 years, medical history variables)

Example: NHS Sub-study

- Here, the blood samples from matched cases and controls were analyzed in the same batch
- If there could be substantial batch-to-batch assay variability, it is especially important that the matching be taken into account in the analysis – batch is another (categorical) matching factor

Logistic Regression for Matched Sets

- Suppose then that there are m matched sets (and in this example we assume matching on at least batch)
- Suppose also that we have n_{1i} cases and n_{2i} controls in the i^{th} matched set
- Let p_{ij} be the probability of disease for the j^{th} subject in the i^{th} matched set
- Let's write this out for, say, 5 matched sets…

Logistic Regression for Matched Sets

- To control for the *m* matched sets, one might choose a "baseline" matched set and then include an overall intercept and *m* – 1 indicator variables (for the non-baseline matched sets)

- An alternate, equivalent approach would be to include no intercept and *m* indicator variables, one for each matched set (let's work with this approach)

Logistic Regression for Matched Sets

- The model is then:

$$\text{logit}(p_{ij}) = \alpha_i + \sum_{k=1}^{K} \beta_k x_{ijk},$$

where

x_{ijk} = value of the k^{th} covariate measured on the j^{th} subject in the i^{th} matched set.

- Let's write this out…

Logistic Regression for Matched Sets

- Since the matched sets are usually small, it is virtually impossible to directly estimate the α_i, since there are so many α_i values
- Therefore, the absolute probability of disease cannot be estimated—we must do something else, clever
- Instead, we use a conditional approach to estimate the other regression parameters (i.e., β_k, $k = 1, \ldots, K$), "conditioning out" the effect of the α_i parameters
- Before demonstrating with an example, let's denote the *likelihood function* for such '*conditional*' models (which we later need for estimation of our model parameters) →

Conditional Logistic Regression

- The overall *conditional likelihood* for logistic regression models that employ 'conditioning' on some aspect, such as matching criteria, is given by:

$$L_C(\beta) = L_1(\beta) \times L_2(\beta) \times \ldots \times L_m(\beta)$$

- We find the conditional maximum likelihood estimates (CMLEs) that maximize this conditional likelihood $L_C(\beta)$.

- Standard methods (based on log likelihoods, p-values, model building, etc.) can be used. *Great!*

Conditional Logistic Regression

- $L_i(\beta)$ is the contribution to this *conditional likelihood* for the i^{th} matched pair

- Note (upcoming) that the term identifying the matched set (the intercept α_i) does not appear in $L_i(\beta)$; it is conditioned out

Conditional Logistic Regression

To demonstrate this 'conditioning-out' of the α intercepts:

- Suppose (for the sake of simplicity) that there is exactly 1 case and 1 control for each matched set,
- And that within each set (pair), subject 1 is the case and subject 2 is the control
- Let $Y_{ij} = 1$ if the j^{th} woman in the i^{th} matched pair is a case, and 0 if she is a control

Conditional Logistic Regression

- Using the multiplication rule of probability (with two independent responses), P(A and B) = P(A)*P(B), and recalling that the PMF for the Binomial distribution is p*(1-p), we have:

- $$\Pr(Y_{i1} = 1 \text{ and } Y_{i2} = 0) = \frac{\exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i1k})}{[1 + \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i1k})]} \times \frac{1}{[1 + \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i2k})]}$$

- Let's write this out for our example…

Conditional Logistic Regression

- Now consider the conditional probability that subject 1 is a case given that exactly 1 out of the 2 subjects in the matched set (pair) is a case, that is

$L_i(\beta_1)$ = P(Y_{i1} = 1 ∩ Y_{i2} = 0 | exactly 1 case (ie Y_{i1} = 1) in matched set)

- Using the formula $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$ for a conditional probability we get:

$$L_i(\beta_1) = \frac{P(Y_{i1} = 1 \cap Y_{i2} = 0)}{P(Y_{i1} = 1 \cap Y_{i2} = 0) + P(Y_{i1} = 0 \cap Y_{i2} = 1)}$$

Conditional Logistic Regression

- Then

$$\Pr(Y_{i1}=1 \text{ and } Y_{i2}=0 \text{ given 1 case in a matched set})$$

$$\equiv \mathrm{L_i} = \frac{\Pr(Y_{i1}=1 \text{ and } Y_{i2}=0)}{\Pr(Y_{i1}=1 \text{ and } Y_{i2}=0)+\Pr(Y_{i1}=0 \text{ and } Y_{i2}=1)}$$

$$= \frac{\exp(\sum_{k=1}^{K}\beta_k x_{i1k})}{\exp(\sum_{k=1}^{K}\beta_k x_{i1k})+\exp(\sum_{k=1}^{K}\beta_k x_{i2k})}$$

- Let's write this out…

Conditional Logistic Regression

- This approach can be generalized to allow for any number of cases (n_{1i}) and any number of controls (n_{2i}) in a matched set
- Also, different matched sets do not have to have the same number of cases and controls, or even the same total number of subjects
- This model is referred to as the *conditional logistic regression model*

Conditional Logistic Regression

- If there is one case ($Y_{i1} = 1$) and two controls ($Y_{i2} = Y_{i3} = 0$) in a matched set, what would be the conditional likelihood contribution for this matched set?

  (challenge yourself!—but you will not be responsible for this.)

Conditional Logistic Regression

- If there is two cases ($Y_{i1} = Y_{i2} = 1$) and two controls ($Y_{i3} = Y_{i4} = 0$) in a matched set, what would be the conditional likelihood contribution for this matched set?

  (challenge yourself!—but you will not be responsible for this.)

Interpretation of Parameters

- Suppose we want to estimate the odds ratio for the effect of a 1 unit increase in the first covariate , holding all other covariates constant, for two subjects in a matched set…

Interpretation of Parameters

- $$\text{logit}(p_{i1}) = \alpha_i + \beta_1(x_1 + 1) + \sum_{k=2}^{K} \beta_k x_{i1k},$$

  $$\text{logit}(p_{i2}) = \alpha_i + \beta_1(x_1) + \sum_{k=2}^{K} \beta_k x_{i1k}.$$

  Thus,

  $$\text{logit}(p_{i1}) - \text{logit}(p_{i2}) = \beta_1,$$

  or

  $$OR_{1 \text{ vs. } 2} = \exp(\beta_1).$$

Interpretation of Parameters

- Thus, our β estimates can be interpreted as log odds ratios for the effect of a one unit increase in a covariate, holding all other covariates to be the same, basically the same as for (ordinary, unconditional) logistic regression.

- Here when we say "holding all other covariates to be the same" we *also* mean being *in the same matched set*, so that the α_i terms cancel out.

Example: NHS Sub-study

```
. summarize case currentpmh ageblood estradiol
```

Variable	Obs	Mean	Std. Dev.	Min	Max
case	510	.3215686	.467537	0	1
currentpmh	510	.1509804	.3583813	0	1
ageblood	510	60.96863	4.989478	45	69
estradiol	510	8.907843	7.741739	2	85

```
. table case
```

case	Freq.
0	346
1	164

Example: NHS Substudy

- There were a total of 510 women in the study, of whom 164 were cases and 346 were controls
- Sample observations are given on the next slide

Example: NHS Substudy

Among 510 women in the study, 164 were cases

	id	matchid	case	curpmh	age	estradiol
19.	100241	107261	0	0	65	11
20.	212974	107261	0	0	64	8
21.	108215	108215	1	0	58	8
22.	106487	108946	0	0	62	6
23.	108946	108946	1	0	61	4
24.	116697	108946	0	0	58	9
25.	102266	109861	0	1	64	6
26.	103214	109861	0	0	65	5
27.	109861	109861	1	1	66	5
28.	100696	110294	0	1	66	3
29.	127187	110294	0	0	68	6

Example: NHS Substudy

- Each subject has an individual identification number (id), and also a matchid which identifies the matched set to which the subject belongs
- A matched set is only useful for the analysis if there is at least one case and at least one control
- Matchid 108215 is not useful because there were no controls matched to this case
- However, matchid 108946 is useful because it has 1 case and 2 controls

Example: NHS Substudy

- The mean estradiol is 8.9 with SD 7.7 and a maximum of 85, suggesting that the distribution of estradiol is quite skewed, so the investigators created a new variable ln_estradiol for the analysis

- Note: Logistic regression does not require that a continuous covariate be normally distributed; here one could try both estradiol and ln_estradiol as covariates and then pick between them (or use splines, quadratic terms, etc.)

Example: NHS Substudy

- $$\text{logit}(p_{ij}) = \alpha_i + \beta_1 \text{ageblood}_{ij} + \beta_2 \text{currentpmh}_{ij} + \beta_3 \ln(\text{estradiol}_{ij}),$$

  where

  $p_{ij} = \Pr(j^{\text{th}}$ subject in the i^{th} matched pair is a case),

  ageblood_{ij} = age at the blood draw for the j^{th} subject in the i^{th} matched pair

  $\text{currentpmh}_{ij} = 1$ if the j^{th} subject in the i^{th} matched pair used postmenopausal hormones at the time of the blood draw, $= 0$ otherwise

  $\text{ln\_estradiol}_{ij} = \log(\text{estradiol for the } j^{\text{th}} \text{ subject in the } i^{\text{th}} \text{ matched pair}).$

Using Stata for Conditional LR

```
. clogit case currentpmh ageblood ln_estradiol,
  group(matchid)

note: multiple positive outcomes within groups encountered.
note: 82 groups (126 obs) dropped due to all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -132.67886
Iteration 1:   log likelihood = -132.50721
Iteration 2:   log likelihood = -132.50714
Iteration 3:   log likelihood = -132.50714

Conditional (fixed-effects) logistic regression   Number of obs    =        384
                                                  LR chi2(3)       =      12.57
                                                  Prob > chi2      =     0.0057
Log likelihood = -132.50714                       Pseudo R2        =     0.0453

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  currentpmh |    .237519   .3007713     0.79   0.430     -.351982      .82702
    ageblood |  -.0248404   .1526609    -0.16   0.871    -.3240501     .2743694
ln_estradiol |   .6826214   .2054062     3.32   0.001     .2800327     1.08521
------------------------------------------------------------------------------
```

Example: NHS Substudy

- We see that 82 “matched sets” (126 women) were not used in the analysis because there were either 0 cases or 0 controls in the matched set
- We also see that ln_estradiol is significant with p = 0.001, while the other two variables in the model are not significant

Using Stata for Conditional LR

To get odds ratios:

```
. clogit, or

Conditional (fixed-effects) logistic regression   Number of obs    =        384
                                                  LR chi2(3)       =      12.57
                                                  Prob > chi2      =     0.0057
Log likelihood = -132.50714                       Pseudo R2        =     0.0453

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  currentpmh |   1.268099   .3814079     0.79   0.430      .7032928    2.286495
    ageblood |   .9754656   .1489154    -0.16   0.871       .723214    1.315701
ln_estradiol |   1.979059   .4065109     3.32   0.001      1.323173    2.960062
------------------------------------------------------------------------------
```

Example: NHS Sub-study

- For each one unit increase in log(estradiol), the estimated odds ratio is 1.98 with 95% confidence interval (1.32, 2.96)
- Suppose we have 2 women in the same matched set, one of whom has log(estradiol) 1 unit higher than the other (e.g., approximately $e^1 = 2.7$ times as high)
- The woman with higher estradiol has a 2-fold higher odds of having breast cancer than the woman with lower estradiol, holding the other variables (age, current post-menopausal hormone use, and the matching factor batch) constant

Using Stata for Conditional LR (here using non-logged estradiol)

```
. clogit case currentpmh ageblood estradiol,
  group(matchid)

note: multiple positive outcomes within groups encountered.
note: 82 groups (126 obs) dropped due to all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -132.84774
Iteration 1:   log likelihood = -132.26253
Iteration 2:   log likelihood = -132.26197
Iteration 3:   log likelihood = -132.26197

Conditional (fixed-effects) logistic regression   Number of obs    =        384
                                                  LR chi2(3)       =      13.06
                                                  Prob > chi2      =     0.0045
Log likelihood = -132.26197                       Pseudo R2        =     0.0470

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  currentpmh |   .2279739   .3019763     0.75   0.450    -.3638888    .8198366
    ageblood |  -.0236678   .1536806    -0.15   0.878    -.3248763    .2775406
   estradiol |    .056603   .0182511     3.10   0.002     .0208316    .0923745
------------------------------------------------------------------------------
```

Using Stata for (Unconditional) Logistic Regression

```
. logistic case currentpmh ageblood ln_estradiol

Logistic regression                               Number of obs   =        510
                                                  LR chi2(3)      =      14.14
                                                  Prob > chi2     =     0.0027
Log likelihood = -313.23517                       Pseudo R2       =     0.0221

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  currentpmh |   1.405057   .3758035     1.27   0.204      .8318163    2.373343
    ageblood |   1.008835   .0199317     0.45   0.656      .9705161    1.048666
ln_estradiol |   1.825742   .3105155     3.54   0.000      1.308188    2.548054
------------------------------------------------------------------------------
```

Example: NHS Substudy

- The standard errors of the β's are smaller for the unconditional logistic regression compared with the conditional regression, especially for the age at blood draw variable

- This is due to the correlation resulting from multiple subjects in the same matched set

- However, the only *appropriate* analysis of a matched dataset is one that takes into account the matching, namely CMLE

Conditional Logistic Regression

Some final words:

- If you match on a factor, you cannot estimate a β for this factor – you can only say that you are controlling for that factor with CMLE
- However, you can interact this variable with another (unmatched) variable
- CMLE arises with matching in case-control studies, but also pairs of eyes or knees, family members, etc.

Coming Up

- Generalized linear models, and seeing how linear and logistic regression fall within this framework
- Overdispersion and other variance estimates
- Poisson regression and more