Lecture 8: Two Related Topics

I. Overdispersion

II.Study Designs

Lin Chen

Department of Public Health Sciences
The University of Chicago

Part I: Overdispersion

Example 1: *Grain beetles* The data is from an experiment studying the effect of ethylene oxide as a pesticide against grain beetles. Ten batches of about thirty insects were exposed to various concentrations (measured in mg/l) of the pesticide. The numbers of insects affected by the pesticide are recorded.

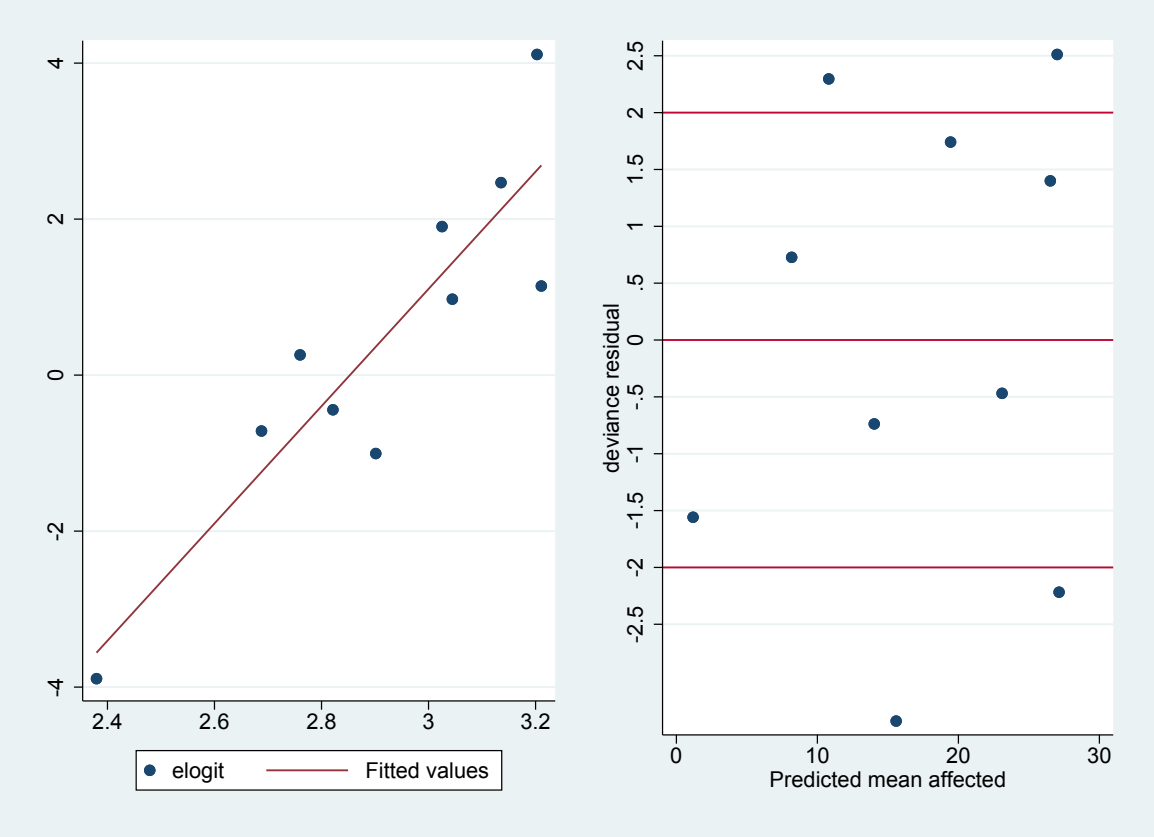
	concen~n	affected	exposed
1.	24.8	23	30
2.	24.6	30	30
3.	23	29	31
4.	21	22	30
5.	20.6	23	26
6.	18.2	7	27
7.	16.8	12	31
8.	15.8	17	30
9.	14.7	10	31
10.	10.8	0	24

Example: *Grain beetles*

```
. generate logconc = log(concentration)

. generate elogit = log((affected+0.5)/(exposed-affected+0.5))

. glm affected logconc, f(bin exposed) nolog

Generalized linear models                         No. of obs         =         10
Optimization     : ML                             Residual df        =          8
                                                  Scale parameter    =          1
Deviance         =  36.44363626                   (1/df) Deviance    =   4.555455
Pearson          =  33.43402849                   (1/df) Pearson     =   4.179254

Variance function: V(u) = u*(1-u/exposed)         [Binomial]
Link function    : g(u) = ln(u/(exposed-u))       [Logit]

                                                  AIC                =   6.801695
Log likelihood   = -32.00847631                   BIC                =   18.02296

------------------------------------------------------------------------------
             |                 OIM
    affected |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     logconc |    6.265436    .778126     8.05   0.000     4.740338    7.790535
       _cons |   -17.86704   2.269017    -7.87   0.000    -22.31423   -13.41985
------------------------------------------------------------------------------

. display chi2tail(8,36.44363626)
.00001456
```

Example: *Grain beetles:* logit model linear in log concentration

```
. twoway (scatter elogit logcon) (lfit elogit logconc), name(A,replace)
. predict yhatl
. predict dresl, d

. twoway (scatter dresl yhatl), ylabel(-2.5(.5)2.5) yline(0 2 -2) name(B, replace)
. graph combine A B
```



Overdispersion in *Grain beetles* example

- The empirical logit plot reveals no clear departures from linearity and no obvious outliers. The deviance (36.4 on 8 d.f.) indicates significant lack of fit. When plotting residuals, we see that 4 out of 10 groups/batches have large residuals.
- The poor fit is not because the models is getting the *means* (logit of prob or log odds) wrong, but that it is getting the *variability* wrong.
- **Overdispersion** occurs when the observed variance is higher than the variance of a theoretical model.
- One possibility is that there were differences in the way that the ten batches of insects were managed during the course of the experiment.
- What may cause overdispersion?
  - Heterogeneity between the response probabilities. Different batches might be treated slightly differently.
  - Positive correlation between responses (some are more similar)
  - Omitted predictors

Overdispersion in *Grain beetles* example

- The empirical logit plot reveals no clear departures from linearity and no obvious outliers. The deviance (36.4 on 8 d.f.) indicates significant lack of fit. When plotting residuals, we see that 4 out of 10 groups/batches have large residuals.
- The poor fit is not because the models is getting the *means* (logit of prob or log odds) wrong, but that it is getting the *variability* wrong.
- **Overdispersion** occurs when the observed variance is higher than the variance of a theoretical model.
- One possibility is that there were differences in the way that the ten batches of insects were managed during the course of the experiment.
- What may cause overdispersion?
  - Heterogeneity between the response probabilities. Different batches might be treated slightly differently.
  - Positive correlation between responses (some are more similar)
  - Omitted predictors

Overdispersion in *Grain beetles* example

- The empirical logit plot reveals no clear departures from linearity and no obvious outliers. The deviance (36.4 on 8 d.f.) indicates significant lack of fit. When plotting residuals, we see that 4 out of 10 groups/batches have large residuals.
- The poor fit is not because the models is getting the *means* (logit of prob or log odds) wrong, but that it is getting the *variability* wrong.
- **Overdispersion** occurs when the observed variance is higher than the variance of a theoretical model.
- One possibility is that there were differences in the way that the ten batches of insects were managed during the course of the experiment.
- What may cause overdispersion?
  - Heterogeneity between the response probabilities. Different batches might be treated slightly differently.
  - Positive correlation between responses (some are more similar)
  - Omitted predictors

Overdispersion: an alternative model

- A simple alternative model for overdispersion is:

$$\mathrm{E}(Y_i) = n_i p_i \tag{1}$$

and

$$\mathrm{Var}(Y_i) = \phi n_i p_i (1 - p_i) \tag{2}$$

where $\phi > 1$ is the scaling factor.

- If this model is correct, then the Pearson's χ^2 statistic (for goodness of fit) divided by its degree of freedom is an estimate for ϕ (McCullagh and Nelder, 1989).

$$\mathrm{E}(X^2) \approx (n - p)\phi$$

where X^2 is the Pearson's χ^2 statistic.

- Stata can fit this model automatically using the option "`scale(x2)`".

Overdisperson model estimates for *Grain beetles* example

```
. glm affected logconc, f(bin exposed) scale(x2) nolog

Generalized linear models                         Number of obs   =         10
Optimization     : ML                             Residual df     =          8
                                                  Scale parameter =          1
Deviance         =  36.44363626                   (1/df) Deviance =   4.555455
Pearson          =  33.43402849                   (1/df) Pearson  =   4.179254

Variance function: V(u) = u*(1-u/exposed)         [Binomial]
Link function    : g(u) = ln(u/(exposed-u))       [Logit]

                                                  AIC             =   6.801695
Log likelihood   = -32.00847631                   BIC             =   18.02296

------------------------------------------------------------------------------
             |                 OIM
    affected |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     logconc |    6.265436    1.59074     3.94   0.000     3.147643     9.38323
       _cons |   -17.86704    4.638601   -3.85   0.000    -26.95853    -8.775553
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

Comparing with the results on page 3, the estimated coefficients are the same, the CIs are much wider now.

Summary of overdispersion

There may exist apparent overdispersion when there are:

1. Omitted important predictors
2. Mis-specification of predictor form (quadratic, polynomial, interaction terms are missing)
3. Outliers
4. Mis-specification of link function (more later)
5. Response probability being too small or too large (more later)

Overdispersion do not affect prediction ($\hat{\beta}$'s) but ignoring it may underestimate variation and lead to inflated type I error rates.

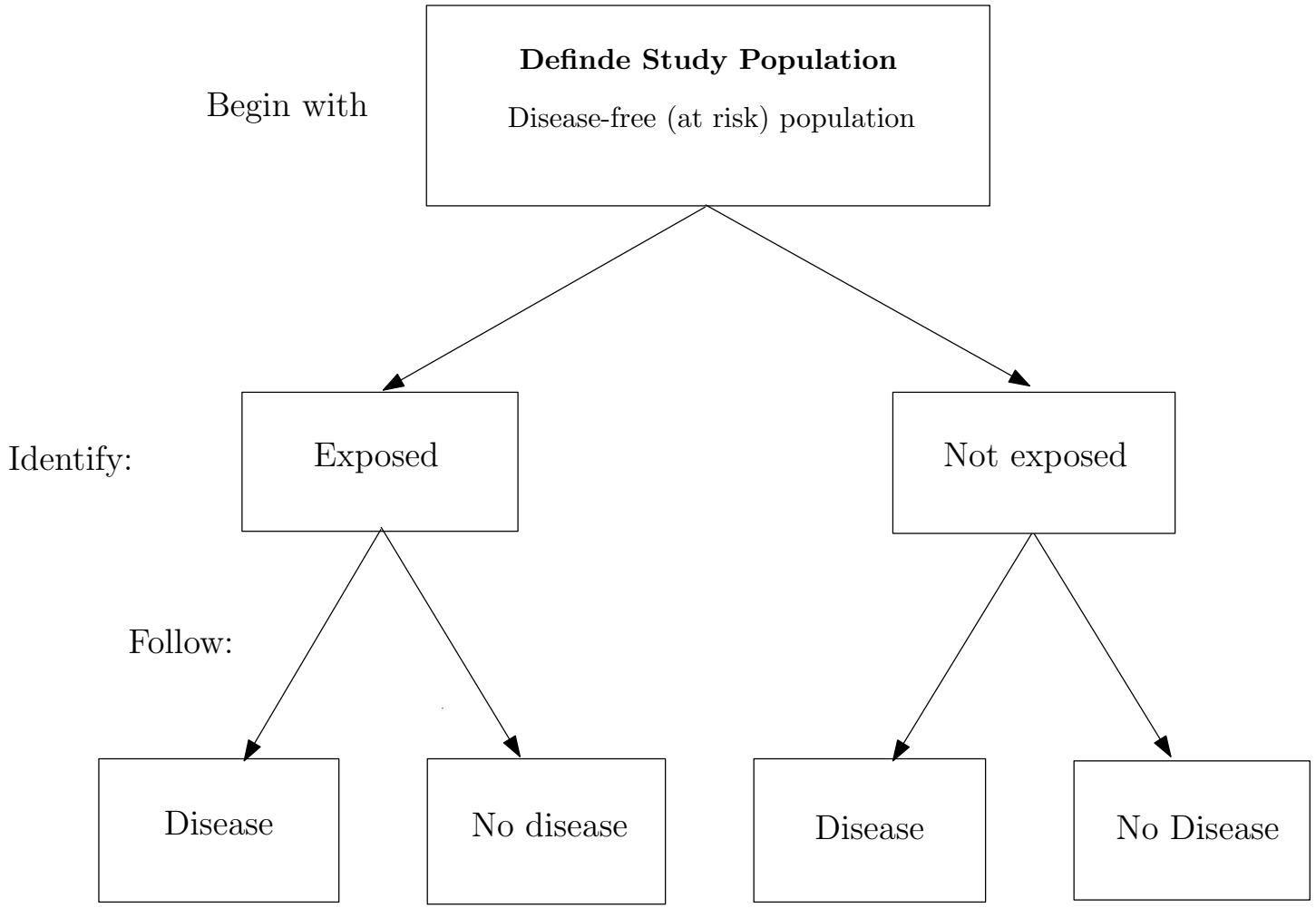
Correcting for overdispersion should be done only after convincing yourself that the model for the mean is correct– curvature on the logit scale is appropriately taken into account, there are no plausible terms (including interactions) missing from the model, and outliers have already been accounted for. Possibility 1-3 should be checked before correcting/adjusting overdispersion.

Part II. Study designs in Epidemiological studies

- Epidemiology is the study of patterns of disease occurrence, the distributions of health related events in defined populations and the factors that influence such patterns/distributions.
- Two commonly used designs in epidemiological studies are the *cohort* study and the *case-control* study.

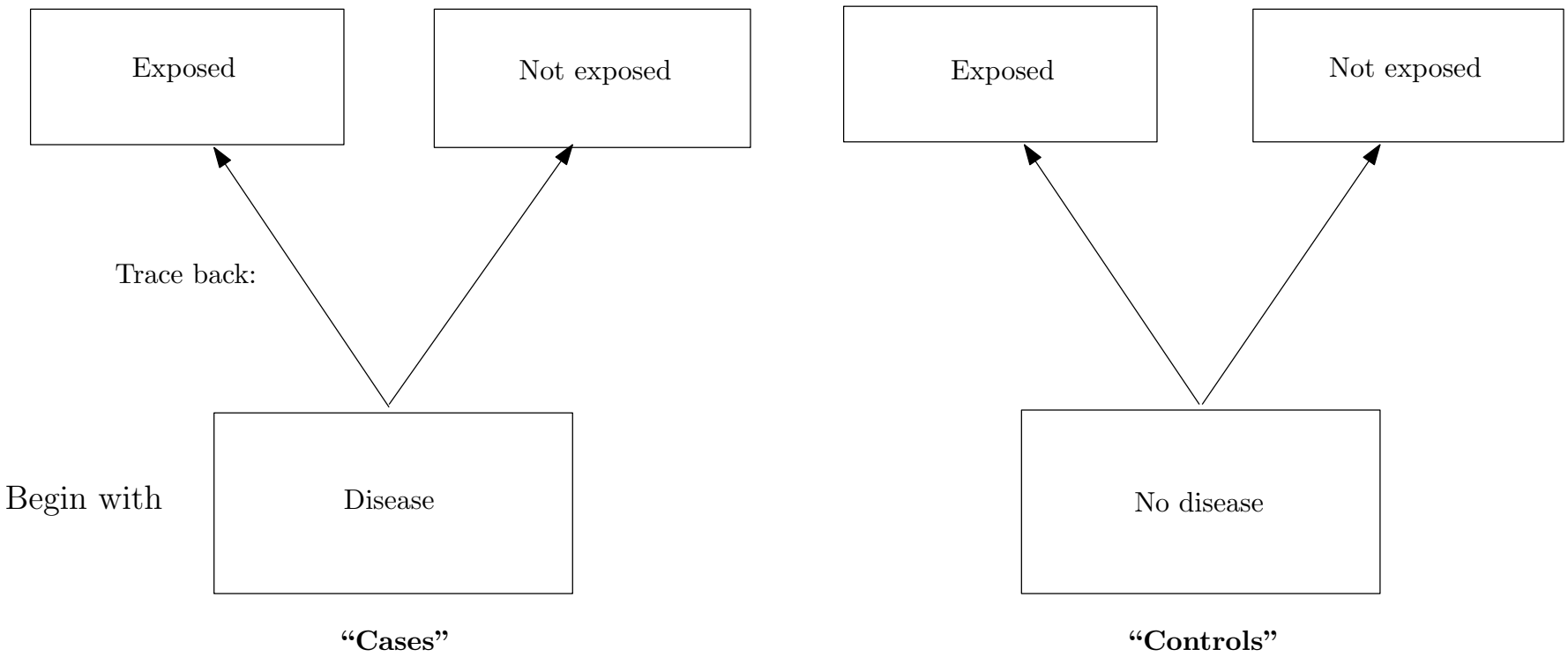
Cohort study

A cohort study is often termed a *prospective study*, since the individuals are followed prospectively in time.



Case-control study

A case-control study is known as a *retrospective study*.



Cohort study vs. *Case-control study*

- Major disadvantages of cohort study:
  1. Cohort study takes longer time to complete.
  2. When the disease under study is rare, a cohort study will need a much larger number of subjects in order to ensure that there is a sufficient number of cases at the end of the study.
  3. Generally, it is costly to conduct a cohort study compared to a case-control study.

Cohort study vs. *Case-control study*

- Major disadvantages of case-control study:
  1. Case-control study is restricted to one disease, whereas information of a number of different diseases could be collected in the cohort study.
  2. *Bias:* In the case-control design, considerable care must be taken to ensure the cases and controls used are representative of the underlying populations of individuals. In particular, the probability that an individuals is included in the case-control study as either a case or a control must not be associated with exposure factors of interest.
  3. In certain studies, case-control design relies on a subject's recall of past experiences or on suitable documentary evidence of past exposure being available, which could lead to unreliable data.

An example of case-control study

- There is a suspicion that zinc oxide, the white non-absorbent sunscreen traditionally worn by lifeguards is more effective at preventing sunburns that lead to skin cancer than absorbent sunscreen lotions.
- A case-control study was conducted to investigate if exposure to zinc oxide is a more effective skin cancer prevention measure. The study involved comparing a group of 147 former lifeguards that had developed cancer on their cheeks and noses (cases) to a group of 173 lifeguards without this type of cancer (controls).
- A questionnaire was administered to each of the 320 lifeguards in the study about the prior exposure to zinc oxide or absorbent lotion.

Table 1: Data from a study to examine the association between skin cancer and zinc oxide exposure.

Exposure	Cases	Controls
lotion	86	78
zinc oxide	61	95
Total	147	173

Measures of association between disease and exposure

- Risk of disease = Incidence of disease
  The probability of a disease occurring during a given period of time.
- Denote the risk of disease among exposed p_e and the risk of disease among unexposed p_u,

$$\textit{Relative Risk of disease } (RR): \rho = p_e / p_u$$

- Besides RR, another useful measure of association is the odds ratio,

$$\phi = \frac{p_e/(1-p_e)}{p_u/(1-p_u)}$$

- Both risk and odds ratio could be estimated/interpreted in a cohort study, but only odds ratio could be estimated from the results of a case-control study. This is a reason why the odds ratio is so important.

Measures of association between disease and exposure

- Risk of disease = Incidence of disease
  The probability of a disease occurring during a given period of time.
- Denote the risk of disease among exposed p_e and the risk of disease among unexposed p_u,

$$\textit{Relative Risk of disease } (RR): \rho = p_e / p_u$$

- Besides RR, another useful measure of association is the odds ratio,

$$\phi = \frac{p_e/(1-p_e)}{p_u/(1-p_u)}$$

- Both risk and odds ratio could be estimated/interpreted in a cohort study, but only odds ratio could be estimated from the results of a case-control study. This is a reason why the odds ratio is so important.

Study design matters in applying the linear logistic model

Let $p(\boldsymbol{x})$ be the probability that the individual develops the disease, and $x_1, ..., x_k$ be k explanatory variables measured,

$$\text{logit}(p(\boldsymbol{x})) = \log(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Let's see what we can learn about this model when the data is from a cohort study, and when the data is from a case-control study respectively.

The linear logistic model for data from a cohort study

- If data is from cohort studies, since the proportions of diseased ($Y = 1$) in the study are comparable to the proportions of diseased people in the population, a fitted logistic regression model using the observed data from a cohort study can be used to estimate the probability of disease.
- The coefficient for the intercept could be interpreted as log odds when covariates are set to be the baseline, the coefficients for covariates can be interpreted as logarithms of odds or log odds ratios, and we can estimate/predict $p(\boldsymbol{x})$ based on log odds.

The linear logistic model for data from case-control studies

- Because in a case-control study, the proportion of cases in the study will not be comparable to the proportion of diseased people in the population, a fitted linear logistic model cannot be used to estimate the probability of disease, odds of disease, nor the relative risk of disease.
- However, the model enables estimates of odds ratios to be determined.

Why odds ratio is still valid in linear logistic model?

What are we actually estimating if we fit the logistic regression model using data obtained from a case-control study (case=1, control=0)?

Consider the probability that an individual in the case-control study has the disease:

$$p_0(\boldsymbol{x}) = \mathrm{P}[\ \textit{diseased}\ |\ \textit{individuals with covariate values } \boldsymbol{x} \textit{ in the study}]$$

What we are actually fitting is the logistic regression model for $p_0(\boldsymbol{x})$:

$$\mathrm{logit}(p_0(\boldsymbol{x})) = \log(\frac{p_0(\boldsymbol{x})}{1-p_0(\boldsymbol{x})}) = \beta_0' + \beta_1' x_1 + \beta_2' x_2 + \cdots + \beta_k' x_k.$$

Why odds ratio is still valid in linear logistic model?

What are we actually estimating if we fit the logistic regression model using data obtained from a case-control study (case=1, control=0)?

Consider the probability that an individual in the case-control study has the disease:

$$p_0(\boldsymbol{x}) = \mathrm{P}[\ \textit{diseased}\ |\ \textit{individuals with covariate values } \boldsymbol{x} \textit{ in the study}]$$

What we are actually fitting is the logistic regression model for $p_0(\boldsymbol{x})$:

$$\mathrm{logit}(p_0(\boldsymbol{x})) = \log(\frac{p_0(\boldsymbol{x})}{1-p_0(\boldsymbol{x})}) = \beta_0' + \beta_1' x_1 + \beta_2' x_2 + \cdots + \beta_k' x_k.$$

The linear logistic model for data from case-control studies

- Recall p is the probability of disease in the population, what we really want is

$$\text{logit}(p(\boldsymbol{x})) = \log(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \quad (3)$$

- This is what we are actually fitting if data is from a case-control study:

$$\text{logit}(p_0(\boldsymbol{x})) = \log(\frac{p_0(\boldsymbol{x})}{1-p_0(\boldsymbol{x})}) = \beta_0' + \beta_1' x_1 + \beta_2' x_2 + \cdots + \beta_k' x_k. \quad (4)$$

- Are there any relationship between those two models?

The linear logistic model for data from case-control studies

Define $\pi_1 = \mathrm{P}[\ \textit{individual is in the study} \mid \textit{diseased}\ ]$,
$\pi_2 = \mathrm{P}[\ \textit{individual is in the study} \mid \textit{diseased-free}\ ]$

By Bayes theorem,

$$p_0(\boldsymbol{x}) = \frac{\pi_1 p(\boldsymbol{x})}{\pi_1 p(\boldsymbol{x}) + \pi_2\{1 - p(\boldsymbol{x})\}}.$$

From this result,

$$\frac{p_0(\boldsymbol{x})}{1 - p_0(\boldsymbol{x})} = \frac{\pi_1}{\pi_2}\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})},$$

The linear logistic model for data from case-control studies (continued)

and it then follows that

$$\text{logit}\{p_0(x)\} = \log(\frac{\pi_1}{\pi_2}) + \text{logit}\{p(x)\} \tag{5}$$

If the logistic model for $p(\boldsymbol{x})$ is

$$\text{logit}(p(\boldsymbol{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

then the model for $p_0(\boldsymbol{x})$ is given by

$$\text{logit}(p_0(\boldsymbol{x})) = \log(\frac{\pi_1}{\pi_2}) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{6}$$

Comparing model (5) with model (6), one could obtain $\beta_0' = \log(\frac{\pi_1}{\pi_2}) + \beta_0$, and $\beta_j' = \beta_j, j = 1, ..., k$. In case-control studies, the β_0 cannot be interpreted as log odds when $X = 0$ either.

The linear logistic model for data from case-control studies (continued)

- When fitting a logistic model for data from case-control studies, $\hat{p}_0(\boldsymbol{x})$ could be determined (which is of no interest).
- What we are truly interested is $p(\boldsymbol{x})$, but it depends on the unknown $\frac{\pi_1}{\pi_2}$, and could not be estimated. It follows that **the relative risk**, $\rho = \frac{p(\boldsymbol{x}_1)}{p(\boldsymbol{x}_0)}$ couldn't be estimated, and **the odds** $\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}$ **couldn't be estimated**.
- However, the ratio of the odds of disease for a person with $\boldsymbol{x}_1$ relative to someone with $\boldsymbol{x}_0$ could be estimated since

$$\phi = \frac{p(\boldsymbol{x}_1)/\{1-p(\boldsymbol{x}_1)\}}{p(\boldsymbol{x}_0)/\{1-p(\boldsymbol{x}_0)\}} = \frac{p_0(\boldsymbol{x}_1)/\{1-p_0(\boldsymbol{x}_1)\}}{p_0(\boldsymbol{x}_0)/\{1-p_0(\boldsymbol{x}_0)\}}$$

A case-control example: Output Comparison

```
. cci 86 61 78 95, woolf

                |    Exposed    Unexposed  |       Total  Proportion exposed
----------------+-------------------------+-------------------------------
          Cases |          86           61  |         147          0.5850
       Controls |          78           95  |         173          0.4509
----------------+-------------------------+-------------------------------
          Total |         164          156  |         320          0.5125
                |                         |
                |      Point estimate     |    [95% Conf. Interval]
                |-------------------------+-------------------------------
     Odds ratio |            1.717108     |    1.101226     2.677433 (Woolf)
 Attr. frac. ex.|            .4176255     |    .0919215     .6265079 (Woolf)
 Attr. frac. pop|            .2443251     |
                +-----------------------------------------------------------
                           chi2(1) =     5.73   Pr>chi2 = 0.0167

. csi 86 61 78 95, woolf

                |    Exposed    Unexposed  |       Total
----------------+-------------------------+-----------
...
           Risk |   .5243902     .3910256  |     .459375
                |                         |
                |      Point estimate     |    [95% Conf. Interval]
                |-------------------------+-------------------------------
 Risk difference|            .1333646     |    .0251717     .2415575
     Risk ratio |            1.341064     |    1.050581     1.711864
 Attr. frac. ex.|            .2543232     |    .0481454     .4158416
 Attr. frac. pop|            .1487877     |
                +-----------------------------------------------------------
                           chi2(1) =     5.73   Pr>chi2 = 0.0167
```

Output Comparison

Exposure	Cases	Controls
lotion	86	78
zinc oxide	61	95
Total	147	173

Stata code/output

```
clear
input x y n
1 86 164
0 61 156
end

 . blogit y n x

Logistic regression for grouped data              Number of obs     =        320
                                                  LR chi2(1)        =       5.75
                                                  Prob > chi2       =     0.0165
Log likelihood = -217.87686                       Pseudo R2         =     0.0130
------------------------------------------------------------------------------
     _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
           x |    .5406415   .2266455     2.39   0.017     .0964245    .9848585
       _cons |   -.443003   .1640724    -2.70   0.007    -.7645791   -.121427
------------------------------------------------------------------------------
```

What are the connections of the output from logistic regression and the output from the 2x2 table?

Summary

- **Overdispersion**:
  - could be caused by extra variation in response probabilities that cannot be explained by the current model.
  - It may cause inflated type I error rates if not being corrected.
  - There are alternative models for correcting overdispersion and this could be a dense topic

- **Cohort study versus case-control study**:
  - Due to “distortion" (manipulation) of diseased probability in the case-control study, risk/probability, relative risk (risk ratio) and odds measures are no longer estimable from the case-control study.
  - Odd ratio estimates are still valid and interpretable.