Lecture 17: Logistic Regression: Testing Homogeneity of the OR

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Testing for homogeneity of the OR across strata

- Recall, in the previous lecture we were interested in estimating the "common" (or adjusted) OR using a logistic model
- In doing so, we assumed that the OR remained the same for each level of our confounding variable j
- Suppose we again think of the data as arising from J, (2×2) tables:

Stratum j (of W)

Variable (X)	Variable (Y): 1	2	
1	Y_{j11}	Y_{j12}	Y_{j1+}
2	Y_{j21}	Y_{j22}	Y_{j2+}
	Y_{j+1}	Y_{j+2}	Y_{j++}

- Let $OR_{XY(j)} = OR_j^{XY.W}$ be the odds ratio for the subtable where $W = j$.
- Previously, we have assumed a common odds ratio across the J strata of W, and we have estimated this common odds ratio, and tested if it equals 0.
- Often, you will not know that there is a common odds ratio across the J strata, and you will want to test for the 'homogeneity' of the odds ratio across the J strata (homogeneity means a common odds ratio):

$$\mathsf{H}_0 : OR_1^{XY.W} = OR_2^{XY.W} = ... = OR_J^{XY.W}$$

- One can rewrite this null hypothesis as

$$\begin{aligned} \mathsf{H}_0: \quad & OR_1^{XY.W} = OR_2^{XY.W}, \\ & OR_1^{XY.W} = OR_3^{XY.W} \\ & \ldots \\ & OR_1^{XY.W} = OR_J^{XY.W} \end{aligned}$$

- This second form implies the first; note, there are only $J-1$ statements in the second form (this is seen by looking to the right of the '=' sign, which goes from 2 to J.)
- Then, we see (from the second form) that this will be a $J-1$ degree of freedom test. ($J-1$ statements)

- This can also be seen using the appropriate logistic regression model
- Recall, the logistic model when there was a common odds ratio was

$$\text{logit}\{P[Y=1|W=j,X=x]\} = \beta_0 + \alpha_j + \beta x$$

$$= \beta_0 + \alpha_1 w_1 + \alpha_2 w_2 + ... + \alpha_{J-1} w_{J-1} + \beta x$$

- and, the common odds ratio was

$$\exp(\beta) = OR_j^{XY.W}.$$

- Now, suppose we add interaction terms in, between W and X. There are $J-1$ covariates corresponding to STRATA (AGE), and 1 corresponding to the covariate of importance (VACCINE). Then, the interaction will have $(J-1)\cdot 1$ terms:

$$\text{logit}\{P[Y=1|W=j,X=x]\} = \beta_0 + \alpha_j + \beta x + \gamma_j x$$

$$= \beta_0 + \alpha_1 w_1 + \alpha_2 w_2 + ... + \alpha_{J-1} w_{J-1} + \beta x + \gamma_1 w_1 x + \gamma_2 w_2 x + ... + \gamma_{J-1} w_{J-1} x$$

  where we impose the constraints $\alpha_J = 0$ and $\gamma_J = 0$, and the γ_j are the interaction parameters.

- In practice, you create $J-1$ dummy variables for AGE (W), and then multiply each dummy age indicator by x to get the $J-1$ interaction columns.

- If the interaction model holds, it means that there is a different odds ratio for each strata (level $W = j$), thus, the odds ratios are not the same (homogeneous) across strata.
- In this model, the log-odds ratio between VACCINE (X) and PARALYSIS (Y), controlling for AGE $(W = j)$, is

$$\log OR_j^{XY.W} =$$

$$\log\left(\frac{P[Y=1|W=j,x=1]}{1-P[Y=1|W=j,x=1]}\right) - \log\left(\frac{P[Y=1|W=j,x=0}{1-P[Y=1|W=j,x=0}\right) =$$

$$\text{logit}\{P[Y=1|W=j,x=1\}-$$
$$\text{logit}\{P[Y=1|W=j,x=0\} =$$

$$[\beta_0 + \alpha_j + \beta(1) + \gamma_j(1)] - [\beta_0 + \alpha_j + \beta(0) + \gamma_j(0)]$$

$$= \beta + \gamma_j,$$

which depends on $W = j$, and thus the log-odds ratios are not constant across strata.

- If the OR is constant in each strata, then each $\gamma_j = 0$ so that

$$\log OR_j^{XY.W} = \beta \ ,$$

  for all j.
- The hypothesis of a constant odds ratio across strata is

$$\mathsf{H}_0 : \gamma_1 = \cdots \gamma_{J-1} = 0.$$

- Note, subtracting terms, you can show that

$$\gamma_j = \log[OR_j^{XY.W}] - \log[OR_J^{XY.W}]$$

- Then, the null hypothesis says that

$$\gamma_j = \log[OR_j^{XY \cdot W}] - \log[OR_J^{XY \cdot W}] = 0,$$

  for $j = 1, ..., J-1$.

- Or, equivalently,

$$\log[OR_j^{XY \cdot W}] = \log[OR_J^{WY \cdot X}],$$

  for $j = 1, ..., J-1$ and, thus all $OR_j^{XY \cdot W}$'s are equal to each other, i.e., homogeneity.

Test Statistics for Homogeneity

- Using maximum likelihood, we can use either the WALD, SCORE, or Likelihood Ratio Statistic.
- Note, the alternative (non-homogeneity) model with interaction is a 'saturated' model.
- If we consider the row margins in each of the J (2×2) table as fixed, then we have $2J$ binomial populations, and thus $2J$ binomial probabilities in the saturated models.
- The interaction model has $2J$ parameters, and thus is a saturated model.

$$1 \quad \beta_0$$

$$(J-1) \quad \alpha_j\text{'s} \qquad (\alpha_J = 0)$$

$$1 \quad \beta$$

$$(J-1) \quad \gamma_j\text{'s} \qquad (\gamma_J = 0),$$

which gives $1 + (J-1) + 1 + (J-1) = J + J = 2J$ parameters.

- Then, under the alternative of non-homogeneity, we have a saturated model so that the estimate

$$\widehat{P}[Y=1|W=j,X=k]=\widehat{p}_{jk}=\frac{Y_{jk}}{n_{jk}}.$$

- If we let $\tilde{p}_{jk}$ be the estimate of p_{jk} under homogeneity, the likelihood ratio statistic for homogeneity is 2 times the difference of the log likelihoods under the saturated and homogeneity models. This is the deviance for the homogeneity model (the model without the interaction terms), $D^2(w_1,...,w_{J-1},x)$,

$$D^2=2\sum_{j=1}^{J}\sum_{k=1}^{2}\left[y_{jk}\log\left(\frac{y_{jk}}{n_{jk}\tilde{p}_{jk}}\right)+(n_{jk}-y_{jk})\log\left(\frac{n_{jk}-y_{jk}}{n_{jk}(1-\tilde{p}_{jk})}\right)\right]$$

- Note, this statistic is again of the form

$$D^2 = 2\sum O \log\left(\frac{O}{E}\right)$$

- Similarly, the score statistic (Pearson's) is

$$\begin{aligned} X^2 &= \sum_{j=1}^{J}\sum_{k=1}^{2}\left(\frac{[y_{jk}-n_{jk}\tilde{p}_{jk}]^2}{n_{jk}\tilde{p}_{jk}(1-\tilde{p}_{jk})}\right) \\ &= \sum_{j=1}^{J}\sum_{k=1}^{2}\left[\frac{[y_{jk}-n_{jk}\tilde{p}_{jk}]^2}{n_{jk}\tilde{p}_{jk}} + \frac{[(n_{jk}-y_{jk})-n_{jk}(1-\tilde{p}_{jk})]^2}{n_{jk}(1-\tilde{p}_{jk})}\right] \end{aligned}$$

- Note, this statistic is again of the form

$$X^2 = \sum\left(\frac{[O-E]^2}{E}\right)$$

Example - Age, Vaccine, Paralysis Data

Age	Salk Vaccine	Paralysis: No	Paralysis: Yes
0-4	Yes	20	14
	No	10	24
5-9	Yes	15	12
	No	3	15
10-14	Yes	3	2
	No	3	2
15-19	Yes	12	3
	No	7	5
20+	Yes	1	0
	No	3	2

SAS Proc Logistic

```
data one;
 input age vac para count;
cards;
1  1  1 20
1  1  0 14
1  0  1 10
1  0  0 24
2  1  1 15
2  1  0 12
2  0  1  3
2  0  0 15
3  1  1  3
3  1  0  2
3  0  1  3
3  0  0  2
4  1  1 12
4  1  0  3
4  0  1  7
4  0  0  5
5  1  1  1
5  1  0  0
5  0  1  3
5  0  0  2
;
run;
```

```
proc logistic descending;
  class age (PARAM=ref) ;
  model para = age vac / aggregate scale=d /* specify for deviance */ ;
 freq count;
run;

/* SELECTED OUTPUT */

       Deviance and Pearson Goodness-of-Fit Statistics

Criterion        DF           Value     Value/DF     Pr > ChiSq

Deviance          4          1.9547       0.4887          0.7441
Pearson           4          1.8456       0.4614          0.7641 (=SCORE)
```

Here DF=4 since we have left out of the model the 4 (=5-1) interaction terms

Class Level Information

Class	Value	Design Variables			
age	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	5	0	0	0	0

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	45.4266	5	<.0001
Score	43.0142	5	<.0001
Wald	38.3369	5	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	4	20.0456	0.0005
vac	1	26.3875	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	0.5245	0.6162	0.7244	0.3947
age	1	1	-1.4273	0.6510	4.8078	0.0283
age	2	1	-1.7473	0.6708	6.7852	0.0092
age	3	1	-0.7181	0.7813	0.8448	0.3580
age	4	1	-0.2905	0.6932	0.1756	0.6752
vac		1	1.2830	0.2498	26.3875	<.0001

Wald Statistic

- Now, let's look at the WALD statistic for no interaction

$$\mathsf{H}_0 : \gamma_1 = \cdots \gamma_{J-1} = 0,$$

in the model

$$\text{logit}\{P[Y = 1|W = j, X = x]\} = \beta_0 + \alpha_j + \beta x + \gamma_j x$$

- Previously, we discussed that

$$\gamma_j = \log[OR_j^{XY.W}] - \log[OR_J^{XY.W}]$$

- Since the model is saturated, the estimated OR's equal the observed OR's, i.e„

$$\widehat{OR}_j^{XY.W} = \frac{y_{j11}y_{j22}}{y_{j12}y_{j21}},$$

and

$$\hat{\boldsymbol{\gamma}}_j = \log[\widehat{OR}_j^{XY.W}] - \log[\widehat{OR}_J^{XY.W}]$$

- If we form the $(J-1) \times 1$ vector

$$\gamma = [\gamma_1, \cdots, \gamma_{J-1}]',$$

then, the MLE of γ is

$$\hat{\boldsymbol{\gamma}} = [\hat{\boldsymbol{\gamma}}_1, \cdots, \hat{\boldsymbol{\gamma}}_{J-1}]',$$

- Then, the WALD statistic for

$$\mathsf{H}_0 : \gamma_1 = \cdots \gamma_{J-1} = 0,$$

is

$$W^2 = \hat{\boldsymbol{\gamma}}^T [\widehat{Var}(\hat{\boldsymbol{\gamma}})]^{-1} \hat{\boldsymbol{\gamma}} \sim \chi^2_{J-1},$$

where $\widehat{Var}(\hat{\boldsymbol{\gamma}})$ is obtained from the inverse of the information matrix [the lower right hand $(J-1) \times (J-1)$ block].

- In the WALD statistic, we are basically comparing the observed log OR's

$$\log[\widehat{OR}_j^{WY.X}]$$

  to each other to see how similar they are.

- To estimate the proposed WALD statistic, we will use a contrast statement in PROC LOGISTIC

SAS Proc Logistic for WALD

Model modification with interaction term:

```
proc logistic descending;
  class age (PARAM=ref) ;
  model para = age vac age*vac ;
  contrast 'homogeneity'
          age*vac 1 0 0 0, age*vac 0 1 0 0, age*vac 0 0 1 0,
          age*vac 0 0 0 1;
 freq count;
run;
```

Although the contrast looks complicated, it essentially represents a test that the sum of the γ's equals 0.

age*vac $\vec{\gamma} = [\gamma_1 \gamma_2 \gamma_3 \gamma_4]$, $c = [1, 1, 1, 1]^T$.

```
/* SELECTED OUTPUT */
             Analysis of Maximum Likelihood Estimates

                                        Standard
Parameter          DF    Estimate         Error    Chi-Square    Pr > ChiSq

Intercept           1      0.4055        0.9129        0.1973        0.6569
age          1      1     -1.2809        0.9874        1.6829        0.1945
age          2      1     -2.0149        1.1106        3.2918        0.0696
age          3      1    -256E-17        1.2910        0.0000        1.0000
age          4      1     -0.0690        1.0845        0.0040        0.9493
vac                 1      9.7976         164.3        0.0036        0.9524
vac*age      1      1     -8.5655         164.3        0.0027        0.9584
vac*age      2      1     -7.9651         164.3        0.0024        0.9613
vac*age      3      1     -9.7976         164.3        0.0036        0.9524
vac*age      4      1     -8.7478         164.3        0.0028        0.9575
WARNING: The validity of the model fit is questionable.
```

–Recall - some cells had low counts, some with zero. That's the "questionable" fit. Also note the parameter estimate for Age = 3. Conditional logistic, which will be studied shortly, will address this limitation.

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
homogeneity	4	1.6112	0.8068

Problems with Test Statistics for Homogeneity

- The WALD, SCORE, and LIKELIHOOD RATIO test statistics are approximately equivalent in large samples under the null. The values for the example are:

```
STATISTIC          DF     Chi-Square     Pr > ChiSq

DEVIANCE (LR)      4         1.9547         0.7441
PEARSON (SCORE)    4         1.8456         0.7641
WALD               4         1.6112         0.8068
```

- However, these statistics require large samples:
- The likelihood ratio and score tests both compare y_{jk} to $n_{jk}\tilde{p}_{jk}$, and require n_{jk} to be fairly large. Even under the null, we need pretty large n_{jk} since we must estimate β_0, α_j, and β.
- The WALD statistic has worse properties than the other 2; if one observed cell count is 0,

$$\log[\widehat{OR}_j^{WY.X}] = \pm\infty$$

and the test statistic is indeterminate (see above example). One solution is to add .5 to each cell count (which, again, people don't like to do since it is like adding 'Fake Data')

- Even when there is sufficient data, tests for interactions tend not to be as powerful as tests for main effects:
- If all the odds ratios are in the same direction (although not identical), we will usually not reject the null of no interaction, and can estimate a 'pooled' OR.
- **Possible Remedies**
- 1. Use statistics that are less sensitive to small cell counts (conditional or exact conditional methods, which we will discuss later);
- 2. If the stratum variable is ordinal (such as age in the above example), you can take the ordering into account with a logistic regression model that does not have indicator variables for each level of W, but, instead, a continuous value w, such as in the test for trend in a $(J \times 2)$ table.

Testing for Homogeneity: Ordinal data

- When the strata are ordinal, such as in the VACCINE data, one can take this ordinality into account via the logistic regression model:

$$\text{logit}\{P[Y=1|W=j,x]\} = \beta_0 + \alpha w_j + \beta x + \gamma w_j x,$$

  where w_j is an ordered value corresponding to level j of W.

- In this model, the log-odds ratio between VACCINE (X) and PARALYSIS (Y), given AGE$= w_j$, is

$$\log\left(\frac{P[Y=1|W=j,x=1]}{1-P[Y=1|W=j,x=1]}\right) - \log\left(\frac{P[Y=1|W=j,x=0}{1-P[Y=1|W=j,x=0}\right) =$$

$$\text{logit}\{P[Y=1|W=j,x=1\}-$$
$$\text{logit}\{P[Y=1|W=j,x=0\} =$$

$$= [\beta_0 + \alpha w_j + \beta(1) + \gamma w_j(1)]-$$
$$[\beta_0 + \alpha w_j + \beta(0) + \gamma w_j(0)]$$

$$= \beta + \gamma w_j,$$

- Then, a test for a common odds ratio here has 1 degree of freedom:

$$H_0 : \gamma = 0.$$

- Then, you can again use a WALD, LIKELIHOOD RATIO, or SCORE TEST.
- Recall the Age, Vaccine, Paralysis data:
- Treating age as categorical (with 5 levels, and 4 dummy variables), the likelihood ratio statistic for homogeneity of the odds ratio is

$$D^2(w_1, ..., w_{J-1}, x) = 1.95 \qquad df = 4, \qquad p = 0.744$$

- Now, we treat age as continuous, giving the midpoint of the age interval as the covariate,

$$w_j = \begin{cases} 2 \text{ if age 0-4} \\ 7 \text{ if age 5-9} \\ 12 \text{ if age 10-14} \\ 17 \text{ if age 15-19} \\ 22 \text{ if age } 20^+ \end{cases} .$$

in the model

$$\text{logit}\{P[Y = 1|w_j, x]\} = \beta_0 + \alpha w_j + \beta x + \gamma w_j x,$$

and testing for

$$H_0 : \gamma = 0$$

Using SAS

```
data one;
 input age vac para count;
 agec = 2 + 5*(age-1);   /* continuous age = midpoint */
                         /* of interval               */
cards;
1  1  1 20
1  1  0 14
1  0  1 10
1  0  0 24
2  1  1 15
2  1  0 12
2  0  1  3
2  0  0 15
3  1  1  3
3  1  0  2
3  0  1  3
3  0  0  2
4  1  1 12
4  1  0  3
4  0  1  7
4  0  0  5
5  1  1  1
5  1  0  0
5  0  1  3
5  0  0  2
;
```

```
proc logistic descending;
   model para = agec vac/ aggregate scale=d /* specify for deviance */   ;
 freq count;
run;

/* SELECTED OUTPUT */

       Deviance and Pearson Goodness-of-Fit Statistics

Criterion        DF           Value     Value/DF     Pr > ChiSq

Deviance          7          5.2776       0.7539         0.6261
Pearson           7          4.8683       0.6955         0.6760
```

There are 10 binomial parameters (2*5), and we have fit a model with 3 ($\beta_0, \beta_{vac}, \beta_{age}$) which leaves 7 parameters free. For this model, we do not reject the null hypothesis of a "good" fit.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.2141	0.2990	16.4896	<.0001
agec	1	0.0745	0.0245	9.2404	0.0024
vac	1	1.1983	0.3002	15.9376	<.0001

```
proc logistic descending;
   model para = agec vac agec*vac/ aggregate scale=d   ;
 freq count;
run;

/* SELECTED OUTPUT */

       Deviance and Pearson Goodness-of-Fit Statistics

Criterion        DF          Value     Value/DF     Pr > ChiSq

Deviance          6         5.1545       0.8591          0.5242
Pearson           6         4.7249       0.7875          0.5796
```

The interaction of age and vaccine has been added to the model. It is not a saturated model, so we will need to compare deviances to assess added fit.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.2851	0.3731	11.8667	0.0006
agec	1	0.0830	0.0346	5.7552	0.0164
vac	1	1.3448	0.5042	7.1136	0.0077
agec*vac	1	-0.0200	0.0526	0.1440	0.7044

- The likelihood ratio statistic for

$$H_0 : \gamma = 0$$

is

$$D^2(w_j, x) - D^2(w_j, x, w_j x) =$$
$$5.28 - 5.15 = .13$$

$$df = 1, \quad p = 0.73$$

(The WALD STAT = $Z^2 = 0.144$ is obtained from the PARAMETER ESTIMATES section of the output since we have df.).

- Thus, for these data, the common odds ratio assumption seems valid.

General Logistic Model

- We have discussed logistic regression in specific situations, i.e., $(R \times 2)$ tables and J, (2×2) tables.
- These two examples are all subsumed in the general logistic regression model.
- The response for individual i is

$$Y_i = \begin{cases} 1 \text{ if success (column 1)} \\ 0 \text{ if failure (column 2)} \end{cases} .$$

- The covariate vector for individual i is

$$\mathbf{x}_i = [x_{i1}, ..., x_{iK}]',$$

  where x_{ik} is the k^{th} covariate for individual i.
- The x_{ik}'s can be continuous or categorical (indicator covariates).

Example–Arthritis Clinical Trial

- This example is from an arthritis clinical trial comparing the drug auranofin and placebo therapy for the treatment of rheumatoid arthritis (Bombardier, et al., 1986).
- The response of interest is the self-assessment of arthritis, classified as (0) poor or (1) good.
- Individuals were also given a self-assessment at baseline (before treatment), which was also classified as (0) poor or (1) good.
- randomized into one of the two treatment groups after baseline self-assessment of arthritis (with the same 2 levels as the response).
- The dataset contains 293 patients who were observed at both baseline and 13 weeks. The data from 25 cases are as follows:

Arthritis Data

Subset of cases from the arthritis clinical trial

				Self assessment[b]	
CASE	SEX	AGE	TREATMENT[a]	BASELINE	13 WK.
1	M	54	A	0	0
2	M	64	P	0	0
3	M	48	A	1	1
4	F	41	A	1	1
5	M	55	P	1	1
6	M	64	A	1	1
7	M	64	P	1	0
8	F	55	P	1	1
9	M	39	P	1	0
10	F	60	A	0	1
11	M	49	A	0	1
12	M	32	A	0	1
13	F	62	P	0	0
			...		

[a] A = Auranofin, P = Placebo , [b] 0=poor, 1=good.

We are interested in seeing how the binary response

$$Y_i = \begin{cases} 1 \text{ if good at 13 weeks} \\ 0 \text{ if poor at 13 weeks} \end{cases}.$$

is affected by the covariates,
1. BASELINE self-assessment:

$$X_i = \begin{cases} 1 \text{ if good at BASELINE} \\ 0 \text{ if poor at BASELINE} \end{cases}.$$

2. AGE IN YEARS,

3. GENDER

$$\text{SEX} = \begin{cases} 1 \text{ if male} \\ 0 \text{ if female} \end{cases} .$$

4. TREATMENT

$$\text{TRT} = \begin{cases} 1 \text{ if auranofin} \\ 0 \text{ if placebo} \end{cases} .$$

The main question is whether the treatment increases the probability of a more favorable response, after controlling for baseline response; secondary questions are whether the response differs by age and sex.

In general, the covariates can be

1. Qualitative (discrete), as is treatment, gender, and baseline response, in this example.

2. Quantitative (continuous), as is age in this example

3. Ordinal (ordered categorical). Suppose, instead of continuous age, we had three age groups

Category	Age Range
0	21-47
1	48-56
2	57-66

These categories correspond to the bottom, middle, and top third of the ages (21 is the minimum, 66 is the maximum). This grouping is categorical, but there is an ordering.

Distribution of Data

- The distribution of the response for subject i is

$$Y_i \sim Bern(p_i),$$

where $p_i = P[Y_i = 1|x_{i1}, ..., x_{iK}]$ follows the logistic regression model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

- The model here contains an intercept; but it doesn't have to.
- As before, the parameter β_k has the general interpretation as a conditional log-odds ratio between the response and a one unit increase in the covariate x_{ik}, conditional on the other covariates,
- For now, for ease of explanation, we drop the subscript i :

$$\text{logit}(P[Y = 1|x_1, ..., x_K]) = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K$$

Interpretation of β_k

- Consider the two logits,

$$\text{logit}(P[Y = 1|x_1, ..., x_k = c + 1, ..., x_K]) = \beta_0 + \beta_1 x_1 + ... + \beta_k (c + 1) ... + \beta_K x_K$$

and

$$\text{logit}(P[Y = 1|x_1, ..., x_k = c, ..., x_K]) = \beta_0 + \beta_1 x_1 + ... + \beta_k c ... + \beta_K x_K,$$

- The covariate values are the same in the two logits, except that we have increased x_k by one unit in the first logit (say, group 1) over the second logit (say, group 2).

- The log-odds ratio for the two groups is the difference in the logits

$$\log\left[\frac{P[Y=1|x_1,\ldots,x_k=c+1,\ldots,x_K]/(1-P[Y=1|x_1,\ldots,x_k=c+1,\ldots,x_K])}{P[Y=1|x_1,\ldots,x_k=c,\ldots,x_K]/(1-P[Y=1|x_1,\ldots,x_k=c,\ldots,x_K])}\right] =$$

$$\text{logit}(P[Y=1|x_1,...,x_k=c+1,...,x_K])-$$
$$\text{logit}(P[Y=1|x_1,...,x_k=c,...,x_K]) =$$

$$[\beta_0+\beta_1x_1+...+\beta_k(c+1)...+\beta_Kx_K]-$$
$$[\beta_0+\beta_1x_1+...+\beta_kc...+\beta_Kx_K]=\beta_k$$

- Thus,

$$\beta_k$$

  is the log-odds ratio for a one-unit increase in covariate x_k, given all the other covariates are the same.
- For example, if x_k is a dichotomous covariate which equals 1 for the new treatment, and 0 for placebo, the β_k is the log-odds ratio for success for new treatment versus placebo, conditional on the other covariates being the same.
- Since we are now in the regression framework, we may be interested in interaction models.

Interaction

- Suppose (without loss of generality) there is an interaction between $x_{i,K-1}$ and x_{iK} :

$$\text{logit}(p_i) =$$

$$\beta_0 + \beta_1 x_{i1} + ... + \beta_{K-1} x_{i,K-1} + \beta_K x_{iK} + \gamma x_{i,K-1} x_{iK}$$

- To give the interpretation of γ, for simplicity, we now drop the subscript i and compare the two logits, when $x_k = c + 1$ and $x_k = c$:

$$\text{logit}(P[Y = 1|x_1, ..., x_{K-1}, x_K = c + 1]) =$$
$$\beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K (c + 1) + \gamma x_{K-1} (c + 1)$$

and

$$\text{logit}(P[Y = 1|x_1, ..., x_{K-1}, x_K = c]) =$$
$$\beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K c + \gamma x_{K-1} c$$

- The covariate values are the same in the two logits, except that we have increased the last covariate, x_K by one unit in the first logit (say, group 1) over the second logit (say, group 2).

- The log-odds ratio is the difference in the logits

$$\text{logit}(P[Y = 1|x_1, ..., x_{K-1}, x_K = c + 1]) -$$
$$\text{logit}(P[Y = 1|x_1, ..., x_{K-1}, x_K = c]) =$$

$$[\beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K (c + 1) + \gamma x_{K-1}(c + 1)] -$$
$$[\beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K c + \gamma x_{K-1} c] =$$

$$\beta_k + \gamma x_{K-1}$$

- Thus, conditional on the first $(K - 1)$ covariates, the log odds ratio for a one unit increase in the K^{th} covariate is

$$\beta_K + \gamma x_{K-1},$$

and depends on the level of the $(K - 1)^{st}$ covariate.

- If

$$\gamma = 0,$$

  then the conditional OR between Y and x_K does not depend on x_{K-1}.

- We could add the rest of the two-way interactions, as well as three way interactions, etc; the main problem with this is that, even with two- way interactions, the interpretation becomes difficult.

Logistic Regression from case-control studies

- Consider the following Esophageal Cancer Case-Control Study
- This is a case-control study to see if esophageal cancer is associated with tobacco and/or alcohol use, controlling for the possible confounding effects of age, since cancer rates in general tend to increase with age.
- Cases in this study were 200 male esophageal cancer patients in regional hospitals; 775 controls were randomly sampled from the same regions.
- After being selected in the study, the subjects were then questioned about the consumption of alcohol and tobacco (as well as other things) in the previous 10 years.

Data Listing – Grouped data

AGE (Years)	ALCOHOL (GM/DAY)	TOBACCO (GM/DAY)	CASES	CONTROLS
25-34	0-39	0-9	0	40
25-34	0-39	10-19	0	10
25-34	0-39	20-29	0	6
25-34	0-39	30+	0	5
25-34	40-79	0-9	0	27
25-34	40-79	10-19	0	7
25-34	40-79	20-29	0	4
25-34	40-79	30+	0	7
25-34	80-119	0-9	0	2
...				
75+	120+	30+	0	0

- Suppose for now, we dichotomize smoking into 2 groups, and we are interested in the odds ratio between smoking and esophageal cancer, controlling for age. (For now, forget about alcohol).
- In this sampling scheme, the response is actually,

$$X_i = \begin{cases} \text{1 if smoked} \\ \text{0 if did not smoke} \end{cases} .$$

And the covariates are

- 1. A covariate for age,

$$z_i = \text{midpoint of age interval}$$

and

- 2. The case/control status,

$$Y_i = \begin{cases} \text{1 if case} \\ \text{0 if control} \end{cases} ,$$

- Then, you formulate the logistic regression model with X_i, the exposure (smoking), as the response,

$$\text{logit}\{P[X_i = 1|Z_i, Y_i]\} = \text{logit}(p_i) = \beta_0 + \beta_1 z_i + \beta_2 y_i$$

- In this model, the conditional OR for SMOKE/NOT SMOKED versus DISEASE/NOT DISEASE, given age (z_i), is

$$\exp(\beta_2),$$

- Suppose we throw the data in a logistic regression program and get the estimates out,

$$\exp(\widehat{\beta}_2),$$

- The estimate of this conditional odds ratio is OK (asymptotical unbiased for the true ODDS RATIO), even though we have case-control data.
- This is an extension of what we discussed for a simple (2×2) table, i.e., that the ratio of the odds for X given Y equals the ratio of the odds for Y given X.
- That is,

$$OR_{\text{Prospective}} = OR_{\text{Retrospective}}$$

- Note, β_1, the conditional log-odds ratio between age (z_i) and smoking (x_i) given disease (y_i), is probably not of interest.

Plug and Chug

- Now, since we actually think of Y_i (case/control status) as the response, even though the study was case-control, suppose we mistakenly ran the logistic regression model with disease as the response, and the exposure as the covariate:

$$\text{logit}\{P^*[Y_i = 1|Z_i, X_i]\} = \alpha_0 + \alpha_1 z_i + \alpha_2 x_i$$

- I have put (P^*) because X_i was actually sampled given Y_i, (which is opposite of the usual way).
- Thus, we do not have a random sample from

$$(Y_i|x_i, z_i),$$

  i.e., P^* is different from the usual probability.

Question of interest

Suppose the true relationship between exposure and disease is given by logistic model

$$\text{logit}\{P[Y_i = 1|Z_i, X_i]\} = \beta_0 + \beta_1 z_i + \beta_2 x_i,$$

Do the α_k's in the model

$$\text{logit}\{P^*[Y_i = 1|Z_i, X_i]\} = \alpha_0 + \alpha_1 z_i + \alpha_2 x_i$$

equal the β_k's in the above true model?

- Even though, in the true underlying distribution, the disease given exposure is generated by the model with the β_k's,

$$\text{logit}\{P[Y_i = 1|Z_i, X_i]\} = \beta_0 + \beta_1 z_i + \beta_2 x_i,$$

  we have collected the dataset via a **selected sampling** mechanism, which does not correspond to the way the data are collected in the **population**.

- 1. We first collect cases and controls.
- 2. We then ascertain their previous exposure.
- Then, we have to include another indicator random variable for the sampling mechanism:

$$\delta_i = \begin{cases} 1 \text{ if subject } i \text{ is selected to be in} \\ \quad \text{case-control sample} \\ 0 \text{ if otherwise} \end{cases},$$

- Then, probability for the observed data, given earlier,

$$P^*[Y_i = 1|Z_i, X_i]$$

  can actually be written as

$$P[Y_i = 1|\delta_i = 1, Z_i, X_i],$$

  i.e., the probability $Y_i = 1$ given (Z_i, X_i) and that the person was selected into the sample $(\delta_i = 1)$.

- You can use Baye's theorem to write

$$P[Y_i = 1|\delta_i = 1, Z_i, X_i],$$

  in terms of the parameters β_1 and β_2.

Logit for selected sampling

- In particular, using Baye's rule, it can be shown that

$$\text{logit}\{P[Y_i = 1|\delta_i = 1, Z_i, X_i]\} = \beta_0^* + \beta_1 z_i + \beta_2 x_i,$$

where β_1 and β_2 are the correct parameters, and

$$\beta_0^* = \beta_0 + \log\left(\frac{\pi_0}{\pi_1}\right),$$

and

$$\begin{aligned} \pi_0 &= P[\delta_i = 1|Y_i = 0] \\ &= P[\text{control selected from population of controls}] \end{aligned}$$

and

$$\begin{aligned} \pi_1 &= P[\delta_i = 1|Y_i = 1] \\ &= P[\text{case selected from population of cases}] \end{aligned}$$

- Thus, we can make case/control (Y_i) status the response, and still get asymptotically unbiased estimates of everything (the OR's) but the intercept (which is usually of little interest anyway).

General Result

- This is a general result which holds for case-control studies with any number of exposures, even when the exposures are continuous.
- Suppose, in the population, the probability of disease given the covariates (exposures) is

$$p_i = P[Y_i = 1 | x_{i1}, ..., x_{iK}]$$

and follows the logistic regression model

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

- However, we do case control sampling with

$$p_i^* = P[Y_i = 1 | x_{i1}, ..., x_{iK}, \delta_i = 1]$$

- Then, p_i^* follows the logistic regression model

$$\text{logit}(p_i^*) = \beta_0^* + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

where β_k, $k > 0$, are the correct parameters, and

$$\beta_0^* = \beta_0 + \log\left(\frac{\pi_0}{\pi_1}\right),$$

- Later in the next lecture, we will look at models for the esophageal cancer data, using CASE/CONTROL status as the bernoulli response, and the covariates TOBACCO, ALCOHOL, and AGE; looking at these covariates as both discrete and continuous.
- We will also look at additional model building technics to assess fit of “nested models”.
- But prior to that, lets consider the proof of our general result.

PROOF

Using Baye's Law –

$$P[Y_i = 1|\delta_i = 1, Z_i, X_i] = \frac{P[\delta_i=1|Y_i=1,Z_i,X_i]P[Y_i=1|Z_i,X_i]}{P[\delta_i=1|Y_i=0,Z_i,X_i]P[Y_i=0|Z_i,X_i]+P[\delta_i=1|Y_i=1,Z_i,X_i]P[Y_i=1|Z_i,X_i]}$$

$$= \frac{P[\delta_i=1|Y_i=1]P[Y_i=1|Z_i,X_i]}{P[\delta_i=1|Y_i=0]P[Y_i=0|Z_i,X_i]+P[\delta_i=1|Y_i=1]P[Y_i=1|Z_i,X_i]}$$

$$= \frac{\pi_1 P[Y_i=1|Z_i,X_i]}{\pi_0 P[Y_i=0|Z_i,X_i]+\pi_1 P[Y_i=1|Z_i,X_i]}$$

$$= \frac{(\pi_1/\pi_0)P[Y_i=1|Z_i,X_i]}{P[Y_i=0|Z_i,X_i]+(\pi_1/\pi_0)P[Y_i=1|Z_i,X_i]}$$

Proof continued

$$
\begin{aligned}
&= \frac{(\pi_1/\pi_0)\frac{e^{\beta_0+\beta_1 z_i+\beta_2 x_i}}{1+e^{\beta_0+\beta_1 z_i+\beta_2 x_i}}}{\frac{1}{1+e^{\beta_0+\beta_1 z_i+\beta_2 x_i}}+(\pi_1/\pi_0)\frac{e^{\beta_0+\beta_1 z_i+\beta_2 x_i}}{1+e^{\beta_0+\beta_1 z_i+\beta_2 x_i}}} \\
&= \frac{(\pi_1/\pi_0)e^{\beta_0+\beta_1 z_i+\beta_2 x_i}}{1+(\pi_1/\pi_0)e^{\beta_0+\beta_1 z_i+\beta_2 x_i}} \\
&= \frac{e^{\beta_0^*+\beta_1 z_i+\beta_2 x_i}}{1+e^{\beta_0^*+\beta_1 z_i+\beta_2 x_i}},
\end{aligned}
$$

where

$$\beta_0^* = \beta_0 + \log\left(\frac{\pi_0}{\pi_1}\right),$$

Notes about the proof

- When doing the proof, you must make the assumption that selection into the sample only depends on case/control status, i.e.

$$P[\delta_i = 1|Y_i = y_i, Z_i, X_i] = P[\delta_i = 1|Y_i = y_i],$$

- If this probability also depends on Z_i, X_i, then switching disease and exposure in logistic regression will not give consistent estimates.