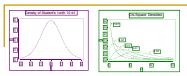
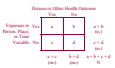
# Statistics and Epidemiology II
# Wednesday PM, June 1, 2011

Deborah Rosenberg, PhD
Research Associate Professor
Division of Epidemiology and Biostatistics
University of IL at Chicago School of Public Health

Training Course in MCH Epidemiology



## The Epidemiologic Framework

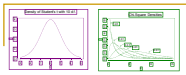


Recall that sample means, proportions, and rates, and differences or ratios of these are mathematically unbiased, but this is not a guarantee of accuracy or meaning.

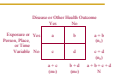
Epidemiology goes beyond the statistical properties of estimates, addressing study design, sampling and data collection strategies, data organization, and data analysis so that conclusions and decision-making are based on accurate and meaningful information.

Which statistics are reported and how they are reported flow from an epidemiologic perspective.





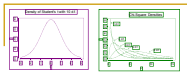
## The Epidemiologic Framework



### Laying the Groundwork

- What do we want to know? What are our hypotheses?
- What indicators should we examine—which risk markers, risk factors, and outcomes?
- Are there relevant existing data sources? Will new data be collected?
- What is the study design? Will we be able to report incidence? prevalence? relative risks? odds ratios?
- What are potential study biases?
- How will the data be organized? What statistical approaches will be used?





## The Epidemiologic Framework

- What is the study design? Will we be able to report incidence? prevalence? relative risks? odds ratios?

  The study design, and the sampling strategy in particular will have an impact on the kinds of statistical analysis that can be carried out:

  - Which measures of occurrence can be reported?
  - Which measures of association can be reported?
  - How will standard errors for confidence intervals and statistical testing be calculated?

## The Epidemiologic Framework

### Common Study Designs

Cohort Design I:
Sampling from a Disease-Free Population
Prior to Knowing Exposure Status

| RF / RM | Disease: Y | Disease: N | |
|---|---|---|---|
| Y | ? | ? | ? |
| N | ? | ? | ? |
| | ? | ? | N disease-free individuals |

Cohort Design II:
Sampling from a Disease-Free Population
According to Exposure Status ($n_1$ and $n_2$)

| RF / RM | Disease: Y | Disease: N | |
|---|---|---|---|
| Y | ? | ? | $n_1$ |
| N | ? | ? | $n_2$ |
| | ? | ? | N disease-free individuals |

**What statistics can be reported?**



## The Epidemiologic Framework

### Common Study Designs

The Case Control Design:
Sampling from $m_1$ and $m_2$

| RF / RM | Disease: Y | Disease: N | |
|---|---|---|---|
| Y | a | b | $n_1$ |
| N | c | d | $n_2$ |
| | $m_1$ | $m_2$ | N individuals |

Cross-sectional Design:
Sampling from the Entire Population

| RF / RM | Disease: Y | Disease: N | |
|---|---|---|---|
| Y | a | b | $n_1$ |
| N | c | d | $n_2$ |
| | $m_1$ | $m_2$ | N individuals |

**What statistics can be reported?**



## The Epidemiologic Framework

**What is the study design when using birth certificate data?**

- Study subjects are not sampled—all live births included
- Data on some risk factors are collected prior to the occurrence of outcomes (from prenatal record);
- Data on some risk factors are collected subsequent to the occurrence of the outcome (self-reported at delivery);



## The Epidemiologic Framework

**What is the study design when the aim is to examine the association between risk factors and elevated blood lead levels?**

- All children with high blood lead levels are included in the study; a sample of those with low blood lead levels is included
- Data on all risk factors are collected subsequent to the occurrence of the outcome (blood test);

## The Epidemiologic Framework

- What are potential study biases?

**Selection Bias:** Either prior to the beginning of the study or during the process of accumulating and retaining study participants—***who*** is studied

**Information Bias:** After study participants are already selected—***how*** information is gathered

**Confounding:** Not really a "bias"—another factor is related to both the outcome and risk factor of interest. The relationship is not due to faulty study design or faulty data collection procedures, but rather to existing relationships in the population.



## Study Bias

### Some examples of selection bias

- a sample of children enrolled in pre-school programs may include mostly children in higher income families
- women who seek early prenatal care may be ***either*** the ***highest risk*** or ***the lowest risk*** women
- responders to a sample survey may be ***different*** than non-responders

In all of these examples,
the assumption of a random sample is violated



## Study Bias

### Some examples of information bias

- Recall bias—
  - parents of children disgnosed with asthma may be more likely than parents of children without this diagnosis to remember and/or report exposures or risk factors
  - women whose infants die may be less likely than women whose infants doe not die to report behaviors such as substance use



## Study Bias

- Misclassification
  - Non-differential: the extent of misclassification of children as obese or not obese (clinical assessment) is the same ***regardless*** of reported levels of physical activity; the extent of misclassification of children's physical activity levels is the same ***regardless*** of whether they are obese or not obese.
  - Differential: obese children are more likely than children who are not obese (clinical assessment) to be classified as not being phycially active; children who are not physically active are more likely than those who are physically active to be classified as obese.

## Confounding and Effect Modification

- **Confounding:** the ***strength*** of association between a risk factor and outcome is different after accounting for another factor, but this 'new' strength of association is the same at all levels of the other factor
  - the other factor is itself independently associated with the risk factor and also with the outcome
- **Effect Modification:** the ***strength*** of association between a risk factor and outcome is ***not*** the same at each level of the other factor



## Confounding and Effect Modification

The association between a risk factor and an outcome is the same after adjusting for another factor

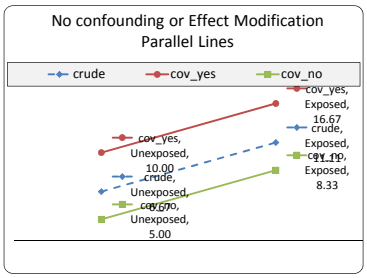
Prevalences and Relative Prevalences

$$\text{Covariate 'yes'} = \frac{16.67}{10.00}$$

$$\text{Crude} = \frac{11.11}{6.67}$$

$$\text{Covariate 'no'} = \frac{8.33}{5.00}$$

$$1.67$$

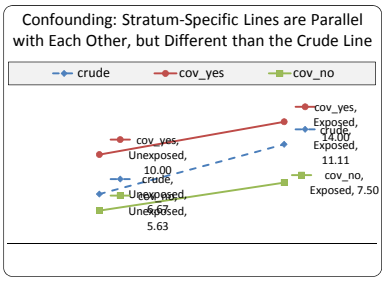





## Confounding and Effect Modification

The association between a risk factor and an outcome changes after adjusting for another factor, but is the same across strata.

Slightly diminished association in both strata after adjustment.

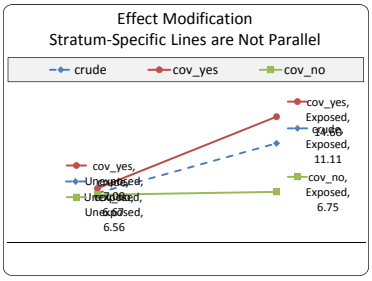





## Confounding and Effect Modification

The association between a risk factor and an outcome varies across strata

The association is much stronger in one stratum than it is in the other.



## Confounding and Effect Modification

1. Randomization
2. Restriction
3. Matching
4. Direct and Indirect Standardization
5. Stratified Analysis
6. Regression Modeling

Approaches 1,2, and 3 are part of the study design, before data analysis—confounding is assumed without regard to effect modification.

Approaches 4, 5, and 6 are used after data are collected, during data analysis—confounding and effect modification can be directly assessed.



## Confounding and Effect Modification

Assessing Confounding

If the adjusted estimate of association differs from the crude estimate of association, then confounding is present.

Determining whether a difference between the crude and adjusted measures is meaningful is a matter of judgment, since there is *no formal statistical test for the presence of confounding*.

By convention, epidemiologists consider confounding to be present if the adjusted measure of association differs from the crude measure by >= 10%



## Confounding and Effect Modification

Assessing Confounding

- Standardization: Does the standardized measure differ from the unstandardized measure?
- Stratified Analysis: Does the adjusted measure of association differ from the crude measure of association?
- Regression Analysis: Does the beta coefficient for a variable in a model that includes a potential confounder differ from the beta coefficient for that same variable in a model that does not include the potential confounder?



## Confounding and Effect Modification

Assessing Effect Modification

If stratum-specific estimates differ, then effect modification may be present and reporting a weighted average makes no sense—it actually masks the important differences that exist. Instead, it is appropriate to report the stratum-specific estimates

Stratum-specific differences can be statistically tested.

## Confounding and Effect Modification

Assessing Effect Modification

- Stratified Analysis: Are the stratum-specific measures of association different (heterogeneous)?
- Regression Analysis: Is the beta coefficient resulting from the multiplication of two variables large?



## Confounding and Effect Modification

### Typical Layout for Stratified Analysis

| | | Outcome Y | Outcome N |
|---|---|---|---|
| Risk Factor / Risk Marker | Y | | |
| | N | | |

Stratum 1: Potential Confounder = 'Y'

| | | Outcome Y | Outcome N |
|---|---|---|---|
| Risk Factor / Risk Marker | Y | | |
| | N | | |

Stratum 2: Potential Confounder='N'

| | | Outcome Y | Outcome N |
|---|---|---|---|
| Risk Factor / Risk Marker | Y | | |
| | N | | |

Break apart the crude 2 x 2 table into new tables for each level of another factor.



## Confounding and Effect Modification

Summary (adjusted) relative risk and odds ratio, and the corresponding statistical test using stratified methods:

$$\frac{\sum_{i=1}^{\#\text{ of strata}} \frac{a_i n_{2i}}{N_i}}{\sum_{i=1}^{\#\text{ of strata}} \frac{c_i n_{1i}}{N_i}} \qquad \frac{\sum_{i=1}^{\#\text{ of strata}} \frac{a_i d_i}{N_i}}{\sum_{i=1}^{\#\text{ of strata}} \frac{b_i c_i}{N_i}}$$

$$\chi^2 = \left( \frac{\sum_{i=1}^{\#\text{of strata}} a_i - \sum_{i=1}^{\#\text{of strata}} \frac{n_{1i} m_{1i}}{N_i}}{\sqrt{\sum_{i=1}^{\#\text{ of strata}} \frac{n_{1i} n_{2i} m_{1i} m_{2i}}{N_i^3}}} \right)^2$$

An adjusted measure is a **weighted average** of estimates across strata of another factor



## Confounding and Effect Modification

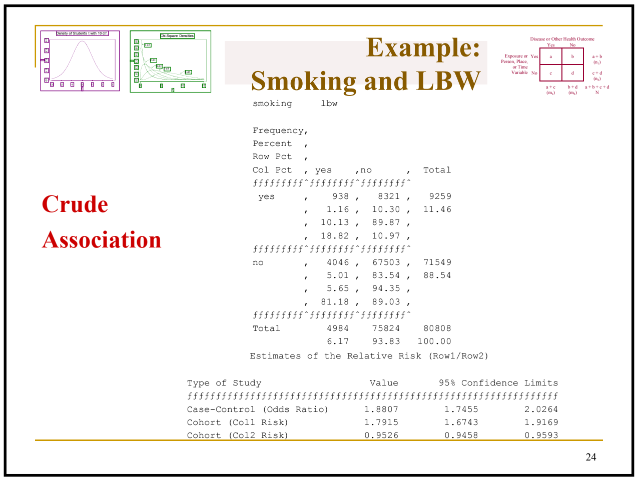
Assessing Effect Modification

The null hypothesis for assessing interaction is:

$$OR_{\text{stratum 1}} = OR_{\text{stratum 2}}$$

For stratified analysis, the statistical test is the Breslow-Day Test of Homogeneity:

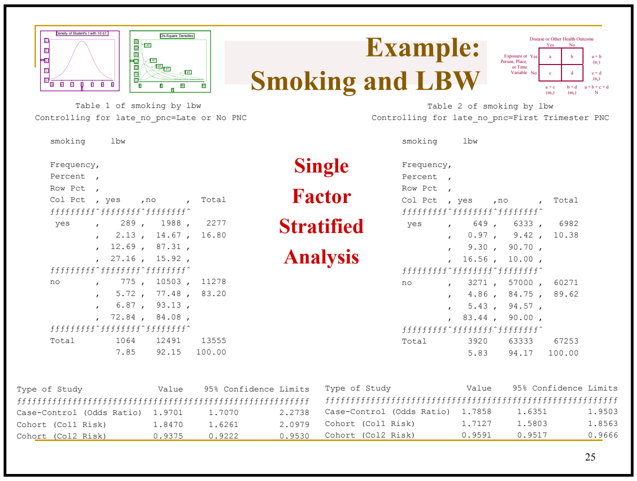
$$\chi^2_{\#\text{strata}-1} = \sum_{i=1}^{\#\text{strata}} \frac{(\text{stratum-specific measure}_i - \text{adjusted measure})^2}{\text{Variance}(\text{stratum-specific measure}_i)}$$



## Example: Smoking and LBW

**Crude Association**

```
smoking     lbw

Frequency,
Percent  ,
Row Pct  ,
Col Pct  , yes    ,no      ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 yes     ,    938 ,   8321 ,   9259
         ,   1.16 ,  10.30 ,  11.46
         ,  10.13 ,  89.87 ,
         ,  18.82 ,  10.97 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
no       ,   4046 ,  67503 ,  71549
         ,   5.01 ,  83.54 ,  88.54
         ,   5.65 ,  94.35 ,
         ,  81.18 ,  89.03 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total        4984    75824    80808
             6.17    93.83   100.00
Estimates of the Relative Risk (Row1/Row2)

Type of Study                 Value       95% Confidence Limits
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control (Odds Ratio)    1.8807       1.7455        2.0264
Cohort (Col1 Risk)           1.7915       1.6743        1.9169
Cohort (Col2 Risk)           0.9526       0.9458        0.9593
```

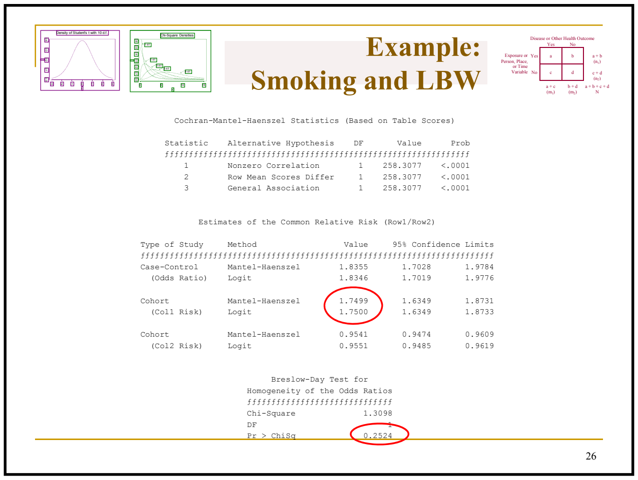


## Example: Smoking and LBW

```
        Table 1 of smoking by lbw
Controlling for late_no_pnc=Late or No PNC

   smoking     lbw

   Frequency,
   Percent  ,
   Row Pct  ,
   Col Pct  , yes    ,no      ,  Total
   ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
    yes     ,    289 ,   1988 ,   2277
            ,   2.13 ,  14.67 ,  16.80
            ,  12.69 ,  87.31 ,
            ,  27.16 ,  15.92 ,
   ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
   no       ,    775 ,  10503 ,  11278
            ,   5.72 ,  77.48 ,  83.20
            ,   6.87 ,  93.13 ,
            ,  72.84 ,  84.08 ,
   ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
   Total        1064    12491    13555
                7.85    92.15   100.00


Type of Study                Value    95% Confidence Limits
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control (Odds Ratio)  1.9701     1.7070         2.2738
Cohort (Col1 Risk)         1.8470     1.6261         2.0979
Cohort (Col2 Risk)         0.9375     0.9222         0.9530
```

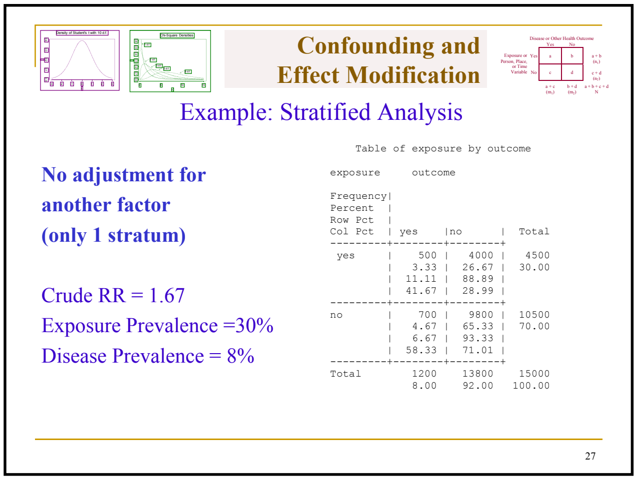
**Single Factor Stratified Analysis**

```
          Table 2 of smoking by lbw
Controlling for late_no_pnc=First Trimester PNC

   smoking     lbw

   Frequency,
   Percent  ,
   Row Pct  ,
   Col Pct  , yes    ,no      ,  Total
   ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
    yes     ,    649 ,   6333 ,   6982
            ,   0.97 ,   9.42 ,  10.38
            ,   9.30 ,  90.70 ,
            ,  16.56 ,  10.00 ,
   ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
   no       ,   3271 ,  57000 ,  60271
            ,   4.86 ,  84.75 ,  89.62
            ,   5.43 ,  94.57 ,
            ,  83.44 ,  90.00 ,
   ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
   Total        3920    63333    67253
                5.83    94.17   100.00


Type of Study                Value    95% Confidence Limits
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control (Odds Ratio)  1.7858     1.6351         1.9503
Cohort (Col1 Risk)         1.7127     1.5803         1.8563
Cohort (Col2 Risk)         0.9591     0.9517         0.9666
```



## Example: Smoking and LBW

```
     Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

  Statistic    Alternative Hypothesis    DF       Value      Prob
  ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
      1        Nonzero Correlation        1    258.3077    <.0001
      2        Row Mean Scores Differ     1    258.3077    <.0001
      3        General Association        1    258.3077    <.0001


         Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                  Value     95% Confidence Limits
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control      Mantel-Haenszel        1.8355       1.7028        1.9784
  (Odds Ratio)    Logit                  1.8346       1.7019        1.9776

Cohort            Mantel-Haenszel        1.7499       1.6349        1.8731
  (Col1 Risk)     Logit                  1.7500       1.6349        1.8733

Cohort            Mantel-Haenszel        0.9541       0.9474        0.9609
  (Col2 Risk)     Logit                  0.9551       0.9485        0.9619


                        Breslow-Day Test for
                    Homogeneity of the Odds Ratios
                    ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
                    Chi-Square              1.3098
                    DF                           1
                    Pr > ChiSq              0.2524
```



## Confounding and Effect Modification

### Example: Stratified Analysis

**No adjustment for another factor (only 1 stratum)**

Crude RR = 1.67

Exposure Prevalence =30%

Disease Prevalence = 8%

```
     Table of exposure by outcome

exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes     |no      |  Total
---------+--------+--------+
 yes     |    500 |   4000 |   4500
         |   3.33 |  26.67 |  30.00
         |  11.11 |  88.89 |
         |  41.67 |  28.99 |
---------+--------+--------+
no       |    700 |   9800 |  10500
         |   4.67 |  65.33 |  70.00
         |   6.67 |  93.33 |
         |  58.33 |  71.01 |
---------+--------+--------+
Total        1200    13800    15000
             8.00    92.00   100.00
```

## Confounding and Effect Modification

Table 1 of exposure by outcome
Controlling for covariate=yes

exposure outcome

| Frequency / Percent / Row Pct / Col Pct | yes | no | Total |
|---|---|---|---|
| yes | 250 | 1250 | 1500 |
| | 5.00 | 25.00 | 30.00 |
| | 16.67 | 83.33 | |
| | 41.67 | 28.41 | |
| no | 350 | 3150 | 3500 |
| | 7.00 | 63.00 | 70.00 |
| | 10.00 | 90.00 | |
| | 58.33 | 71.59 | |
| Total | 600 | 4400 | 5000 |
| | 12.00 | 88.00 | 100.00 |

Table 2 of exposure by outcome
Controlling for covariate=no

exposure outcome

| Frequency / Percent / Row Pct / Col Pct | yes | no | Total |
|---|---|---|---|
| yes | 250 | 2750 | 3000 |
| | 2.50 | 27.50 | 30.00 |
| | 8.33 | 91.67 | |
| | 41.67 | 29.26 | |
| no | 350 | 6650 | 7000 |
| | 3.50 | 66.50 | 70.00 |
| | 5.00 | 95.00 | |
| | 58.33 | 70.74 | |
| Total | 600 | 9400 | 10000 |
| | 6.00 | 94.00 | 100.00 |

RR=1.67 **Adj. RR=1.67** RR=1.67

Exp. Prev.=0.30 Exp. Prev.=0.30

Dis. Prev.=0.12 Dis. Prev.=0.06



## Confounding and Effect Modification

Summary Statistics for exposure by outcome
Controlling for covariate

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

| Statistic | Alternative Hypothesis | DF | Value | Prob |
|---|---|---|---|---|
| 1 | Nonzero Correlation | 1 | 85.4574 | <.0001 |
| 2 | Row Mean Scores Differ | 1 | 85.4574 | <.0001 |
| 3 | General Association | 1 | 85.4574 | <.0001 |

Estimates of the Common Relative Risk (Row1/Row2)

| Type of Study | Method | Value | 95% Confidence Limits | |
|---|---|---|---|---|
| Case-Control (Odds Ratio) | Mantel-Haenszel | 1.7619 | 1.5607 | 1.9891 |
| | Logit | 1.7618 | 1.5606 | 1.9891 |
| Cohort (Coll Risk) | Mantel-Haenszel | 1.6667 | 1.4951 | 1.8579 |
| | Logit | 1.6667 | 1.4952 | 1.8578 |
| Cohort (Col2 Risk) | Mantel-Haenszel | 0.9524 | 0.9415 | 0.9634 |
| | Logit | 0.9575 | 0.9472 | 0.9680 |

Breslow-Day Test for
Homogeneity of the Odds Ratios

| | |
|---|---|
| Chi-Square | 0.1108 |
| DF | 1 |
| Pr > ChiSq | 0.7392 |



## Confounding and Effect Modification

Table 1 of exposure by outcome
Controlling for covariate=yes

exposure outcome

| Frequency / Percent / Row Pct / Col Pct | yes | no | Total |
|---|---|---|---|
| yes | 350 | 2150 | 2500 |
| | 7.00 | 43.00 | 50.00 |
| | 14.00 | 86.00 | |
| | 58.33 | 48.86 | |
| no | 250 | 2250 | 2500 |
| | 5.00 | 45.00 | 50.00 |
| | 10.00 | 90.00 | |
| | 41.67 | 51.14 | |
| Total | 600 | 4400 | 5000 |
| | 12.00 | 88.00 | 100.00 |

Table 2 of exposure by outcome
Controlling for covariate=no

exposure outcome

| Frequency / Percent / Row Pct / Col Pct | yes | no | Total |
|---|---|---|---|
| yes | 150 | 1850 | 2000 |
| | 1.50 | 18.50 | 20.00 |
| | 7.50 | 92.50 | |
| | 25.00 | 19.68 | |
| no | 450 | 7550 | 8000 |
| | 4.50 | 75.50 | 80.00 |
| | 5.63 | 94.38 | |
| | 75.00 | 80.32 | |
| Total | 600 | 9400 | 10000 |
| | 6.00 | 94.00 | 100.00 |

RR=1.4 **Adj. RR=1.37** RR=1.33

Exp. Prev.=0.50 Exp. Prev.=0.20

Dis. Prev.=0.12 Dis. Prev.=0.06



## Confounding and Effect Modification

Summary Statistics for exposure by outcome
Controlling for covariate

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

| Statistic | Alternative Hypothesis | DF | Value | Prob |
|---|---|---|---|---|
| 1 | Nonzero Correlation | 1 | 28.7931 | <.0001 |
| 2 | Row Mean Scores Differ | 1 | 28.7931 | <.0001 |
| 3 | General Association | 1 | 28.7931 | <.0001 |

Estimates of the Common Relative Risk (Row1/Row2)

| Type of Study | Method | Value | 95% Confidence Limits | |
|---|---|---|---|---|
| Case-Control (Odds Ratio) | Mantel-Haenszel | 1.4194 | 1.2487 | 1.6134 |
| | Logit | 1.4172 | 1.2465 | 1.6112 |
| Cohort (Coll Risk) | Mantel-Haenszel | 1.3721 | 1.2219 | 1.5407 |
| | Logit | 1.3714 | 1.2214 | 1.5399 |
| Cohort (Col2 Risk) | Mantel-Haenszel | 0.9696 | 0.9584 | 0.9810 |
| | Logit | 0.9726 | 0.9616 | 0.9837 |

Breslow-Day Test for
Homogeneity of the Odds Ratios

| | |
|---|---|
| Chi-Square | 0.3182 |
| DF | 1 |
| Pr > ChiSq | 0.5727 |

## Confounding and Effect Modification

Table 1 of exposure by outcome
Controlling for covariate=yes

| exposure | outcome: yes | outcome: no | Total |
|---|---|---|---|
| yes | 365 (7.30, 14.60, 67.59) | 2135 (42.70, 85.40, 47.87) | 2500 (50.00) |
| no | 175 (3.50, 7.00, 32.41) | 2325 (46.50, 93.00, 52.13) | 2500 (50.00) |
| Total | 540 (10.80) | 4460 (89.20) | 5000 (100.00) |

Table 2 of exposure by outcome
Controlling for covariate=no

| exposure | outcome: yes | outcome: no | Total |
|---|---|---|---|
| yes | 135 (1.35, 6.75, 20.45) | 1865 (18.65, 93.25, 19.97) | 2000 (20.00) |
| no | 525 (5.25, 6.56, 79.55) | 7475 (74.75, 93.44, 80.03) | 8000 (80.00) |
| Total | 660 (6.60) | 9340 (93.40) | 10000 (100.00) |

RR=2.08 **Adj. RR=1.51** RR=1.03

Exp. Prev.=0.50 Exp. Prev.=0.20

Dis. Prev.=0.108 Dis. Prev.=0.066



## Confounding and Effect Modification

Summary Statistics for exposur by outcome
Controlling for covariate

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

| Statistic | Alternative Hypothesis | DF | Value | Prob |
|---|---|---|---|---|
| 1 | Nonzero Correlation | 1 | 43.8370 | <.0001 |
| 2 | Row Mean Scores Differ | 1 | 43.8370 | <.0001 |
| 3 | General Association | 1 | 43.8370 | <.0001 |

Estimates of the Common Relative Risk (Row1/Row2)

| Type of Study | Method | Value | 95% Confidence Limits | |
|---|---|---|---|---|
| Case-Control (Odds Ratio) | Mantel-Haenszel | 1.5677 | 1.3756 | 1.7865 |
| | Logit | 1.5498 | 1.3525 | 1.7759 |
| Cohort (Col1 Risk) | Mantel-Haenszel | 1.5091 | 1.3342 | 1.7068 |
| | Logit | 1.4978 | 1.3218 | 1.6973 |
| Cohort (Col2 Risk) | Mantel-Haenszel | 0.9631 | 0.9524 | 0.9740 |
| | Logit | 0.9723 | 0.9618 | 0.9829 |

Breslow-Day Test for
Homogeneity of the Odds Ratios

| | |
|---|---|
| Chi-Square | 32.6678 |
| DF | 1 |
| Pr > ChiSq | <.0001 |



## Confounding and Effect Modification

Highest Level of Parent Education and being overweight/obese
African American children 10-17 years old, stratified by gender

| Controlling for male=Female | | | |
|---|---|---|---|
| lesscol | Ovwt: Overweight /Obese | Ovwt: Normal Weight | Total |
| College Graduate | 313570 33.13 | 632859 66.87 | 946429 |
| Less than college | 383231 46.91 | 433754 53.09 | 816985 |
| Total | 696801 | 1066613 | 1763414 |

| Controlling for male=Male | | | |
|---|---|---|---|
| lesscol | Ovwt: Overweight /Obese | Ovwt: Normal Weight | Total |
| College Graduate | 408699 39.84 | 617133 60.16 | 1025832 |
| Less than college | 264553 39.75 | 400990 60.25 | 665543 |
| Total | 673252 | 1018124 | 1691375 |

Without the help of statistical results, do you think gender is an effect modifier or a confounder of this association? What would you guess is the adjusted measure of association?



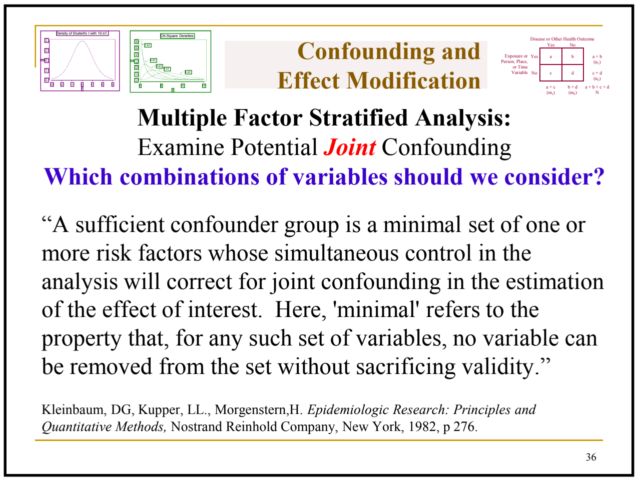
## Confounding and Effect Modification

Physical activity and being overweight/obese
African American children 10-17 years old, stratified by gender

| Controlling for male=Female | | | |
|---|---|---|---|
| phys04 | Ovwt: Overweight/ Obese | Ovwt: Normal Weight | Total |
| 0-4 days/wk | 464925 43.69 | 599203 56.31 | 1064128 |
| 5-7 days/wk | 221502 33.52 | 439273 66.48 | 660775 |
| Total | 686428 | 1038475 | 1724903 |

| Controlling for male=Male | | | |
|---|---|---|---|
| phys04 | Ovwt: Overweight /Obese | Ovwt: Normal Weight | Total |
| 0-4 days/wk | 317466 41.80 | 441970 58.20 | 759435 |
| 5-7 days/wk | 344459 37.86 | 565475 62.14 | 909935 |
| Total | 661925 | 1007445 | 1669370 |

Without the help of statistical results, do you think gender is an effect modifier or a confounder of this association? What would you guess is the adjusted measure of association?



## Confounding and Effect Modification

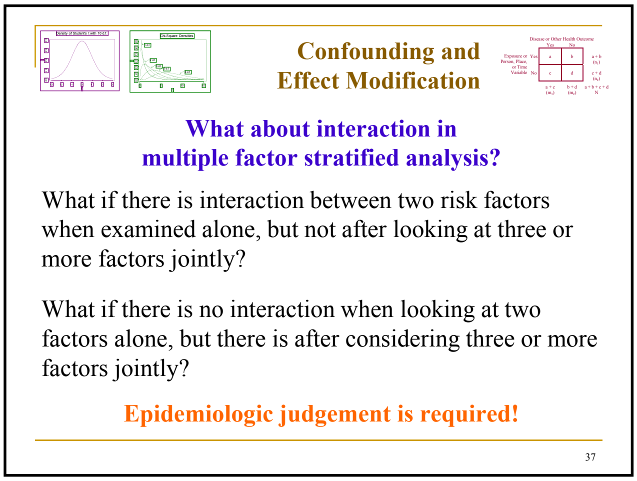
### Multiple Factor Stratified Analysis:

Examine Potential ***Joint*** Confounding

**Which combinations of variables should we consider?**

"A sufficient confounder group is a minimal set of one or more risk factors whose simultaneous control in the analysis will correct for joint confounding in the estimation of the effect of interest. Here, 'minimal' refers to the property that, for any such set of variables, no variable can be removed from the set without sacrificing validity."

Kleinbaum, DG, Kupper, LL., Morgenstern,H. *Epidemiologic Research: Principles and Quantitative Methods,* Nostrand Reinhold Company, New York, 1982, p 276.



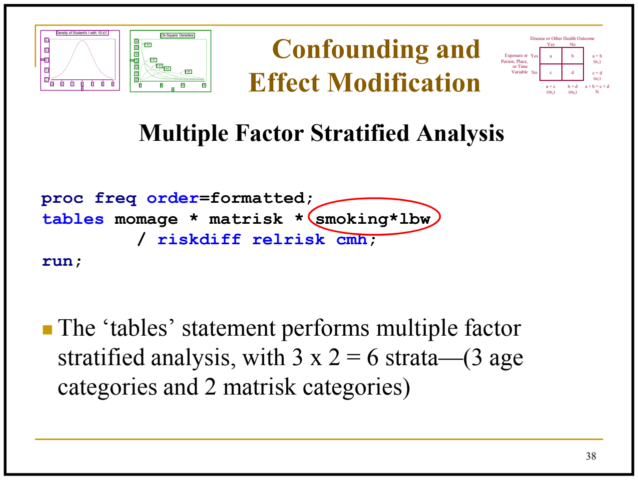
## Confounding and Effect Modification

### What about interaction in multiple factor stratified analysis?

What if there is interaction between two risk factors when examined alone, but not after looking at three or more factors jointly?

What if there is no interaction when looking at two factors alone, but there is after considering three or more factors jointly?

**Epidemiologic judgement is required!**

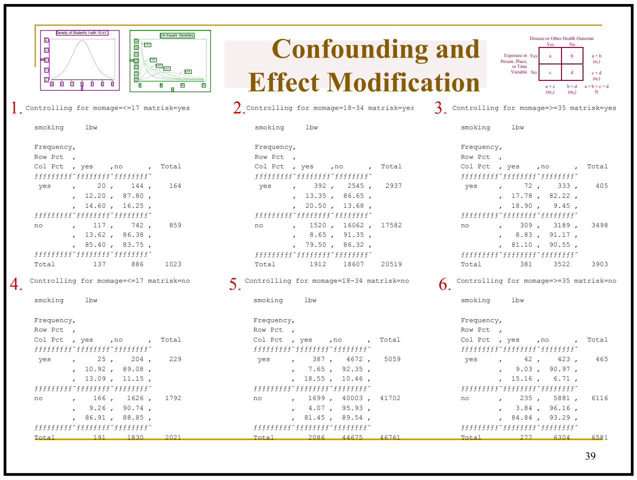


## Confounding and Effect Modification

### Multiple Factor Stratified Analysis

```
proc freq order=formatted;
tables momage * matrisk * smoking*lbw
         / riskdiff relrisk cmh;
run;
```

- The 'tables' statement performs multiple factor stratified analysis, with 3 x 2 = 6 strata—(3 age categories and 2 matrisk categories)



## Confounding and Effect Modification

1.
```
Controlling for momage=<=17 matrisk=yes

 smoking     lbw

 Frequency,
 Row Pct  ,
 Col Pct  , yes    ,no      ,  Total
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
  yes     ,     20 ,    144 ,    164
          ,  12.20 ,  87.80 ,
          ,  14.60 ,  16.25 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 no       ,    117 ,    742 ,    859
          ,  13.62 ,  86.38 ,
          ,  85.40 ,  83.75 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total         137      886     1023
```

2.
```
Controlling for momage=18-34 matrisk=yes

 smoking     lbw

 Frequency,
 Row Pct  ,
 Col Pct  , yes    ,no      ,  Total
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
  yes     ,    392 ,   2545 ,   2937
          ,  13.35 ,  86.65 ,
          ,  20.50 ,  13.68 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 no       ,   1520 ,  16062 ,  17582
          ,   8.65 ,  91.35 ,
          ,  79.50 ,  86.32 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total        1912    18607    20519
```

3.
```
Controlling for momage=>=35 matrisk=yes

 smoking     lbw

 Frequency,
 Row Pct  ,
 Col Pct  , yes    ,no      ,  Total
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
  yes     ,     72 ,    333 ,    405
          ,  17.78 ,  82.22 ,
          ,  18.90 ,   9.45 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 no       ,    309 ,   3189 ,   3498
          ,   8.83 ,  91.17 ,
          ,  81.10 ,  90.55 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total         381     3522     3903
```

4.
```
Controlling for momage=<=17 matrisk=no

 smoking     lbw

 Frequency,
 Row Pct  ,
 Col Pct  , yes    ,no      ,  Total
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
  yes     ,     25 ,    204 ,    229
          ,  10.92 ,  89.08 ,
          ,  13.09 ,  11.15 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 no       ,    166 ,   1626 ,   1792
          ,   9.26 ,  90.74 ,
          ,  86.91 ,  88.85 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total         191     1830     2021
```

5.
```
Controlling for momage=18-34 matrisk=no

 smoking     lbw

 Frequency,
 Row Pct  ,
 Col Pct  , yes    ,no      ,  Total
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
  yes     ,    387 ,   4672 ,   5059
          ,   7.65 ,  92.35 ,
          ,  18.55 ,  10.46 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 no       ,   1699 ,  40003 ,  41702
          ,   4.07 ,  95.93 ,
          ,  81.45 ,  89.54 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total        2086    44675    46761
```

6.
```
Controlling for momage=>=35 matrisk=no

 smoking     lbw

 Frequency,
 Row Pct  ,
 Col Pct  , yes    ,no      ,  Total
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
  yes     ,     42 ,    423 ,    465
          ,   9.03 ,  90.97 ,
          ,  15.16 ,   6.71 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 no       ,    235 ,   5881 ,   6116
          ,   3.84 ,  96.16 ,
          ,  84.84 ,  93.29 ,
 ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total         277     6304     6581
```

## Confounding and Effect Modification

### Multiple Factor Stratified Analysis

Estimates of the Common Relative Risk (Row1/Row2)

| Type of Study | Method | Value | 95% Confidence Limits | |
|---|---|---|---|---|
| Case-Control (Odds Ratio) | Mantel-Haenszel | 1.7759 | 1.6473 | 1.9144 |
| | Logit | 1.7961 | 1.6663 | 1.9361 |
| Cohort (Col1 Risk) | Mantel-Haenszel | 1.6889 | 1.5791 | 1.8064 |
| | Logit | 1.7047 | 1.5936 | 1.8234 |
| Cohort (Col2 Risk) | Mantel-Haenszel | 0.9562 | 0.9495 | 0.9630 |
| | Logit | 0.9585 | 0.9521 | 0.9650 |

Is there confounding or effect modification?

Breslow-Day Test for Homogeneity of the Odds Ratios

| | |
|---|---|
| Chi-Square | 21.5276 |
| DF | 5 |
| Pr > ChiSq | 0.0006 |



## Confounding and Effect Modification

| Strata: Maternal Age and Maternal Risk Status | Stratum-Specific Relative Prevalences for Smoking and LBW | 95% Confidence Interval |
|---|---|---|
| 1. <=17, high risk | 0.90 | 0.57-1.40 |
| 2. 18-34, high risk | 1.54 | 1.39-1.71 |
| 3. >=35, high risk | 2.01 | 1.59-2.55 |
| 4. <=17, low risk | 1.18 | 0.79-1.75 |
| 5. 18-34, low risk | 1.88 | 1.69-2.09 |
| 6. >=35, low risk | 2.35 | 1.72-3.22 |



## Confounding and Effect Modification

Summary

| Confounding | Effect Modification |
|---|---|
| Compare crude v. adjusted OR/RR | Compare stratum-specific OR/RR |
| No statistical testing | Statistical testing |

For confounding, the association between a risk factor and a health outcome is the same (or close to the same) in each stratum.

For effect modification, the association between a risk factor and a health outcome ***varies*** from stratum to stratum.



## Stratified Analysis—Summary

Issues to think about when planning stratified analysis:

- Defining categorical variables on data collection instruments
- Categorization schemes for continuous variables
- Misclassification of exposures or confounders
- Sample size / sparse data
- Missing values
- 'Overcontrolling'
- Software / coding issues

# Regression Modeling Overview

## Linear Models: General Considerations

The utility of regression models is their ability to assess the effect of many independent variables simultaneously, better mirroring the complexity of the real world.

- For **means,** regression analysis is an alternative to and extension of t-tests and F tests from classical analysis of variance (ANOVA).
- For **proportions or rates, regression** analysis is an alternative to and extension of chi-square tests from contingency tables – crude and stratified analysis.



## Linear Models: General Considerations

**For proportions/rates, why not just do stratified analysis?**

**Like stratified analysis, regression modeling:**

- allows examination of multiple factors (independent variables) simultaneously in relation to an outcome (dependent variable)
- provides a way to assess and test effect modification and to assess and control confounding.



## Linear Models: General Considerations

**Unlike stratified analysis, regression modeling:**

- handles more variables more efficiently
- accommodates both continuous and discrete variables, both as outcomes and as independent variables
- No need to identify a primary independent variable and a stratification variable—variables are mutually adjusted and if interaction is present, stratum-specific measures can be defined in whichever way is most informative

## Linear Models: General Considerations

### The Purpose of Modeling

Sometimes, regression modeling is carried out in order to assess **one association;** other variables are included to adjust for confounding or account for effect modification. In this scenario, the focus is on obtaining the 'best' estimate of the single association.

Sometimes, regression modeling is carried out in order to assess **multiple, competing exposures**, or to identify a **set of variables** that together predict the outcome.



## Linear Models: General Considerations

The regression models most commonly used to analyze health data express a hypothesized association between risk or other factors and an outcome as a linear (straight line) relationship.

A linear model has the advantage of interpretability—for each unit change in the value of an independent variable, there is a unit change in the value of the outcome.

When an independent variable is ordinal or continuous, the test of the beta coefficient is a test of linear trend.



## Linear Models: General Considerations

The usual linear equation is:

$$\text{Outcome}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

**Dependent = -----Independent Variables----- Error Variable**

This equation is relevant to any ***linear*** model; **what differentiates one modeling approach from another is**

- ***the structure of the outcome variable, and***
- ***the corresponding structure of the error terms***

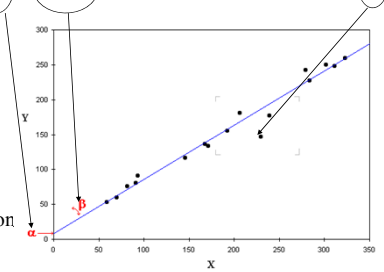


## Linear Models: General Considerations

$$\text{Outcome}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

The straight line relationship includes an intercept and one or more slope parameters.

The differences between the actual data and the regression line are the errors.



## Linear Models: General Considerations

The Traditional, 'Normal' Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

This model has the following properties:

- The outcome "Y" is continuous & normally distributed.
- The Y values are independent.
- The errors are independent, normally distributed; their sum equals 0, with constant variance across levels of X.
- The expected value (mean) of the Y's is linearly related to X (a straight line relationship exists).



## Linear Models: General Considerations

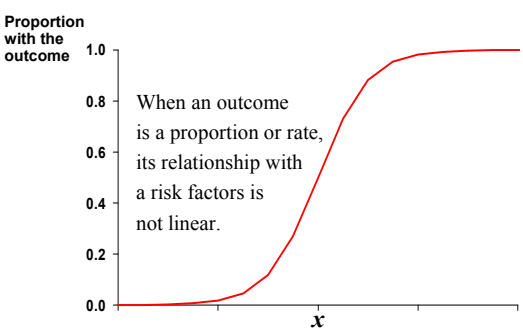
When the outcome variable is **not** continuous and normally distributed, a linear model cannot be written in the same way, and the properties listed above no longer hold.

For example, if the outcome variable is a proportion or rate:

- The errors are **not** normally distributed
- The variance across levels of X is **not** constant. (By definition, **p(1-p)** changes with **p**, and **r** changes with **r**).
- The expected value (proportion or rate) is **not** linearly related to X (**a straight line relationship does not exist**).



## Linear Models: General Considerations






## Linear Models: General Considerations

***Is there a way to use a linear modeling approach with the many health outcomes that are proportions or rates?***

YES—we can define a **"link function"** to **transform** an outcome variable from any of these distributions so that it is linearly related to a set of independent variables; the error terms can also be defined to correspond to the form of the outcome variable.

This is possible given that the normal, binomial, Poisson, exponential, chi-square, F, and multinomial distributions are all in the **exponential family**.

## Linear Models: General Considerations

### General Linear Models

**Some common link functions:**

- identity (untransformed)
- natural log (natural log of proportions / rates)
- logit (natural log of odds)
- cumulative logit
- generalized logit

The interpretation of the parameter estimates—the beta coefficients—changes depending on whether and how the outcome variable has been transformed (which link function has been used).



## Linear Models: General Considerations

'Normal' Regression—Link=Identity, Dist=Normal

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Logistic Regression—Link=Logit, Dist=Binomial

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Log Binomial or Poisson Regression—
Link=Log, Dist=Binomial or Dist=Poisson

$$\ln(\pi) \text{ or } \ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$



## Linear Models: General Considerations

### Regression Modeling Results

Measures of Occurrence
Predicted Values, Crude or Adjusted

- A 'normal' model yields means
- A Logistic model yields ln(odds)
- A binomial / Poisson model yields ln(proportions / rates)

Predicted values are the ***points on the regression line*** given particular values of the set of independent variables



## Linear Models: General Considerations

### Regression Modeling Results

Measures of Association
Beta Coefficients, Crude or Adjusted
(Difference Measures)

$$CI = b_1 \times \text{diff} \pm z_{1-\alpha/2} \times \text{s.e.}(b_1)$$

$$\text{Test Statistic} = \frac{\text{Observed Beta} - \text{Expected Beta}}{\text{Standard Error(Observed Beta)}}$$

Measures of association are ***comparisons of points*** on the regression line at different values of the independent variables

The purpose of the modeling dictates the results we report.

# Common Linear Regression Models

"Normal" and Logistic Regression

## 'Normal' Regression

Predicted Values (Means):

$$\hat{Y}_{x=\text{Value A}} = b_0 + (b_1 \times \text{Value A})$$

$$\hat{Y}_{x=\text{Value B}} = b_0 + (b_1 \times \text{Value B})$$

Predicted values use the entire regression equation, including the intercept.

Measures of Association (Differences Between Means):

When comparing two predicted values —a measure of association— the intercept terms cancel out.

$$\hat{Y}_{x=\text{Value A}} - \hat{Y}_{x=\text{Value B}}$$
$$= b_0 + (b_1 \times \text{Value A}) - b_0 + (b_1 \times \text{Value B})$$
$$= b_1 \times (\text{Value A} - \text{Value B})$$



## Logistic Regression

### Predicted Values

When the outcome is a proportion with a logistic transformation, the predicted values are log odds

$$\ln\left(\frac{p_{x=\text{Value A}}}{1-p_{x=\text{Value A}}}\right) = b_0 + (b_1 \times \text{Value A})$$

$$\ln\left(\frac{p_{x=\text{Value B}}}{1-p_{x=\text{Value B}}}\right) = b_0 + (b_1 \times \text{Value B})$$



## Logistic Regression

### Measures of Association
### Differences Between Log Odds, and the Odds Ratio

The beta coefficient is the change in the log odds for every unit change in X.

$$\ln\left(\frac{p_{x=\text{Value A}}}{1-p_{x=\text{Value A}}}\right) - \ln\left(\frac{p_{x=\text{Value B}}}{1-p_{x=\text{Value B}}}\right)$$
$$= b_0 + (b_1 \times \text{Value A}) - b_0 + (b_1 \times \text{Value B})$$
$$= b_1 \times (\text{Value A} - \text{Value B})$$

$$e^{[b_0 + (b_1 \times \text{Value A}) - b_0 + (b_1 \times \text{Value B})]}$$
$$= \frac{e^{b_0 + (b_1 \times \text{Value A})}}{e^{b_0 + (b_1 \times \text{Value B})}}$$
$$= e^{b_1 \times (\text{Value A} - \text{Value B})}$$

## 'Normal' Regression with OLS in SAS

### Well Child Visits and Insurance Type

```
TTEST PROCEDURE

Variable: CVISITS

MEDICAID        N            Mean         Std Dev       Std Error
-----------------------------------------------------------------
 private       34      4.97058824      3.17647884      0.54476163
medicaid       26      7.00000000      3.35857112      0.65866999

Variances         T        DF     Prob>|T|
-----------------------------------------
Unequal     -2.3743      52.3       0.0213
Equal       -2.3923      58.0       0.0200

For H0: Variances are equal, F' = 1.12    DF = (25,33)
                             Prob>F' = 0.7545
```



## 'Normal' Regression with OLS in SAS

```
proc reg data=one;
   model cvisits = medicaid;
   run;
```

"cvisits" = Number of well child visits

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 60.67941 | 60.67941 | 5.72 | 0.0200 |
| Error | 58 | 614.97059 | 10.60294 | | |
| Corrected Total | 59 | 675.65000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 3.25622 | R-Square | 0.0898 |
| Dependent Mean | 5.85000 | Adj R-Sq | 0.0741 |
| Coeff Var | 55.66181 | | |

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 7.00000 | 0.63860 | 10.96 | <.0001 |
| medicaid | private v. medicaid | 1 | -2.02941 | 0.84833 | -2.39 | 0.0200 |



## Logistic Regression

### Well Child Visits

```
Table of medicaid by cviscat

Frequency,
Percent  ,
Row Pct  ,
Col Pct  , < 5 vis,>= 5 vis,  Total
         ,its     ,its     ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 private ,     15 ,     19 ,     34
         ,  25.00 ,  31.67 ,  56.67
         ,  44.12 ,  55.88 ,
         ,  68.18 ,  50.00 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
medicaid ,      7 ,     19 ,     26
         ,  11.67 ,  31.67 ,  43.33
         ,  26.92 ,  73.08 ,
         ,  31.82 ,  50.00 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total          22       38       60
            36.67    63.33   100.00
```

| | | | | |
|---|---|---|---|---|
| Case-Control | Mantel-Haenszel | 2.1429 | 0.7135 | 6.4353 |



## Logistic Regression

```
proc logistic;
   model cviscat = medicaid;
   run;
```

"cviscat": 1= <5 visits
0= 5+ visits

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Standard Estimate | Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -1.7602 | 0.9493 | 3.4381 | 0.0637 |
| medicaid | 1 | 0.7619 | 0.5610 | 1.8443 | 0.1745 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| private v. medicaid | 2.142 | 0.713 | 6.434 |

## 'Normal' Regression with OLS in SAS

```
proc reg data=one;
   model cvisits = medicaid inadpnc;
   run;
```

"cvisits" = Number of well child visits

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 100.38021 | 50.19010 | 4.97 | 0.0102 |
| Error | 57 | 575.26979 | 10.09245 | | |
| Corrected Total | 59 | 675.65000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 3.17686 | R-Square | 0.1486 |
| Dependent Mean | 5.85000 | Adj R-Sq | 0.1187 |
| Coeff Var | 54.30533 | | |

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 7.35269 | 0.64791 | 11.35 | <.0001 |
| medicaid | private v. medicaid | 1 | -1.73481 | 0.84088 | -2.06 | 0.0437 |
| inadpnc | | 1 | -1.83400 | 0.92470 | -1.98 | 0.0522 |



## Logistic Regression

```
Table 1 of medicaid by cviscat
Controlling for inadpnc=yes

medicaid     cviscat

Frequency,
Percent  ,
Row Pct  ,
Col Pct  , < 5 vis,>= 5 vis,  Total
         ,its    ,its    ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
 private ,      7 ,      5 ,     12
         ,  41.18 ,  29.41 ,  70.59
         ,  58.33 ,  41.67 ,
         ,  77.78 ,  62.50 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
medicaid ,      2 ,      3 ,      5
         ,  11.76 ,  17.65 ,  29.41
         ,  40.00 ,  60.00 ,
         ,  22.22 ,  37.50 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total          9        8       17
            52.94    47.06   100.00


Odds Ratio   2.1000   0.2507  17.5941

             Case-Control      Mantel-Haenszel
```

```
Table 2 of medicaid by cviscat
Controlling for inadpnc=no

medicaid     cviscat

Frequency,
Percent  ,
Row Pct  ,
Col Pct  , < 5 vis,>= 5 vis,  Total
         ,its    ,its    ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
 private ,      8 ,     14 ,     22
         ,  18.60 ,  32.56 ,  51.16
         ,  36.36 ,  63.64 ,
         ,  61.54 ,  46.67 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
medicaid ,      5 ,     16 ,     21
         ,  11.63 ,  37.21 ,  48.84
         ,  23.81 ,  76.19 ,
         ,  38.46 ,  53.33 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total         13       30       43
            30.23    69.77   100.00


Odds Ratio  1.8286   0.4847  6.8985

            1.9006   0.6164   5.8606
```



## Logistic Regression

```
proc logistic;
   model cviscat = medicaid inadpnc;
   run;
```

"cviscat": 1= <5 visits
0= 5+ visits

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -1.8282 | 0.9659 | 3.5822 | 0.0584 |
| medicaid | 1 | 0.6426 | 0.5747 | 1.2502 | 0.2635 |
| inadpnc | 1 | 0.8504 | 0.5996 | 2.0117 | 0.1561 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| private v. medicaid | 1.901 | 0.616 | 5.865 |
| inadpnc | 2.341 | 0.723 | 7.580 |



## 'Normal' Regression with OLS in SAS

```
proc reg data=one;
   model dbirwt = smoking;
   run;
```

"dbirwt" = Birthweight (continuous) from vital records

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 317799174 | 317799174 | 981.24 | <.0001 |
| Error | 80806 | 26171015889 | 323875 | | |
| Corrected Total | 80807 | 26488815063 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 569.09987 | R-Square | 0.0120 |
| Dependent Mean | 3330.17903 | Adj R-Sq | 0.0120 |
| Coeff Var | 17.08917 | | |

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 3352.73853 | 2.12758 | 1575.84 | <.0001 |
| smoking | | 1 | -196.88822 | 6.28538 | -31.32 | <.0001 |

## Logistic Regression

### Low Birthweight

Table of smoking by lbw

```
Frequency,
Percent  ,
Row Pct  , yes    ,no      ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 yes     ,    938 ,    8321 ,   9259
         ,   1.16 ,   10.30 ,  11.46
         ,  10.13 ,   89.87 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
no       ,   4046 ,   67503 ,  71549
         ,   5.01 ,   83.54 ,  88.54
         ,   5.65 ,   94.35 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total        4984    75824    80808
             6.17    93.83   100.00
```

| | | | | |
|---|---|---|---|---|
| Case-Control | Mantel-Haenszel | 1.8807 | 1.7455 | 2.0264 |



## Logistic Regression

**Output from proc logistic:**

Response Profile

| Ordered Value | lbw | Total Frequency |
|---|---|---|
| 1 | yes | 4984 |
| 2 | no | 75824 |

Probability modeled is lbw='yes'.

| | | |
|---|---|---|
| -2 Log L | 37423.364 | 37177.164 |

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.8144 | 0.0162 | 30236.4216 | <.0001 |
| smoking | 1 | 0.6317 | 0.0381 | 275.5238 | <.0001 |

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| smoking → | 1.881 | 1.746 | 2.026 |



## 'Normal' Regression with OLS in SAS

```
proc reg data=one;
   model dbirwt = smoking late_no_pnc;
   run;
```

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 375465933 | 187732967 | 580.92 | <.0001 |
| Error | 80805 | 26113349130 | 323165 | | |
| Corrected Total | 80807 | 26488815063 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 568.47605 | R-Square | 0.0142 |
| Dependent Mean | 3330.17903 | Adj R-Sq | 0.0142 |
| Coeff Var | 17.07044 | | |

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 3364.04030 | 2.28746 | 1470.64 | <.0001 |
| smoking | | 1 | -190.55737 | 6.29636 | -30.26 | <.0001 |
| late_no_pnc | | 1 | -71.69983 | 5.36744 | -13.36 | <.0001 |



## Logistic Regression

Table 1 of smoking by lbw
Controlling for late_no_pnc=Late or No PNC

```
Frequency,
Percent  ,
Row Pct  , yes    ,no      ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 yes     ,    289 ,    1988 ,   2277
         ,   2.13 ,   14.67 ,  16.80
         ,  12.69 ,   87.31 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
no       ,    775 ,   10503 ,  11278
         ,   5.72 ,   77.48 ,  83.20
         ,   6.87 ,   93.13 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total        1064    12491   13555
             7.85    92.15  100.00
```

Odds Ratio 1.9701 1.7070-2.2738

Table 2 of smoking by lbw
Controlling for late_no_pnc=First Trimester PNC

```
Frequency,
Percent  ,
Row Pct  , yes    ,no      ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 yes     ,    649 ,    6333 ,   6982
         ,   0.97 ,    9.42 ,  10.38
         ,   9.30 ,   90.70 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
no       ,   3271 ,   57000 ,  60271
         ,   4.86 ,   84.75 ,  89.62
         ,   5.43 ,   94.57 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total        3920    63333   67253
             5.83    94.17  100.00
```

Odds Ratio 1.7858 1.6351-1.9503

Case-Control Mantel-Haenszel 1.8355 1.7028-1.9784

## Logistic Regression

**Output from proc logistic:**

```
                Response Profile

                Ordered                  Total
                  Value     lbw      Frequency

                      1     yes           4984
                      2     no           75824

          Probability modeled is lbw='yes'.

     -2 Log L      37423.364      37122.331

                                Standard          Wald
Parameter       DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept        1     -2.8622      0.0176     26390.9295        <.0001
smoking          1      0.6064      0.0382       251.5499        <.0001
late_no_pnc      1      0.2739      0.0362        57.3687        <.0001

                          Point          95% Wald
Effect                 Estimate      Confidence Limits

smoking                   1.834       1.701       1.977
late_no_pnc               1.315       1.225       1.412
```



## Logistic Regression

Computing the OR for the association between smoking and low birthweight adjusting for late/no prenatal care:

$$= \frac{e^{-2.8622+(0.6064\times 1)+(0.2739\times 0)}}{e^{-2.8622+(0.6064\times 0)+(0.2739\times 0)}} \qquad \frac{e^{-2.8622+(0.6064\times 1)+(0.2739\times 1)}}{e^{-2.8622+(0.6064\times 0)+(0.2739\times 1)}}$$

$$= e^{0.6064\times(1-0)} \qquad = e^{0.6064\times(1-0)}$$

$$= e^{0.6064} \qquad = e^{0.6064}$$

$$= 1.834 \qquad = 1.834$$

The result is the same ***regardless of*** whether we use '1' or '0' for the value of the prenatal care variable—this is the meaning of 'adjustment for confounding'.



## Modeling Effect Modification

Effect modification / interaction is typically assessed under the multiplicative model, by entering the product of the values on two variables into the model as a new variable.

```
proc logistic order=formatted;
   model outcome = v1 v2 v1*v2;
run;
```



## Modeling Effect Modification

Assuming two dichotomous variables, values of the modeled variables are:

| v1 | v2 | int_1_2 |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 0 |

## Modeling Effect Modification

Assuming one dichotomous and one continuous variable, values of the modeled variables might be:

| v1 | Age | int_1_2 |
|---|---|---|
| 1 | 22 | 22 |
| 1 | 65 | 65 |
| 0 | 53 | 0 |
| 0 | 47 | 0 |



## Modeling Effect Modification

Suppose we are assessing interaction between two dichotomous variables. The hypothesis of no multiplicative interaction is:

$$OR_{\text{stratum 1}} = OR_{\text{stratum 2}}$$



## Modeling Effect Modification

```
/*product term in model */
proc logistic order=formatted;
   model lbw = smoking late_no_pnc
               smoking*late_no_pnc;
run;
```

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.8580 | 0.0180 | 25267.2474 | <.0001 |
| smoking | 1 | 0.5799 | 0.0450 | 166.2930 | <.0001 |
| late_no_pnc | 1 | 0.2514 | 0.0413 | 36.9870 | <.0001 |
| smoking*late_no_pnc | 1 | 0.0987 | 0.0858 | 1.3219 | 0.2503 |

What are the adjusted estimates?

What are the stratum-specific estimates?



## Modeling Effect Modification

Only a Model Considering Effect Modification Yields Stratum-Specific Measures of Association

$$= \frac{e^{-2.8858+(0.5799\times1)+(0.2514\times1)+(0.0987\times1)}}{e^{-2.8858+(0.5799\times0)+(0.2514\times1)+(0.0987\times0)}}$$
$$= e^{0.5799\times(1-0)+0.0987\times(1-0)}$$
$$= e^{0.5799\times(1)+0.0987\times(1)}$$
$$= 1.971$$

--Computation for 1st stratum

The association between smoking and LBW among women with late or no PNC.

$$= \frac{e^{-2.8858+(0.5799\times1)+(0.2514\times0)+(0.0987\times0)}}{e^{-2.8858+(0.5799\times0)+(0.2514\times0)+(0.0987\times0)}}$$
$$= e^{0.5799\times(1-0)}$$
$$= e^{0.5799(1)}$$
$$= 1.7859$$

Computation for 2nd stratum—

The association between smoking and LBW among women with early PNC.

## Modeling Effect Modification

Table 1 of exposure by outcome
Controlling for covariate=yes

| exposure | outcome: yes | outcome: no | Total |
|---|---|---|---|
| yes | 365 (7.30%, Row 14.60%, Col 67.59%) | 2135 (42.70%, Row 85.40%, Col 47.87%) | 2500 (50.00%) |
| no | 175 (3.50%, Row 7.00%, Col 32.41%) | 2325 (46.50%, Row 93.00%, Col 52.13%) | 2500 (50.00%) |
| Total | 540 (10.80%) | 4460 (89.20%) | 5000 (100.00%) |

Table 2 of exposure by outcome
Controlling for covariate=no

| exposure | outcome: yes | outcome: no | Total |
|---|---|---|---|
| yes | 135 (1.35%, Row 6.75%, Col 20.45%) | 1865 (18.65%, Row 93.25%, Col 19.97%) | 2000 (20.00%) |
| no | 525 (5.25%, Row 6.56%, Col 79.55%) | 7475 (74.75%, Row 93.44%, Col 80.03%) | 8000 (80.00%) |
| Total | 660 (6.60%) | 9340 (93.40%) | 10000 (100.00%) |

RR=2.08 **Adj. RR=1.51** RR=1.03

Exp. Prev.=0.50 Exp. Prev.=0.20

Dis. Prev.=0.108 Dis. Prev.=0.066



## Modeling Effect Modification

```
proc logistic data = interaction order=formatted;
  model outcome = exposure covariate;
run;
```

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.7491 | 0.0438 | 3930.7625 | <.0001 |
| exposure | 1 | 0.4289 | 0.0651 | 43.4616 | <.0001 |
| covariate | 1 | 0.4054 | 0.0643 | 39.6934 | <.0001 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| exposure | 1.535 | 1.352 | 1.744 |
| covariate | 1.500 | 1.322 | 1.702 |

We know there is interaction, so these are inappropriate estimates even though they are statistically significant



## Modeling Effect Modification

```
proc logistic data = interaction order=formatted;
  model outcome = exposure covariate exposure*covariate;
run;
```

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.6559 | 0.0452 | 3460.2770 | <.0001 |
| exposure | 1 | 0.0302 | 0.0999 | 0.0912 | 0.7626 |
| covariate | 1 | 0.0692 | 0.0905 | 0.5857 | 0.4441 |
| exposure*covariate | 1 | 0.7903 | 0.1390 | 32.3012 | <.0001 |

**Covariate=Yes**

$$\frac{e^{-2.6559+0.0302(1)+0.0692(1)+0.7903(1)}}{e^{-2.6559+0.0302(0)+0.0692(1)+0.7903(0)}} = e^{0.0302(1)+0.7903(1)} = 2.27$$

**Covariate=No**

$$\frac{e^{-2.6559+0.0302(1)+0.0692(0)+0.7903(0)}}{e^{-2.6559+0.0302(0)+0.0692(0)+0.7903(0)}} = e^{0.0302(1)} = 1.03$$



## Modeling with Dummy Variables

What if you want to model a nominal independent variable?

For example, what if you want to model three categories of race/ethnicity—African American, Hispanic, and White?

If the outcome variable were continuous, this situation would be ANOVA, testing the hypothesis:

$$H_o: \mu_1 = \mu_2 = \mu_3$$

## Modeling with Dummy Variables

In logistic regression, the hypothesis could be written:

$$H_o: \ln(Odds_1) = \ln(Odds_2) = \ln(Odds_3)$$

In either case, we need to "trick" the modeling procedure into handling the nominal variable appropriately.



## Modeling with Dummy Variables

Race/ethnicity as a single variable coded, for example 1=African American, 2=Hispanic, and 3=White, will be treated as ordinal in a regression model.

```
proc logistic order=formatted;
  model outcome = ethnicity;
run;
```

The **incorrect** interpretation of the resulting beta coefficient for 'ethnicity' would be, "for every unit change in 'ethnicity', there is a ____ change in the log odds of the outcome".



## Modeling with Dummy Variables

So, what's the trick?

Dummy variables, or indicator variables, are a set of dichotomous variables which together capture the nominal construct of interest.

For a nominal variable with k categories, a set of k-1 dummy variables will capture the entire construct.

If variables for all k categories are created, there will be redundancy in the model.



## Modeling with Dummy Variables

Each dichotomous variable is indeed assumed to be 'ordinal' by the modeling procedure, but this will work when there are only two categories

For example, we know that 'sex' can be appropriately modeled even though it is a nominal variable.

The beta coefficient for sex is interpreted as the difference between means (OLS) or the difference between log odds (logistic) for males and females.

## Modeling with Dummy Variables

Example: Dummy variables for race/ethnicity:

So, we only create 2 variables for our 3 category race/ethnicity variable.

| | af_am | hisp |
|---|---|---|
| African American | 1 | 0 |
| Hispanic | 0 | 1 |
| White | 0 | 0 |

Here, whites are being considered the reference group.



## Modeling with Dummy Variables

Explicit coding in SAS: If we're not sure which level we want as the reference group, we can code 'k' dummies and then decide which k-1 we will model:

```
if race = 1 then af_am = 1;
  else if race ^= . then af_am = 0;
if race = 2 then hisp = 1;
  else if race ^= . then hisp = 0;
if race = 3 then white = 1;
  else if race^= . then white = 0;
```



## Modeling with Dummy Variables

Now, race/ethnicity can be modeled as follows:

```
proc logistic order=formatted;
  model outcome = af_am hisp;
run;
```

The beta coefficient for af_am is the difference in log odds between African Americans and **whites**; the beta coefficient for hisp is the difference in log odds between Hispanics and **whites**. Exponentiating the betas, we get the odds ratios for African Americans v. whites & for Hispanics v. whites.



## Modeling with Dummy Variables

With this model, there is no direct way to compare African Americans and Hispanics, but we could rerun the model with African American as the reference group:

```
proc logistic order=formatted;
  model outcome = hisp white;
run;
```

Now, exponentiating the betas, we get the odds ratios for Hispanics v. African Americans & for whites v. African Americans.

## Modeling with Dummy Variables

Suppose you have an apparently ordinal variable such as income:

- Should you include it in a model in this ordinal form?
- Should you create dummy variables?

And, if you decide to create dummy variables:

- How many dummy variables will you create?
- What is the reference group? Why?



## Modeling with Dummy Variables

Suppose income is coded:

1 = < $25,000
2 = $25,000 - $49,999
3 = $50,000 - $74,999
4 = >= $75,000

```
If income = 1 then under25k = 1;
  else if income ^= . then under25k = 0;
If income = 2 then twentyfiveto50k = 1;
  else if income ^= . then twentyfiveto50k = 0;
If income = 3 then fiftyto75k = 1;
  else if income ^= . then fiftyto75k = 0;
If income = 4 then morethan75k = 1;
  else if income ^= . then morethan75k = 0;
```



## Modeling with Dummy Variables

Possible Models for Income:

```
proc logistic order=formatted;
  model outcome = under25k  fiftyto75k  over75k;
run;
```

OR

What are the pros and cons of each approach?

```
proc logistic order=formatted;
  model outcome = twentyfiveto50k   fiftyto75k   over75k;
run;
```



## Modeling with Dummy Variables

Dummy variables can also be used to create composite variables combining multiple variables. Sometimes it makes conceptual sense to create an "index" by combining several variables.

For example, dummy variables could be used to reflect a combined variable for age and education or for several SES measures.

## Model Building Strategies

Preparation for Modeling

- Be sure you understand the coding of variables
- Think about the directionality of variables in relation to what you want to report—directionality for both outcome and independent variables
- Determine whether you will be using the entire dataset or whether you will be restricting your analysis to particular subgroups



## Model Building Strategies

Multivariable modeling should be the culmination of an analytic strategy that includes articulating a conceptual framework and carrying out preliminary analysis.

**BEFORE any multivariable modeling—**

1. Select variables of interest
2. Define categories, sometimes more than once, for a given variable
3. Examine univariate distributions
4. Examine bivariate associations



## Model Building Strategies

**BEFORE any multivariable modeling—**

5. Perform single factor stratified analysis for the primary association of interest, with each potential confounder / effect modifier
6. Rethink variables and categories
7. Perform multiple factor stratified analysis for the primary association of interest with different combinations of potential confounders / effect modifiers

**These steps should never be skipped!**



## Model Building Strategies

Use the results of stratified analysis to inform the initial model-building phase.

Continue exploring confounding and effect modification with more variables.

Decide on "rules" for inclusion of variables in a model, including components of interactions: statistical testing, conceptual rationale, etc.

## Model Building Strategies

Some approaches to organizing variables and reducing the analysis burden:

- Group according to domain:
  - Sociodemographic
  - Behavioral
  - Medical risk
  - Health care system
- Group according to evidence from previous studies
- Produce correlation matrices and identify proxy variables
- Subset analysis



## Model Building Strategies

Conceptual issues

- risk factors that only apply to some subjects, e.g. history of …
- causal pathway
- combining variables; index construction

Coding Issues

- Level of measurement
  - Dummy variable coding—how many? what's the reference group?
  - Ordinal—how refined?
  - Continuous



## Model Building Strategies

Automated Model Selection Procedures

- Stepwise
- Forward
- Backward
- Best Subsets
- Chunkwise

Can manually mimic these approaches ...



## A Note on a Few Other Linear Models

- Log binomial regression: an alternative to logistic regression for many MCH indicators for which we have prevalence or cumulative incidence data; it produces predicted prevalences and relative prevalences rather than odds and odds ratios.
- Polytomous logistic regression: makes it possible to model outcomes with more than two categories, e.g. adequacy of prenatal care, types of stressful life events, categories of gestational age at delivery, categories of body mass index.
- Multilevel modeling: simultaneously models individual level and contextual level indicators, such as school, hospital, neighborhood, county, or state level variables.