3. When do you need to stratify?
   a. Heruristically: when stratifier is a confounder
      i. That is, it is related to both exposure and disease
      ii. Empirically, the odds ratio will change if both row and column proportions differ according to stratifier.
      R Code SAS Code

A: 4.3.5

C. Varying odds ratios
1. Varying odds ratios represent interactions.
   a. If θ for the various strata are different, there is an interaction between the confounder and exposure.
   b. Use Breslow and Day statistic to test homogeneity of odds ratio in a series of I 2×2 tables:

$$\sum_{i,j,k}(X^i_{jk} - \hat{E^i_{jk}})^2/\hat{E^i_{jk}} - C \sim \chi^2_{I-1}$$

      i. $\hat{E^i_{jk}}$ satisfy
         - $\hat{E^i_{+k}} = X^i_{+k} \forall j, i$,
         - $\hat{E^i_{j+}} = X^i_{j+} \forall k, i$,
         - $(\hat{E^i_{11}}\hat{E^i_{00}})/(\hat{E^i_{10}}\hat{E^i_{01}}) = \hat{\theta} \forall i$, for $\hat{\theta}$ the Mantel Haneszel estimator.
      ii. $C = \sum_i (X^i_{00} - \hat{E^i_{00}})^2 / \sum_i (1/\hat{E^i_{00}} + 1/\hat{E^i_{10}} + 1/\hat{E^i_{01}} + 1/\hat{E^i_{11}})^{-1}$
         - Called Tarone's correction.
         - Agresti says that that generally C is small
         - SAS appears to ignore C .

         - Necessary, because Mantel Haenszel estimator does not minimize the quadratic form.
   2. Checking for confounding via hypothesis test
      a. Procedure
         i. test for association betw. C and D and betw. C and E ,
         ii. adjust if these are significant
      b. Uses significance as a proxy for strength of effect
      c. To make it work at all, typically make very loose criteria for significance
      d. Fails to control Type 1 error R Code SAS Code

A: 8-8.2

D. Matching
1. Matching is extreme case of stratification
   a. Can either be case–control pairs or exposed–unexposed pairs
   b. Exposed-Unexposed
      i. Let $n_{il} =$ number of pairs with unexposed at response level i , exposed at response level l
         - Pairs with the same response levels for exposed and unexposed are called *concordant*.
         - Pairs with different response levels for exposed and unexposed are called *discordant*.
   c. Case-Control
      i. Let $n_{il} =$ number of pairs with case at exposure level i , control at exposure level l
         - Pairs with the same exposure levels for case and control are called *concordant*.

         - Pairs with different exposure levels for case and control are called *discordant*.
   2. Assumption (exposed–unexposed pairs):
      a. Let π^i_k be the probability of event in exposure group k for pair i
      b. Assume $\pi^i_1(1-\pi^i_0)/[\pi^i_0(1-\pi^i_1)] = \theta \forall i$
   3. Use Mantel–Haenszel test
      a. For concordant pairs
         i. Expected values are exactly observed
         ii. Variance is zero
         iii. Hence contribution is zero
      b. For discordant pairs
         i. Expected is all $\frac{1}{2}$
         ii. Obsd-expected is
            - $(1 - \frac{1}{2}) = \frac{1}{2}$ for pairs with + association
            - $(0 - \frac{1}{2}) = -\frac{1}{2}$ for pairs with - association
         iii. Using hypergeometric distribution, null variance contribution for pair is $(1 \times 1 \times 1 \times 1)/(2 \times 2 \times (2-1)) = \frac{1}{4}$
            - Total variance is $\frac{1}{4}(n_{10} + n_{01})$.
      c. Test statistic is $(n_{10} - n_{01})/\sqrt{n_{10} + n_{01}}$
         i. same as test that binomial proportion equals $\frac{1}{2}$
         ii. Compare to standard normal
      d. Test is called *McNemar's test* SAS Code R Code
         i. Test where units are pairs
         ii. Each pair has two measurements
         iii. This is NOT a test of whether the two pairs agree
         SAS Code R Code

   4. What should we match on?
      a. Often match on traits that are expected to impact disease
      b. Matching is to remove effect of something associated with both putative cause and effect
      c. Matching can reduce efficiency:
         i. Matching on something correlated to exposure,

$$\begin{array}{c} E \rightarrow D \\ \downarrow \\ C \end{array}$$

            - you get pairs with similar exposure
            - that don't give much info about effect of exposure on disease
         ii. Matching on an intermediate step in causal chain,

$$E \rightarrow C \rightarrow D$$

            - make exposed more similar to non-exposed.
            - artificially deflate effect of exposure
         iii. Both are known as *over-matching*
         iv. Sometimes matched pairs are multiple observations on one individual.

A: 2.4.3

   5. Estimation for Matched pairs
      a. Pairs have probabilities

	0	1
0	$\psi_{00}\psi_{10}$	$\psi_{00}\psi_{11}$
1	$\psi_{01}\psi_{10}$	$\psi_{01}\psi_{11}$

      b. $n_{01}|n_{10}+n_{10} \sim \text{Bin}(\psi_{00}\psi_{11}/(\psi_{00}\psi_{11}+\psi_{01}\psi_{10}), n_{10}+n_{01}) = \text{Bin}(\theta/(1+\theta), n_{10}+n_{01})$ after conditioning on $n_{10}+n_{01}$.

i. $\omega = \theta/(1+\theta)$; $\theta = \omega/(1-\omega)$.

c. Hence $\hat{\theta} = n_{01}/n_{10}$

d. And get CI for θ by transforming binomial CI

e. This is also Mantel–Haenszel estimator R Code

6. Sometimes it is hard to make matched pairs,

a. because collection of subjects doesn't contain pair

b. or setting up pairs is a lot of work

c. Many models we will employ later will allow us to adjust for confounders without matching.

Se: 9 pp. 279–280

7. When matched groups are larger than 2

a. and not necessarily all the same size

b. still use Mantel-Haenszel procedure

c. exact binomial results no longer hold

d. Returns in efficiency from many control matches to a single case diminish

A: 4.3

V. Rates depending on covariates

A. Introduction

1. Previous methods in this course

a. Exposure dichotomous, or categorical with few levels

b. Simple model allowed disease rates to vary from exposure group to exposure group

2. Now

a. want covariate with more levels

b. Suppose L covariates

i. Includes constant 1

c. Identify K relatively homogeneous groups

i. ie., same (or similar) values for all covariates

3. Need some structure betw. rates at different exposure levels

a. Interpret ability

b. stability of estimates

c. We will assume linearity on log scale

B&D2: 4.3a

4. Assume that

a. numbers of events in an interval are Poisson

i. $\mathrm{P}[X_j = x_j] = \exp(-\lambda_j)\lambda_j^{x_j}/x_j!$

b. Implies that each person has chance $\exp(-\Delta\lambda_j)$ of surviving interval Δ without an event.

c. As before, assume individuals act independently.

5. Log linear model for effect of covariates

a. Suppose that z_{kl} is covariate l in group k

b. model says $\log(\lambda_k) = \sum_{l=1}^{L} z_{kl}\beta_l = \boldsymbol{z}_k\boldsymbol{\beta}$

c. Bold faced quantities are vectors

d. Multiplication in last expression is inner product.

6. Model is an example of a *generalized linear model*.

a. More specifically, *Poisson regression*

:

B. Preivious models as regressions

1. One dimension:

a. $\lambda_k = \exp(\alpha_k)$

b. $\boldsymbol{\beta} = (\alpha_0, \ldots, \alpha_{K-1})$, $\boldsymbol{z}_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, with the 1 in position k .

i. Model now has one parameter for every observation: *saturated*

c. $L(\boldsymbol{\alpha}) = \prod_{k=0}^{K-1} \exp([\omega_k+\alpha_k]X_k - \exp([\omega_k+\alpha_k]))/X_k!$

d. $l(\boldsymbol{\alpha}) = \sum_{k=0}^{K-1}[\{\alpha_k+\omega_k\}X_k - \exp(\alpha_k+\omega_k) - \log(X_k!)]$

e. $l^k(\boldsymbol{\alpha}) = X_k - \exp(\alpha_k+\omega_k)$

f. Maximizer satisfies $\hat{\alpha_k} = \log(X_k) - \omega_k$

g. For the submodel with all α's equal, $l(\alpha) = \alpha X_+ + \sum_{k=0}^{K-1}\omega_k X_k - \exp(\alpha)\sum_{k=0}^{K-1}\exp(\omega_k) - \sum_{k=0}^{K-1}\log(X_k!)$

i. $l'(\alpha) = X_+ - \exp(\alpha)\sum_{k=0}^{K-1}\exp(\omega_k)$

ii. $\hat{\alpha} = \log(X_+/\sum_{k=0}^{K-1}\exp(\omega_k))$.

iii. Profile score statistic is $l^k(\hat{\alpha}) = X_k - X_+\exp(\omega_k)/\sum_{k=0}^{K-1}\exp(\omega_k)$

h. After conditioning on X_+ ,

i. distribution is now multinomial with probabilities $\pi_k = \exp(\omega_k+\alpha_k)/\sum_{m=0}^{K-1}\exp(\omega_m+\alpha_m)$

ii. Increasing or decreasing all of the α_k by the same amount gives the same probabilities.

iii. Hence one can not identify all of the α_k .

iv. Pick one of these (ie., $\alpha_0 = 0$), or set sum to zero (PROC CATMOD)

2. Model contains log of time at risk as an *offset*

a. Fit component is added to every log rate

b. If you know something that rates might be proportional to, log of this could be added to the offset as well

i. For ex, rate in unexposed population by age SAS Code R Code

3. Complications:

a. Do iterations bounce back and forth without converging?

b. Sometimes best fits for parameters are $\pm\infty$

c. Tests can mislead when some groups have small expected value