Se: 9 pp. 285–287

i. Test where units are pairs
ii. Each pair has two measurements
iii. Note that this is NOT a test of whether the two pairs agree

e. A measure of association might be constructed by taking
i. observed proportion agreeing
ii. minus expected proportion agreeing p_e
- Expectation same as for χ^2 test

iii. All divided by its maximal value $1 - p_e$
iv. Result is called *kappa statistic*.

f. What should we match on?
i. Often match on traits that are expected to impact disease
ii. Matching is to remove effect of something associated with both putative cause and effect
iii. Matching can reduce efficiency:
- If you match on something correlated to exposure,

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

▷ you get pairs with similar exposure
▷ that don't give much info about effect of exposure on disease
- 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

- Both are known as *over-matching*
- Sometimes matched pairs are multiple observations on one individual.

g. Estimation for Matched pairs
i. From (1), pairs have probabilities

	0	1
0	$\frac{\pi_0 P_0}{(P_1\pi_1+P_0\pi_0)}\frac{[1-\pi_0]P_0}{P_1[1-\pi_1]+P_0[1-\pi_0]}$	$\frac{\pi_0 P_0}{(P_1\pi_1+P_0\pi_0)}\frac{[1-\pi_1]P_1}{P_1[1-\pi_1]+P_0[1-\pi_0]}$
1	$\frac{\pi_1 P_1}{(P_1\pi_1+P_0\pi_0)}\frac{[1-\pi_0]P_0}{P_1[1-\pi_1]+P_0[1-\pi_0]}$	$\frac{\pi_1 P_1}{(P_1\pi_1+P_0\pi_0)}\frac{[1-\pi_1]P_1}{P_1[1-\pi_1]+P_0[1-\pi_0]}$

ii. $n_{10}|n_{10}+n_{10} \sim \text{Bin}(\pi_1(1-\pi_0)/[\pi_1(1-\pi_0)+\pi_0(1-\pi_1)], n_{10}+n_{01}) = \text{Bin}(\psi/(1+\psi), n_{10}+n_{01})$ after conditioning on $n_{10}+n_{01}$.
- $\omega = \psi/(1+\psi)$; $\psi = \omega/(1-\omega)$.

iii. Hence $\hat{\psi} = n_{10}/n_{01}$
iv. And get CI for ψ by getting binomial CI and transforming.

Se: 9 pp. 282–285

h. This is also Mantel–Haenszel estimator
i. Sometimes it is hard to make matched pairs,
i. because collection of subjects doesn't contain pair
ii. or setting up pairs is a lot of work

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

Se: 9 pp. 279–280

k. When matched groups are larger than 2
i. and not necessarily all the same size
ii. still use Mantel-Haenszel procedure
iii. exact binomial results no longer hold

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

B&D2: 4.1

E. Modeling disease rates in terms of covariates
1. Before
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
i. Suppose L covariates
- Includes constant 1
- For nickel smelters, might be indicators of exposure group
- For car example, might be age of driver, time of day of accident, etc.
- Includes dochotomous "response", if present.

b. Identify K relatively homogeneous groups
i. ie., same (or similar) values for all covariates
c. Need some structure betw. rates at different exposure levels
i. Interpret ability
ii. stability of estimates
d. We will assume linearity on log scale

B&D2: 4.3a

3. Assume that
a. numbers of events in an interval are Poisson
i. $\text{P}[O_j = d] = \exp(-\lambda_j Q_j)(\lambda_j Q_j)^d/d!$

ii. Implies that each person has chance $\exp(-\Delta\lambda_j)$ of surviving interval Δ without an event.
iii. As before, assume individuals act independently.
iv. Assume effectively $P_j = \infty$.
- Might not be true for communicable diseases.

b. Log linear model for effect of covariates
i. Suppose that x_{kl} is covariate l in group k
c. Fit model that says $\log(\lambda_k) = +\sum_{l=1}^{L} x_{kl}\beta_l = \boldsymbol{x}_k\boldsymbol{\beta}$
i. Bold faced quantities are vectors
ii. Multiplication in last expression is inner product.
iii. Choice of stratification vs interest variables is arbitrary
d. $O_k = P_k \exp(\boldsymbol{x}_k\boldsymbol{\beta}) + \epsilon_k$ for
i. Approximately, $\epsilon_{k-} \sim \mathcal{N}(0, P_k\lambda_{k-})$

4. Fitting the model
a. Start with a guess of best values $\boldsymbol{\beta}$
i. call them $\boldsymbol{\beta}^0$
ii. almost any value (like 0) will do
b. $O_k \approx P_k\lambda_k^0[1 + \boldsymbol{x}_k(\boldsymbol{\beta} - \boldsymbol{\beta}^0)] + \epsilon_k$,
i. $\lambda_k^0 = \exp(\boldsymbol{x}_k\boldsymbol{\beta}^0)$
ii. $\epsilon_k \sim \mathcal{N}(0, \lambda_k^0)$
iii. Now this looks like a regular regression problem
- except that variances of errors are not equal.

iv. $(O_k - P_k\lambda_k^0)/(P_k\lambda_k^0) \approx \boldsymbol{x}_k(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \epsilon_k^*$
v. $\text{Var}[\epsilon_k^*] \approx 1/[P_k\lambda_k^0]$
c. Use multiple regression to update guess
i. Do multiple times
ii. Method is called *iteratively reweighted least squares*.

5. Model is an example of a *generalized linear model*.
   a. More specifically, *Poisson regression*
   b. Parameter estimates are logs of relative risk
   c. Testing done via
      i. standard errors, which come from Delta method (Wald test)
         - Also gives CI

         B&D2: 4.3c–d
      ii. likelihood ratio
   d. Complications:
      i. Do iterations bounce back and forth without converging?
      ii. Sometimes best fits for parameters are $\pm\infty$
      iii. Tests can mislead when some groups have small expected value
   e. Does model fit well?
      i. Predicted mean values for each of the groups ought to be about right
      ii. Hence $\sum_k (O_k - e_k)^2/e_k$ ought to be approximately χ^2
         - DF is number of groups - number of parameters

         B&D1: 6.4
      iii. Alternatively, use likelihood ratio
         - Write down probability for data
         - Express as function of unknown parameters
           ▷ Function L is called *likelihood*.
         - Parameter value that maximizes L is called the *maximum likelihood estimate*

         - H_0 is plausible if L is not much higher somewhere else.
         - Hence test hypothesis by comparing maximized value to value at null
           ▷ compare with ratio to get *likelihood ratio test*
           ▷ usually take log: $l = \log(L)$.
           ▷ $2\times$ difference in l generally approximately $\sim \chi^2_k$ for k the difference in number of unknown parameters.
6. Fitting multiple regression
   a. Setup: Response Y_j, explanatory variables x_{ij}
      i. Maybe $x_{1j} = 1$ for all j
   b. Want $Y_j = \boldsymbol{x}_j\boldsymbol{\beta} + \epsilon_j$
   c. A way to do the fitting:
      i. Let $R_j = Y_j$
      ii. Choose $\hat\beta_1$ to make x_{1j} best fit R_j:
         - $\hat\beta_1$ minimizes $\sum_j (R_j - \beta_1 x_{1j})^2$
         - $\hat\beta_1 = \sum_j x_{1j}R_j / \sum_j x_{1j}^2$
         - Now change R_j to what you haven't explained: $R_j =$ old $R_j - \hat\beta_1 x_{1j}$: residuals
      iii. Choose $\hat\beta_2$ to make x_{2j} best fit R_j:
         - after removing information about x_{1j} from x_{2j}:
           ▷ New $x_{2j} = x_{2j} - (\sum_l x_{1l}x_{2l})/(\sum_l x_{1l}^2)x_{1j}$
         - $\hat\beta_2 = \sum_j x_{2j}Y_j / \sum_j x_{2j}^2$
         - Adjust $\hat\beta_1$ for the fact that x_{2j} hat some x_{1j} in it.
      iv. Iterate

   d. Example: $x_{1j} = 1 \forall j$
      i. $\hat\beta_1 = (\sum_j 1 \times Y_j)/\sum_j 1^2 = \bar Y$
      ii. New x_{2j} is $x_{2j} - \bar x_2$ for $\sum_l x_{2l}/n$
      iii. $\hat\beta_2 = \sum_l (x_{2j} - \bar x_2)(Y_j - \bar Y)/\sum_l (x_{2j} - \bar x_2)^2$
      iv. New $\hat\beta_2$ is $\bar Y - \bar x_2 \hat\beta_1$.
      v. Subexample: For each j, either x_{2j} or x_{3j} is 1 and the other is 0.
         - Corresponds to model allowing for intercept and effect of membership in two groups
         - Then new $x_{3j} = 0$
         - Then $\hat\beta_3 = 0/0$
      vi. Hence can't estimate separate parameter values for intercept and all groups.

This page intentionally left blank.