

**Due Date:** Thursday, Sept. 15, in class

**Reading Assignment:** Collett, Chapters 1 and 2

1. In a study of the effect of oral contraceptive use on the incidence of breast cancer in New Zealand, Paul et al (1986) identified all breast cancer cases reported to the New Zealand National Cancer Registry over a particular two-year period. A random sample of women from the electoral rolls was used to identify women with no recorded history of breast cancer. The data are as follows

Used Oral Contraceptives	Breast Cancer Patients	Non-breast Cancer women
Yes	310	708
No	123	189
Total	433	897

Use these data to help answer the following questions.

- Is this an observational study or a randomized experiment?
  - Is this a case-control or a cohort study?
  - What is the sampling unit that provides the response?
  - Estimate the odds ratio (OR) corresponding to the odds that an oral contraceptive user develops breast cancer versus the odds that a non-user develops breast cancer.
  - Construct an approximate 95% confidence interval for the true odds ratio. (Use an estimate of the variance of the estimated log-odds based on the large sample normal approximation to the distribution of the estimator.)
  - Describe conditions under which the odds ratio (OR) considered in parts (d) and (e) provides a good estimate of relative risk of developing breast cancer for oral contraceptive users? Can you directly estimate relative risk from the data collected in this study? Explain.
  - Suppose the objective of the study was to estimate the relative risk of developing breast cancer for oral contraceptive users in New Zealand. Identify possible sources of bias for this study.
2. Framingham is an industrial city located about 20 miles west of Boston, Mass. In 1948, a cohort study was begun with the objective of identifying potential risk factors for coronary heart disease (CHD). Individuals found to be free of heart disease at the beginning of the study were followed for 12 years and those who developed CHD during that period were identified. At the beginning of the study measurements were also made on a number of other variables, including age, serum cholesterol level, gender, systolic blood pressure, and smoking history. The following table shows information on proportions of individuals who developed CHD for various levels of some potential risk factors.

Serum Cholesterol Level	Age group (years)	Sex	CHD	No CHD
< 190	30-49	Female	6	536
		Male	13	327
	50-62	Female	9	49
		Male	18	110
190-219	30-49	Female	6	547
		Male	18	390
	50-62	Female	12	123
		Male	33	143
220-249	30-49	Female	10	402
		Male	40	381
	50-62	Female	21	197
		Male	35	139
≥ 250	30-49	Female	18	339
		Male	57	305
	50-62	Female	48	347
		Male	49	134

These data are stored in the file CHD.txt on the assignment section of the course web page. There is one line for each count and five columns corresponding to initial serum cholesterol level, age group, sex, CHD status, and the count, respectively.

- (a) Consider just the data for males aged 50-62.

Initial Serum Cholesterol Level	CHD	No CHD
< 250	86	392
≥ 250	49	134

Directly estimate the relative risk (RR) of 60-62 year old males developing CHD for serum cholesterol levels of at least 250 versus those with serum cholesterol levels below 250. Do not use an odds ratio.

- (b) Construct an approximate 95% confidence interval for RR in part A. Show how to derive the large sample variance of the estimate of  $\log(\text{RR})$ .
- (c) Obtain an approximate estimate of RR using an appropriate odds ratio.
- (d) Construct an approximate 95% confidence interval for the “true” odds ratio.
- (f) Does an odds ratio provide a good approximation to RR in this situation? Comment.

3. Consider the data in part (a) of problem 2. Fit a logistic regression model of the form

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \beta_1 Z_j$$

where 
$$Z_j = \begin{cases} 0 & \text{if initial serum cholesterol} < 250 \text{ (j=1)} \\ 1 & \text{if initial serum cholesterol} \geq 250 \text{ (j=2)} \end{cases}$$

and

$$\pi_j = P(\text{CHD} | Z_j)$$

- Give an interpretation of  $e^{\beta_1}$ .
- Evaluate the m.l.e. for  $\beta_1$  and the mle for  $e^{\beta_1}$ .
- Construct an approximate 95% confidence interval for  $e^{\beta_1}$ . How do these results compare with those from parts (b) and (d) of problem 2?

4. Consider the data from the first table in problem 2. Find maximum likelihood estimates for the parameters in the model

$$\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \mu + \alpha_i + \beta_1 Z_j + \beta_2 W_k$$

where

$$Z_j = \begin{cases} 0 & \text{females (j=1)} \\ 1 & \text{males (j=2)} \end{cases}$$

$$W_k = \begin{cases} 0 & \text{for the 30-49 age group (k=1)} \\ 1 & \text{for the 50-62 age group (k=2)} \end{cases}$$

$\alpha_i$  corresponds to the  $i$ -th cholesterol level ( $i=1,2,3,4$ )

- Report values for the m.l.e.'s of the parameters and their standard errors.
- Give an interpretation of  $e^{\alpha_2 - \alpha_1}$  and construct an approximate 95% confidence interval.
- What are the major restrictions that this model imposes on the potential association between the risk factors (age, sex, cholesterol level) and relative odds of developing CHD? Explain how you could assess the validity of those restrictions.

5. In a prospective cohort study to determine whether or not low calcium intake among older women affects their risk of hip fracture, there is concern that study participation rates may be higher among women with high calcium intake than among women with low calcium intake. It is believed that participation rates are 90% for those with calcium intake above the median and 60% for those with calcium intake below the median. Discuss how this differential participation might have affected the results, which indicate that the ten-year incidence of hip fracture among women with calcium intake below the median is approximately 50% larger than the ten-year incidence of hip fracture among women with calcium intake above the median.
6. One objective of a study described by Kelsey and Hardy (1975, Amer. J. of Epid) was to determine if driving motor vehicles is a risk factor for lower back pain caused by acute herniated lumbar intervertebral discs. Cases were selected from adults between the ages of 20 and 64 living in the area of New Haven, Conn., who had X-rays taken of the lower back between June, 1971 and May, 1973. Those diagnosed with having acute herniated lumbar intervertebral discs, and who only recently had acquired symptoms of the disease, were used as cases. Referents were drawn from patients who were admitted to the same hospital or who presented at the same clinic as a case with a condition unrelated to the spine. Cases and referents were further matched on the basis of sex and age, so that the difference in the ages of the case and referent did not exceed 10 years in any matched pair. A total of 217 matched pairs were recruited, consisting of 89 female pairs and 128 male pairs. Information on whether or not the person was a driver was obtained from each member of each pair. The data are shown below.

		Referent	
		Driver	Non-Driver
Case	Driver	144	41
	Non-Driver	19	13

- (a) Perform McNemar's test and state your conclusions.
- (b) Perform an exact Binomial test. How does the result of this test compare with the result from McNemar's test? Is this what you expected?
- (c) Use a logistic regression model to estimate the conditional odds ratio for acute herniated lumbar intervertebral discs for drivers versus non-drivers, given the pair. Construct a 95% confidence interval.
- (d) Using the row and column totals in table, construct an estimate of the marginal odds ratio of acute herniated lumbar intervertebral discs for drivers versus non-drivers. Using the incorrect assumption of independent samples of cases and referents, construct a 95% confidence interval. How does ignoring the matching affect your results?

7. The file posted as lworld.dat on the hw part of the course web page contains data on contains 315 records with data on cases and controls from the Leisure World study of eudiometrical cancer as related to treatment with estrogens for menopausal symptoms and other risk factors. There are 63 groups with one eudiometrical cancer case matched to four controls. Matching was done by age. See the article by Mack et al in NEJM 294:1262-1267, 1976 for a full description. There is one line in the data file for each subject and the values of the nine variables are in the following order on each line.

Order	Variable Name	Description	Codes/Range
1	GROUP	Matched group indicator	1-63
2	CASE	Case-control indicator	1 = Case; 0 = control
3	AGE	Age in years	55-84
4	AGEG	Age group indicator	1 = 55-64; 2 = 65-74; 3 = 75+
5	EST	Estrogen usage	1 = No; 2 = Yes
6	GALL	Gallbladder disease	1 = No; 2 = Yes
7	HYP	Hypertension	1 = No; 2 = Yes
8	OB	Obesity	1 = No; 2 = Yes; 3 = Unknown
9	NON	Non-estrogen drug	1 = No; 2 = Yes

AGEG is based on the age of the case in each matched set.

SAS code that reads the data file is posted as lworld.sas and R code that enters the data into a data frame is posted as lworld.R. Each of these programs recodes the EST, GALL, HYP, and NON variables as 0=NO and 1=Yes. The obesity variable is recoded as two binary variables.

Use conditional logistic regression analysis to examine the association between incidence of eudiometrical cancer and estrogen usage (EST).

- Fit a model that includes only EST as a risk factor. State your conclusion about the conditional association between estrogen treatment and incidence of eudiometrical cancer. Use an appropriate odds ratio to quantify your conclusion in terms of relative risk.
- What does the model in part (a) assume about the conditional association between use of estrogen to treat menopausal symptoms and incidence of eudiometrical cancer?
- Fit a model that includes only EST, GALL, HYP, NON, and the two dummy variables for obesity. State your conclusion about the association between estrogen treatment and incidence of eudiometrical cancer. Use an appropriate odds ratio to quantify your conclusion in terms of relative risk. Was your conclusion about the association between estrogen use and risk of eudiometrical cancer altered by including the other potential risk factors in the model? Why might you expect to obtain different estimates of odds ratios in parts (a) and (c)?
- Using the model in part (c), state your conclusions about the other potential risk factors: presence of gallbladder disease (GALL), presence of hypertension (HYP), obesity, and use of non-estrogen drugs for menopausal symptoms (NON).