Biostatistics 513 Spring 2007

Midterm Exam NAME:

1. A study was conducted to investigate the association between high coffee consumption and non-fatal myocardial infarction (MI) in women coffee drinkers. The study design was a case-control design where MI patients (**y** = 1) and community controls (**y** = 0) were asked about their coffee drinking habits.

Suppose the exposure variable to drinking coffee was coded as: **x** = +1 (high coffee consumption), and **x** = -1 (low coffee consumption). A logistic regression model relating **x** to non-fatal MI was fitted and yielded the following estimates:

```
------------------------------------------------------------------------------
           y |    Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .3466     etc.
       _cons |   .7456     etc.
------------------------------------------------------------------------------
```

a) Provide an estimate of the odds ratio of disease (non-fatal MI) to exposure to disease to non-exposure. Does the exposure appear to increase or decrease the risk of non-fatal MI?

b) Suppose the exposure variable, **x**, had been coded as 0 = unexposed and 1 = exposed. What would the estimated coefficient for the exposure variable be?

2. A study was conducted to investigate whether longer duration of hormone replacement therapy (HRT) use was associated with a lower risk for myocardial infarction (MI) in postmenopausal women. Data from a sample of 1000 case subjects (i.e., a random sample of post-menopausal women enrolled in a large HMO with incident fatal or nonfatal myocardial infarction from January 1990 through December 1999) were collected using medical records and telephone interviews with consenting survivors. Control subjects were a random sample of postmenopausal women enrolled in the HMO without MI and matched individually to case subjects by age and calendar year. All postmenopausal women not on HRT were excluded from this study. The use of hormones was then ascertained using the HMO's computerized pharmacy database. HRT exposure was dichotomized as long duration and short duration use.

What method would you use to statistically compare long duration of HRT to incident non-fatal or fatal MI? Please justify your response and reference appropriate tests or estimates you would use.

3. Graham et al. (1981) studied dietary factors in the epidemiology of cancer of the larynx. Interviews were carried out with 338 male patients at Roswell Park Memorial Institute with cancer of the larynx, and with 359 male controls with diseases other than of the digestive or respiratory system (and without newplasms).

This table compares vitamin A (IU/month) intake for cases and the controls.

	Cases	Controls	Total
<50,500	98	78	176
≥50,500	240	281	521
Total	338	359	697

a) What are the appropriate null and alternative hypotheses for testing the association between vitamin A intake and cancer?

b) Give a point estimate for the association between vitamin A intake and cancer (i.e., relative risk, odds ratio, risk difference) and interpret your estimate.

An additional analysis of vitamin C showed the following results:

	Cases	Controls	Total
<1000 (mg/month) vit C	112	75	187
≥1000 (mg/month) vit C	226	284	510
Total	338	359	697

The following statistics are available:

```
           |      Point estimate    |    [95% Conf. Interval]
           |------------------------+------------------------
Odds ratio |         1.876578       |    1.316656     2.67981
           +-----------------------------------------------
                        chi2(1) =    13.30  Pr>chi2 = 0.0003
```

c) Interpret the χ^2 statistic with respect to the relationship between disease and vitamin C intake. What can you conclude from the test?

The actual data presented in Graham et al. (1981) are given as follows:

Vitamin C	Cases	Controls	Unadjusted OR
< 1000	112	75	1.00 (reference)
1000-1400	116	138	0.56
1400-1800	74	85	0.58
> 1800	36	61	0.40
Total	338	359	

```
Test of homogeneity (equal odds): chi2(3)  =    15.79
                                  Pr>chi2  =   0.0013

Score test for trend of odds:     chi2(1)  =    12.45
                                  Pr>chi2  =   0.0004
```

d) State (in words or in symbols that you define) the null hypothesis and alternative hypotheses for testing whether there is a trend in disease status with vitamin C consumption.

H_0:

H_1:

Consider how a logistic regression model could be used to test for a trend in the odds of disease with increased vitamin C consumption.

e) Define a covariate, X_1, representing vitamin C consumption, and define a logistic regression model using X_1, that could be used to test for trend.

f) Define the null hypothesis and alternative hypothesis based on your logistic regression model that would be used to test for trend.

g) What test statistic would you use to execute the test of the hypothesis given in part (f) above? (Please be explicit.)

h) Additional analyses found that vitamin C could be reasonably modeled as a "grouped linear" variable. Formulate a logistic regression model to investigate whether vitamin A consumption modifies the association between vitamin C consumption and cancer. Define your vitamin A variable (given in problem 3(a)) as X_2 in your logistic regression model. Also state the null hypothesis to investigate the association (using parameters from your stated logistic regression model).

4. The following data were taken from the manuscript: "Breast Cancer, Lactation History, and Serum Organochlorines" by Romieu et al. (2000) *AJE*. Recent studies have suggested that exposure to low levels of the toxins DDT and DDE (organochlorines) is associated with breast cancer. A case-control study of women who had given birth to at least one child was conducted in Mexico City, Mexico.

```
The following variables are reported in Romieu et al.:

DDT:  1 = 0.023-0.070   micro g / g lipids      (serum measurement)
      2 = 0.071-0.10    micro g / g lipids
      3 = 0.11-0.18     micro g / g lipids
      4 = 0.19-5.41     micro g / g lipids

DDE:  1 = 0.20-1.16     micro g / g lipids      (serum measurement)
      2 = 1.17-1.96     micro g / g lipids
      3 = 1.97-3.48     micro g / g lipids
      4 = 3.49-14.84    micro g / g lipids

POST: 0 = premenopause
      1 = postmenopause

CASE: 0 = control
```

```
      1 = case (breast cancer)

COUNT: number of subjects
```

The goal of the study was to assess the relationship between exposure and the risk of breast cancer. A total of 126 cases were obtained and 120 community controls were also recruited. A dichotomous exposure variable was created:

```
      DDEhigh=1 if DDE 1.97-14.84   micro g / g lipid
      DDEhigh=0 if DDE 0.20-1.96    micro g / g lipid
```

a) A crude analysis of the relationship between DDEhigh and CASE yielded:

```
                          DDEhigh=1 (high)    DDEhigh=0 (low)
Case=1 (breast cancer)           82                  38
Case=0 (control)                 63                  63

Odds ratio estimate: 2.16
95% Confidence Interval for the OR: (1.29, 3.62).
```

Interpret the odds ratio, and interpret the confidence interval for the odds ratio (is it a significant association?)

Additional analysis revealed that CASE status and menopause status (POST) were associated (OR=1.178), and menopause status was associated with exposure (OR=5.899).

A stratified analysis yielded:

```
Odds Ratios comparing CASE odds among DDEhigh=1 (high) to DDEhigh=0 (low):

      Strata       OR           95% Conf. Interval
      POST=0       1.907        (0.910, 3.997)
      POST=1       3.093        (1.257, 7.581)

Test of Homogeneity: (Breslow-Day) chi2(1) statistic = 0.64, p-value = 0.422

      Crude Odds Ratio Estimate:                    2.158
      Mantel-Haenszel Common Odds Ratio estimate:   2.326
      95% Confidence Interval for Common OR:        (1.309, 4.132)
```

b) Is a common odds ratio estimate appropriate based on these statistics? Justify your answer.

c) Give an explicit interpretation of the common odds ratio estimate (OR estimate = 2.326).

d) If a similar stratified analysis was performed to evaluate DDEhigh but using the levels of DDT as the stratifying variable, then what would be the hypothesis of homogeneity of the odds ratios and what would be the degrees of freedom for a test of this homogeneity hypothesis? (Please be explicit.)

e) Given the crude and adjusted analyses, would you conclude that menopause status is a confounder? Justify your answer.

A subsequent analysis used logistic regression with dummy variables to code for the variable DDE. The results of this model are:

Note: DDE=1 is the reference category and no dummy variable is included.

Name	OR	s.e.	Z	p-val	95% Conf. Interval
DDE=2	1.107	0.457	0.246	0.806	(0.493, 2.488)
DDE=3	2.213	0.876	2.007	0.045	(1.019, 4.809)
DDE=4	2.814	1.186	2.455	0.014	(1.232, 6.429)
POST	0.796	0.236	-0.770	0.441	(0.445, 1.423)

log likelihood = -81.206

f) Interpret the odds ratio for DDE=4. (Describe the specific comparison that is made).

g) A model with only the POST variable gave a log likelihood of -170.23. Complete the following expressions that refers to a likelihood ratio test comparing the model above to the null model that only has the POST variable:

Likelihood Ratio Statistic = LR = _____

Degrees of freedom for the LR Test = _____

Null hypothesis H_0: _____

Alternative hypothesis H_1: _____

Further analysis found that a linear model ("grouped linear") for DDE was appropriate (when compared to the dummy variable model using a LR test). Logistic regression was then used to assess whether the effect of DDE exposure appeared to depend on menopause status by fitting the model:

$$\text{logit}[\pi(X)] = -0.722 + .269\ \text{DDE} - 0.889\ \text{POST} + 0.242\ \text{POST}\times\text{DDE}$$

h) Based on this model, what is the estimated odds ratio comparing premenopausal women with DDE=3 (POST=0, DDE=3) to premenopausal women with DDE=1 (POST=0, DDE=1)?

i) Based on this model, what is the estimated odds ratio comparing postmenopausal women with DDE=3 (POST=1, DDE=3) to postmenopausal women with DDE=1 (POST=1, DDE=1)?

j) Likelihood ratio testing indicated that the DDE×POST interaction was not significant. However, additional interest is in the effect of DDE adjusting for both POST and DDT. What logistic regression model could be used for this question? What is(are) the parameter(s) in your model that would describe the effect of interest (adjusted DDE)?