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?

The OR estimate comparing high consumption coffee drinkers to low consumption coffee drinkers is obtained using the logistic regression model as follows:

$$\log[OR] = \log it[\pi(x=+1)] - \log it[\pi(x=-1)] = [\beta_0 + \beta_1] - [\beta_0 - \beta_1] = 2\beta_1.$$

The estimated odds ratio comparing the two groups is then

$$\widehat{OR} = \exp[2\widehat{\beta_1}] = \exp[2(0.3466)] = \exp[0.6932] = 2.00.$$

High coffee consumption appears increase the risk of non-fatal MI in women coffee drinkers.

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?

The dummy variable coding would provide a direct estimate of the log odds ratio comparing the high coffee consumption group to the low coffee consumption group. The estimated OR would not change from part 1(a), so the coefficient under the dummy variable coding would be:

$$\widehat{\beta}_{1\,new} = \log[\widehat{OR}] = \log[2.00] = 0.6932 = [2(0.3466)] = 2\widehat{\beta}_{1\,old}$$

The estimated regression coefficient is double the estimated regression coefficient in part 1(a).

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.

This is a matched case-control study (matched by age and calendar year). It would be appropriate to use McNemar's test. One would assume potential confounders are reasonable balanced by matching. Two alternate methods of analysis are (1) standard logistic regression analysis (that includes HRT exposure as the predictor of interest and adjusts for age and calendar year) or conditional logistic regression (that more formally takes into account the matching variables). The solution sought was McNemar's test. The latter two methods are more advanced (and would be investigated formally in Biost 536).

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?

Acceptable solutions for the null hypothesis are: cancer of the larynx and vitamin A intake are not associated, $P[Exposure \mid Disease] = P[Exposure \mid \overline{Disease}]$*, or the odds of cancer for low vitamin C intake (<50,500) to the odds of cancer for high vitamin C intake (≥50,500) is equal to 1.0. Acceptable alternative hypotheses are the compliments of the null hypotheses (i.e., there is an association between cancer of the larynx and vitamin A intake,* $P[Exposure \mid Disease] \neq P[Exposure \mid \overline{Disease}]$*, or* $OR \neq 1$.

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.

The odds ratio is an appropriate summary given the design was a case-control design. The relative risk and risk difference are not directly estimable and would not be appropriate here..

*The OR estimate is 98*281/(78*240) = 1.47.*

The estimated odds of cancer of the larynx among men that report vitamin A intake less than 50,500 IU/month to the odds of cancer of the larynx among men that report vitamin A intake greater than or equal to 50,500 IU/month was 1.47. The estimate indicates there is an increased risk of cancer of the larynx associated with low vitamin A intake.

(A relative risk interpretation of the odds ratio estimate would be acceptable, too, provided one also indicates that the cancer of the larynx is rare, and therefore, the odds ratio estimate is approximately equal to the relative risk.)

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 chi-square statistic can be used to test the association between cancer of the larynx and vitamin C intake (< 1000 mg/month vs 1000+ mg/month). The chi-square statistic, under the null hypothesis (i.e., there is no association between cancer of the larynx and vitamin C intake) is distributed as a chi-square random variable with degrees of freedom =(2-1)(2-1) = 1. The test statistic for these data is 13.3 and has a p-value of 0.003. The probability of observing this result given the null hypothesis is true is very unlikely. We would therefore reject the null hypothesis in favor of the alternative hypothesis. There is an association between cancer of the larynx and vitamin C intake.*

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.

You can write the null hypothesis in terms of odds

$$H_0: odds_1 = odds_2 = odds_3 = odds_4$$

Where $odds_1$ *corresponds to the odds for vitamin C intake < 1000 mg/mo,*
$odds_2$ *corresponds to the odds for vitamin C intake [1000-1400] mg/mo,*
$odds_3$ *corresponds to the odds for vitamin C intake [1400-1800] mg/mo,*
$odds_4$ *corresponds to the odds for vitamin C intake > 1800 mg/mo.*

One may also write the null hypothesis in terms of odds ratios by selecting a stratum as the reference category (e.g., < 1000 mg/mo)

$$H_0: odds_1/odds_1 = odds_2/odds_1 = odds_3/odds_1 = odds_4/odds_1$$

Equivalently,

$$H_0: 1 = OR_2 = OR_3 = OR_4$$

Each odds ratio in this example is in reference to the vitamin C intake group (< 1000 mg/mo).

The alternative hypothesis for the test of trend would be:

$$H_1: odds_1 \leq odds_2 \leq odds_3 \leq odds_4 \text{ or } odds_1 \geq odds_2 \geq odds_3 \geq odds_4$$

with at least one strict inequality (< or >) above.

In terms of odds ratios, we have

$$H_1: 1 \leq OR_2 \leq OR_3 \leq OR_4 \text{ or } 1 \geq OR_2 \geq OR_3 \geq OR_4$$

with at least one strict inequality (< or >).

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.

X_1 would be defined as follows: 1 if vitamin C is < 1000 mg/mo; 2 if vitamin C is between [1000, 1400]; 3 if vitamin C is between [1400, 1800]; 4 if vitamin C is > 1800.

A logistic regression model that could be used to test for trend is $\log it[\pi(X_1)] = \beta_0 + \beta_1 X_1$

where $\pi(X_1) = E[Y = 1 | X_1] = \mathrm{P}[cancer | X_1]$. *$X_1$ is defined in terms of the levels (1,2,3,4) and modeled as if they were values of a continuous variable.*

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

The null and alternative hypotheses used in testing for trend via the logistic regression model reduces to testing the regression coefficient of X_1 in 3(e) above,

H_0: $\beta_1 = 0$ versus H_0: $\beta_1 \neq 0$

The null hypothesis indicates the trend is a horizontal (flat) line while the alternative hypothesis indicates the trend is either increasing (or decreasing) with the levels of the ordered categories of vitamin C.

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

The Wald statistic $z = \hat{\beta}_1 / \widehat{se}(\hat{\beta}_1)$ *can be used to test the hypothesis given in part 3(f). Assuming the null hypothesis in 3(f) is true, the Wald statistic, z, is approximately normal with mean 0 and standard deviation equal to 1. (One could also square the z statistic to obtain a test statistic that has a chi-square distribution with one degree of freedom if the null hypothesis is true.*

Alternatively, one could employ a likelihood ratio (LR) test. One would (1) compute a full (i.e, a model that includes X_1) and a reduced (i.e., a model that excludes X_1) model, (2) obtain the likelihoods under the two models, (3) compute the absolute value of the difference in the two estimated likelihoods and then multiply that number by 2. This LR statistic will be approximately distributed as a chi-square random variable with degrees of freedom equal to 2-1 = 1.

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).

The question requests one investigates whether vitamin A (X_2) is an effect modifier of the relationship between cancer of the larynx and vitamin C (X_1). Because vitamin C can be modeled as a grouped linear variable, we can investigate this question with a relatively simple model. We write

$$\log it[\pi(X_1, X_2)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 .$$

The interaction term would be used to investigate effect modification. We would test for effect modification by examining if β_3 is non-zero, i.e., $H_0 : \beta_3 = 0$.

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?)

The estimated odds of breast cancer among women that report high DDE exposure are 2.16-fold larger than the odds of breast cancer among women that report low DDE exposure. The 95 percent confidence interval for the estimated odds ratio is (1.29, 3.62). Because the 95 percent confidence interval does not contain the value 1.0 (which would indicate there is no association between breast cancer and DDE exposure), we conclude there is a significant association between breast cancer and DDE exposure.

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.

Yes, a common odds ratio for the two strata (i.e., premenopause and postmenopause status) is appropriate. The Breslow-Day test of homogeneity indicates the odds ratio between the two strata are not statistically different (p-value > 0.400). Even though the two estimates may seem different, their difference is considered to be within the range of the sampling error.

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

The estimated odds of breast cancer among women that report high DDE exposure are 2.326-fold larger than the odds of breast cancer among women that report low DDE exposure, after adjusting for women's menopause status.

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.)

Stratifying by DDT (a 4-level categorical variable) in investigating the association between breast cancer status and DDEhigh, the hypotheses of homogeneity is

H_0: $OR_1 = OR_2 = OR_3 = OR_4$
H_1: *at least one odds ratio is different from the others.*

The degrees of freedom is equal to the number of levels of the stratifying variable less one (df = 4-1=3).

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

One could conclude that menopause status is a confounder because the difference between the crude OR (2.16) and the adjusted OR (2.33) is a meaningful difference. (It is about an 8 percent change). One may decide that menopause status is a weak confounder. Finally, one could conclude that menopause status is not a confounder because the change between the two estimates is not meaningful (this might need defending) or because the change is less than 10 percent. The last argument, by itself, is the weakest argument of the four.

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).

The estimated odds of breast cancer among women with the highest DDE exposure level [3.49-14.84 micro g per g lipids] is 2.814-fold larger than the odds of breast cancer among women with the lowest DDE exposure level [0.20-1.16 micro g per g lipids] in comparing women with the same menopause status.

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 = *2*(-81.206 - (-170.23)) = 178.048*

Degrees of freedom for the LR Test = DF = 4-1=3

Null hypothesis H_0: $\beta_{DDE(2)} = \beta_{DDE(3)} = \beta_{DDE(4)} = 0$

Alternative hypothesis H_1: *at least 1 of the 3 parameters in H_0 is non-zero*

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)?

Using the stated logistic regression model,

log(OR) = logit(pi(Post=0,DDE=3)) - logit(pi(Post=0,DDE=1))
*= (b0 + 3 b1 + 0 b2 + (0*3) b3) - (b0 + 1 b1 + 0 b2 + (0*1) b3)*
= 2 b1.

*The estimated OR is exp(2*b1) = exp(0.538) = 1.72.*

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)?

Following the procedure used above in 4(h),

log(OR) = logit(pi(Post=1,DDE=3)) - logit(pi(Post=1,DDE=1))
*= (b0 + 3 b1 + 1 b2 + (1*3) b3) - (b0 + 1 b1 + 1 b2 + (1*1) b3)*
= 2 b1 + 2 b3.

The estimated OR is exp(2(b1+b3)) = exp(1.022) = 2.78.*

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)?

The question asks one to investigate whether menopause status and DDT together confound the association between DDE and breast cancer. The logistic regression model to investigate this question is

logit(pi(x)) = b0 + b1 DDE + b2 Post + b3 DDT(2) + b4 DDT(3) + b5 DDT(4).

We would examine the estimated coefficient for DDE in the model that included only DDE, logit(pi(x)) = b0 + b1 DDE (i.e., the unadjusted model), and compare it to the estimated coefficient for DDE in the full model above. Note: we use DDE (ordered categorical variable) as a grouped linear variable (see top of page 9 of the exam). There was no indication that DDT should be treated as a grouped linear variable. DDT was treated as an ordinary categorical variable.