Midterm Exam NAME: **KEY**

1. A subset of data from the Ille-et-Vilaine study was analyzed to investigate the association between esophageal cancer (alcoholic) cider consumption.

agegp (Age groups)
1 = 25-34 (years)
2 = 35-44
3 = 45-54
4 = 55-64
5 = 65-74
6 = 75+

cider 1 = High consumption
0 = Low consumption

case 1 = esophageal cancer
0 = control

The output from an analysis investigating this association is given below:

```
cc case cider, by(agegp)

     age group |     OR    [95% Conf. Int.]
---------------+--------------------------
         25-34 |   1.00
         35-44 |   8.52      1.09   381.84 (exact)
         45-54 |   2.33      1.10     5.45 (exact)
         55-64 |   1.62      0.89     9.67 (exact)
         65-74 |   1.63      0.80     6.94 (exact)
           75+ |   3.36      0.72    16.33 (exact)
---------------+--------------------------
         Crude |   1.95      1.39     2.72 (exact)
  M-H combined |   2.02      1.44     2.85
---------------------------------------------

Test of homogeneity (Tarone)
            chi2(5) = 4.40 Pr>chi2 = 0.4938

Test that combined OR = 1:
            Mantel-Haenszel chi2(1) =  16.66
                           Pr>chi2 = 0.0000
```

a) [4pt] Which of the following statements is true? Circle one.

i. The ORs for esophageal cancer comparing high cider consumption to low cider consumption are different among the age groups. There is evidence that age is an effect modifier.

ii. The p-value for the test that the Mantel-Haenszel pooled OR = 1 is less than 0.0001. There is evidence that age is an effect modifier.

iii. The p-value for the Tarone test of homogeneity is 0.4938. There is no evidence age is an effect modifier.

iv. None of the above.

b) [4pt] Which of the following statements is true? Circle one.

i. The ORs for esophageal cancer (comparing high cider consumption to low cider consumption) are different among the age groups. There is evidence that age is a confounder.

ii. The p-value for the test that the Mantel-Haenszel pooled OR = 1 is less than 0.0001. There is evidence age is a confounder.

iii. The results for the crude and Mantel-Haenszel OR estimates demonstrate that age is not a strong confounder in the association between high cider consumption and esophageal cancer.

iv. None of the above.

c) [6pt] In a complete sentence, please provide and interpretation for the Mantel-Haenszel combined odds ratio estimate that you might give in the results section for a journal submission.

After adjusting for age, the odds of esophageal cancer for individuals who consume high levels of alcoholic cider is more than two times the odds of individuals who do not consume high levels of alcoholic cider (OR_{MH} = 2.02, 95% CI = [1.44, 2.85]).

d) [6pt] A reviewer of your results believes adjusting for age (in years) would provide a fairer comparison between cancer status and cider consumption. Briefly explain how you might investigate this question and adjust for age as a continuous confounder.

Since one cannot directly stratify for a continuous confounder in a stratified analysis, one should adjust for a continuous age variable as a confounder – for estimating the association between esophageal cancer and alcoholic cider consumption – by fitting a logistic regression model. The model would include age, as a second predictor variable in a logistic regression model,

$$logit(\pi) = \beta_0 + \beta_1\, cider + \beta_2\, age$$

Here, $logit(\pi)$ represents the log odds of esophageal cancer, *cider* is the 0/1 indicator variable for high cider consumption, and age represents the *age* of the individuals in years.

2. [6pt] 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. A random sample of 1000 postmenopausal women without MI were selected as controls. 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.

Given this is a case-control study, one cannot directly estimate the risk of disease (MI) for the exposure (duration length of HRT use) groups. To investigate the association between MI status and HRT use, one could use the standard Pearson X^2 statistic (test of homogeneity).

One could also estimate an odds ratio (OR) for an effect estimate of the magnitude of the association between HRT use and the odds of having an MI. If one were to employ a test of association using the OR summary, the null hypothesis (of no association) would be H0: OR = 1 versus the alternative hypothesis that the OR is not equal to unity.

Finally, one could also test for an association between MI status and HRT use by constructing a test that the probabilities of *exposure* differ between the disease groups. That is, H0: P[HRT | MI case] = P[HRT | MI control], versus the alternative hypothesis that the two probabilities are not equal. The test will be valid and yield identical results to the Pearson X^2 test above, although the summary statistic for the exposure probabilities is less interesting.

3. The following data are taken from Palmer et al. (1995) who investigated the relationship between coffee consumption and non-fatal myocardial infarction (MI) in women. The study is a case-control study where MI patients and community controls are asked about their coffee drinking habits. The table presented here refer only to the caffeinated coffee drinkers.

level	Cups/day	Cases	Controls	% Cases
1=	None	94	108	47
2=	1-2	195	193	50
3=	3-4	134	140	49
4=	5-6	82	41	67
5=	7-9	36	9	80
6=	> 9	29	8	78
	Total	570	499	53

a) [6pt] State (in words or in symbols that you define) the null hypothesis and alternative hypotheses for testing whether there is a trend between MI odds and coffee consumption.

H_0: **$odds_1 = odds_2 = odds_3 = odds_4 = odds_5 = odds_6$**

H_1: **$odds_1 \le odds_2 \le odds_3 \le odds_4 \le odds_5 \le odds_6$** ***or*** **$odds_1 \ge odds_2 \ge odds_3 \ge odds_4 \ge odds_5 \ge odds_6$**

Consider how a logistic regression model could be used to test for a trend in the odds of MI with coffee consumption.

b) [6pt] Define a covariate, X_1, representing coffee 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 no coffee consumption is 1-2 cups/day; 2 if coffee consumption is 1-2 cups/day; 3 if coffee consumption is 3-4 cups/day;…6 if coffee consumption is greater than 9 cups/day.

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 *logit(π)* is the log odds of a non-fatal MI, X_1 is defined in terms of the levels (1,2,3,4, 5, 6) and modeled as if they were values of a continuous variable.

c) [6pt] 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(b) above,

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

d) [6pt] What test statistic would you use to execute the test of the hypothesis given in part (c) above? (Please be explicit.)

The Wald statistic $z = \beta_1/se(\beta_1)$ can be used to test the hypothesis given in part 3(c). Assuming the null hypothesis 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.

4. A cohort study of N=609 white males was undertaken to investigate the association between high catecholamine levels and coronary heart disease (CHD). Study participants were disease free at the beginning of the study. They were followed for seven years, during that time incident CHD events (1=yes, 0=no) were ascertained. The outcome, exposure and other potentially important variables used in the analysis are noted as:

chd = coronary heart disease (1 = incident CHD, 0 = no CHD)
cat = catecholamine status (1 = high level, = low level)
age = age in years
chol = serum cholesterol (mg/dL)
htn = hypertension status (1 = hypertensive, 0 = normal)

The study investigators fit the following logistic regression model to investigate the associations between CHD risk and high catecholamine levels.

$$\textbf{logit}(\pi(X)) = \beta_0 + \beta_1\, cat + \beta_2\, age + \beta_3\, chol + \beta_4\, htn + \beta_5\, cat \times htn$$

The output from an analysis investigating the associations is given below:

```
Logistic regression             Number of obs =    609
                                LR chi2(5)    =  37.04
                                Prob > chi2   = 0.0000
Log likelihood = -200.7         Pseudo R2     = 0.0845

------------------------------------------------------
         chd |     OR  Std. Err.      z    P>|z|
-------------+----------------------------------------
         cat |   6.00   0.4525     3.96   0.000
         age |   1.03   0.0148     1.99   0.047
        chol |   1.01   0.0033     2.52   0.012
         htn |   2.48   0.3254     2.81   0.005
     catXhtn |   0.50   0.2357    -2.94   0.003
------------------------------------------------------
```

a) [6pt] Using the above model, obtain an estimate of the odds ratio of CHD comparing high to low catecholamine levels for individuals that *do not have* hypertension and are the same age and cholesterol levels.

The estimated odds ratio is exp(β_{cat}) = 6.0.

b) [6pt] Using the above model, obtain an estimate of the odds ratio of CHD comparing high to low catecholamine levels for individuals that *have* hypertension and are the same age and cholesterol levels.

The estimated odds ratio is

exp($\beta_{cat} + \beta_{catXhtn}$) = exp($\beta_{cat}$)*exp($\beta_{catXhtn}$) = 6*(1/2) = 3.

c) [6pt] Investigate whether the association between CHD and high catecholamine levels varies with hypertension. (Formally state your hypotheses, significance level, test statistics, test results and conclusions.)

To test effect modification, we test whether the interaction term is statistically different from zero,

H_0: $\beta_{catXhtn} = 0$ versus the alternative H_1: $\beta_{catXhtn}$ is not equal to zero.

Rejecting the null hypothesis is evidence of effect modification. From the fitted logistic regression model, we can formally test for effect modification using the Wald test statistic, z = $\beta_{catXhtn}$/se($\beta_{catXhtn}$) = -2.94. Using an alpha level of one percent (or five percent), we would reject the null hypothesis and conclude that hypertension levels do modify the association between the MI status and catecholamine levels.

5. In a case-control study to determine risk factors of colo-rectal cancer, the analysis below attempts to evaluate the association between colo-rectal cancer risk and age.

agegp (Age groups)
1 = 25-34 (years)
2 = 35-44
3 = 45-54
4 = 55-64
5 = 65-74
6 = 75-84

case 1 = colo-rectal cancer
0 = control

age in years

Two logistic regression analyses were fitted to investigate the association between age and colo-rectal cancer. The first analysis modeled age as a continuous variable and the second analysis modeled age as a grouped linear variable. The results are given below.

MODEL 1: Linear in Age

```
Logit regression           Number of obs =    975
                           LR chi2(5)    =  82.85
                           Prob > chi2   = 0.0000
Log likelihood = -453.3    Pseudo R2     = 0.0837

------------------------------------------------
       case |   Coef.  Std. Err.      z    P>|z|
------------+-----------------------------------
        age |   0.056   0.0066     8.47   0.0000
      _cons |  -4.427   0.3894   -11.37   0.0000
------------------------------------------------
```

MODEL 2: Grouped Linear with respect to Age

```
Logit regression           Number of obs =    975
                           LR chi2(5)    =  87.29
                           Prob > chi2   = 0.0000
Log likelihood = -451.1    Pseudo R2     = 0.0882

------------------------------------------------
       case |   Coef.  Std. Err.      z    P>|z|
------------+-----------------------------------
      agepg |   0.574   0.0660     8.69   0.0000
      _cons |  -3.410   0.2679   -12.73   0.0000
------------------------------------------------
```

a) [6pt] Please provide an interpretation for the coefficient for **age** in Model 1.

The odds of colo-rectal cancer is approximately 1.056 times the odds of colo-rectal cancer for a person that is one year older than another person.

b) [6pt] Interpret the coefficient for **agegp** in Model 2.

The odds of colo-rectal cancer is approximately 1.78 times the odds of colo-rectal cancer for a person that is approximately 10 years older than another person.

c) [4pt] Judging from the fit of these two models, does it appear that the association between age and the odds of colo-rectal cancer is linear (on the logit scale)? Give a brief justification for your answer.

Informally, we see the estimated regression coefficient for the age variable in the grouped linear model is approximately 10 times the coefficient in the logistic regression model that scales age in years. The scale for age in the grouped linear model is in approximately 10 year increments. If we were using the linear model (in age) and wished to estimate the odds ratio for two individuals whose age differed by 10 years, we would have

$$\text{logit}(\pi \mid age = 10) - \text{logit}(\pi \mid age = 0) = \beta_0 + 10\,\beta_1 - \beta_0 = 10\beta_1 = 0.560$$

This is approximately equal to comparing two individuals, in the grouped linear model, that are in to adjacent age groups

$$\text{logit}(\pi \mid agegp = x+1) - \text{logit}(\pi \mid agegp = x) = \beta_1 = 0.574$$

6. You were asked to participate in a study to investigate whether chronic estrogen use was a risk factor for Parkinson's disease among women. Parkinson's disease was denoted as **case** (1 = Parkinson's disease, 0 = no disease) and chronic estrogen use was denoted as **est** (1 = chronic estrogen use, -1 = not chronic estrogen use). *Please note estrogen was coded in an atypical fashion.* Your collaborators suggested you include hypertensive status, **htn** (1 = hypertensive, 0 = normal) in the analysis as a potential confounder.

The output from your analysis investigating this association is given below:

```
Logistic regression             Number of obs =    609
                                LR chi2(5)    =  37.04
                                Prob > chi2   = 0.0000
Log likelihood = -200.7         Pseudo R2     = 0.0845

------------------------------------------------
        case |     OR     SE*      z    P>|z|
-------------+----------------------------------
         est |   2.00   0.1690     4.10   0.000
         htn |   1.14   0.1310     1.00   0.320
------------------------------------------------
```

* *SE = standard errors for log(OR) estimates.*

a) [4pt] Provide an estimate of the odds ratio of disease (Parkinson's disease) for the exposure (chronic estrogen use) to disease for the unexposed.

The log odds ratio estimate comparing the two exposure groups is defined as

$\text{Logit}(\pi \mid \text{est} = 1, \text{htn}) - \text{Logit}(\pi \mid \text{est} = -1, \text{htn}) =$

$[\beta_0 + \beta_{est}(1) + \beta_{htn}\,\text{htn}] - [\beta_0 + \beta_{est}(-1) + \beta_{htn}\,\text{htn}] = 2\,\beta_{est}$

So the odds ratio is $\exp[2\,\beta_{est}] = (\exp[\beta_{est}])^2 = [OR_{est}]^2 = 2^2 = 4$

b) [6pt] You received a review from the journal and were asked to respond to reviewer comments. Referee A stated, "Given that hypertension status was not significantly associated with Parkinson's disease (when chronic estrogen use was included in the model), there is no reason to adjust for hypertensive status and it should be omitted from the model." Do you agree or disagree with the referee's request? Please explain why.

Referee A appears to be confused about the definition of a confounding variable. A confounder is (1) associated with the outcome in the population, (2) associated with the exposure in the sample (although some claim it should be causally associated with the exposure), and (3) not a mediating variable (i.e., in the causal path). Statistical association between the confounder and the outcome (here, in the sample) is not a criterion to proving a variable is a confounder. I disagree with this referee's claim.

c) [6pt] The second reviewer, Referee B, noted that, "It is well established that hypertension is a risk factor for Parkinson's disease, but new evidence also strongly suggests hypertension may result from chronic estrogen use." The referee recommends omitting hypertensive status from the model. Do you agree or disagree with the referee's request? Please explain why.

By the definition of a confounder, Referee B states that hypertension satisfies one of the three criteria of a confounder. That is, (1) hypertension is associated with Parkinson's disease. However, he indicates that chronic estrogen is *causing* hypertension, which violates the third criterion, i.e., hypertension is in the causal path between chronic estrogen use Parkinson's disease. One therefore should not adjust for hypertension in the logistic regression model. I agree with Referee B.