

## Lab #9 Discussion

**Question 1.** Note the estimated odds ratios for nonsmokers and smokers and their 95% confidence intervals, and the estimated summary (combined) odds ratio and its 95% confidence interval. What do the results of the test for homogeneity and inspection of the odds ratios allow you to conclude about the effect of smoking on the association between type A behavior and CHD? Is there any evidence for interaction? If we assume that no interaction is present, is there any evidence that smoking has a confounding influence on the measured association between *chd69* and *dibpat*?

```
. cs chd69 dibpat, by(smoke) or
```

smoke	OR	[95% Conf. Interval]		M-H Weight
nonsmoker	2.941176	1.882394	4.594855	12.10169 (Cornfield)
smoker	1.962697	1.385228	2.780702	23.66644 (Cornfield)
-----				
Crude	2.372929	1.804034	3.121147	
M-H combined	2.293753	1.741568	3.021016	
-----				
Test of homogeneity (M-H)		chi2(1) =	1.938	Pr>chi2 = 0.1639
Test that combined OR = 1:				
	Mantel-Haenszel	chi2(1) =	36.68	
		Pr>chi2 =	0.0000	

The estimated odds ratios (95% C.I.s) for the association between *chd69* and *dibpat* among non-smokers and smokers are 2.94 (1.88, 4.59) and 1.96 (1.39, 2.78), respectively. The odds ratio for the association between behavior pattern and CHD adjusted for the binary measure of smoking is 2.29 (1.74, 3.02). The *test of homogeneity* of odds ratios across the strata of the smoking variable is not significant at the 5% level. Visual inspection of the component odds ratios indicates that both represent positive associations between type-A behavior and CHD, and that the 95% confidence intervals overlap substantially. Recall that rejection of this test indicates that the effect of type A behavior on CHD differs in smokers and non-smokers, obligating us to consider smoking when investigating the association of the outcome with behavior. If we operate under the assumption that the indication of interaction is not strong enough to warrant further consideration, we can summarize the behavior-CHD association with the adjusted (MH-combined) odds ratio estimate. If this differs from the unadjusted (Crude) estimate, we can conclude that smoking has a confounding influence on the association of interest. In this case, the degree of confounding is very slight, with the adjusted measure of association differing from the crude estimate by 0.08.

The issue of what level of significance constitutes sufficient evidence for interaction is a delicate one and is somewhat context dependent. As mentioned in lecture, power to detect interactions in observational studies can be limited. This is related both to the fact that the test for homogeneity focuses on detecting all possible departures from the null hypothesis of equal odds ratios (rather than on detecting a particular departure) and that the additional stratification involved in evaluating interaction frequently leads to sparse

data in strata. Adopting a less stringent standard for rejection than the conventional 0.05 cut-off is one way to avoid missing potentially important interactions. However, uncritical adoption of this procedure may also lead to inappropriate identification of interaction. In this example, further investigation of this issue would include use of the underlying quantitative measure of smoking (`ncig`). We may also want to evaluate the impact of controlling for the interaction in further analyses involving additional predictors. As it stands, the relatively large P-value and the observation that both the component odds ratios reveal fairly strong positive associations indicate that this particular interaction is not likely to be important.

**Question 2.** *How do the estimated odds ratio and 95% confidence interval for `dibpat` compare to their counterparts from the results reported above?*

```
. logistic chd69 dibpat smoke
Logistic regression               Number of obs   =       3154
                                  LR chi2(2)       =       60.51
                                  Prob > chi2      =       0.0000
Log likelihood = -860.36587       Pseudo R2      =       0.0340
```

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
dibpat	2.301319	.3238066	5.92	0.000	1.746661 3.03211
smoke	1.8002	.2422852	4.37	0.000	1.3828 2.343593

The estimated odds ratio and 95% C.I. for `dibpat` from the logistic regression fit is 2.30 (1.75, 3.03). These results differ very slightly from the results above (to two decimal places). In fact, there is no advantage to using logistic regression in this example where only two binary predictors are involved. This will not be the case when we consider age as a continuous measure (below). As noted in the lab handout: We don't expect an exact correspondence between the frequency table and logistic regression results when there are multiple risk factors involved. (This is related to the fact that the logistic model represents the effect of risk factors on the outcome as additive on the log odds scale.) Also, a complete analysis of the relationship between CHD, smoking and behavior type would involve assessing the effect of smoking using the quantitative measure `ncig`.

**Question 3.** *What do the above results allow you to conclude about the effect of age on the association between `arcus` and CHD risk? About the effect of `arcus` on the association between age and CHD risk?*

```
. cs chd69 arcus, by(dage) or
      dage |      OR      [95% Conf. Interval]      M-H Weight
-----+-----
      39-49 |  1.911643   1.348286  2.710533   20.34964 (Cornfield)
      50-59 |  1.0575     .7086192  1.578275   23.00885 (Cornfield)
-----+-----
      Crude |  1.63528     1.257732  2.126197
M-H combined |  1.458379     1.118748  1.901115
```

```
-----
Test of homogeneity (M-H)      chi2(1) =    4.742  Pr>chi2 = 0.0294

      Test that combined OR = 1:
              Mantel-Haenszel chi2(1) =    8.05
                          Pr>chi2 =    0.0046
```

. cs chd69 dage, by(arcus) or

arcus senilis	OR	[95% Conf. Interval]		M-H Weight	
absent	2.4431	1.745663	3.419325	18.83853	(Cornfield)
present	1.351497	.8957664	2.03919	18.99575	(Cornfield)
Crude	2.04938	1.58146	2.655782		
M-H combined	1.89503	1.457833	2.463341		

```
-----
Test of homogeneity (M-H)      chi2(1) =    4.747  Pr>chi2 = 0.0294

      Test that combined OR = 1:
              Mantel-Haenszel chi2(1) =   23.99
                          Pr>chi2 =    0.0000
```

The results presented above display the association between CHD and arcus, stratified by age, and CHD and age, stratified by arcus. Comparison of the stratum-specific odds ratio estimates and the (identical) tests of homogeneity reveal that a significant interaction between the two predictors is operating. This implies that any further conclusions about the association between either and CHD risk must consider the presence of the other to be valid. CHD risk appears to be elevated among younger participants with arcus relative to their older counterparts.

**Question 4.** Compare these results with those obtained from the last two *cs* commands and comment on any correspondences. What do the results imply about the presence of interaction?

```
. logistic chd69 arcus dage arcdage
Logistic regression                               Number of obs   =    3152
                                                  LR chi2(3)      =    40.33
                                                  Prob > chi2     =    0.0000
Log likelihood = -865.43251                    Pseudo R2      =    0.0228
-----
```

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
arcus	1.911643	.3419233	3.62	0.000	1.346349	2.714286
dage	2.4431	.4205154	5.19	0.000	1.743529	3.423365
arcdage	.5531892	.1505928	-2.17	0.030	.3244547	.9431773

```
-----
```

The odds ratios for *arcus* and *dage* correspond. The fact that the 95% confidence interval for the product variable *arcdage* excludes 1, (and the associated P-value of 0.03 is smaller than 0.05) corresponds to the result of the test for homogeneity presented above, indicating a significant interaction between *arcus* and this binary representation of age in influencing CHD risk. As demonstrated in lab, identifying the component odds ratios from the *logistic* output involves a bit more work.

**Question 5.** *Interpret each of the above odds ratios using the definition of the logistic model given above*

Using the `lincom` procedure after fitting a logistic regression model allows estimation and inference about combinations of regression coefficients and particular values of predictors using the estimated regression equation. For this example, this has the form:

$$\log(\text{odds of CHD}) = \beta_0 + \beta_1 \text{arcus} + \beta_2 \text{dage} + \beta_3 \text{arcdage}$$

As discussed in lecture, the component log odds ratios for the interaction between arcus and age can be obtained by differences between appropriate pairs of log-odds. For example, the log odds ratio comparing CHD in individuals with arcus to those without can be obtained by taking the difference in the log odds in these two groups. Because the interaction with age is operating (as represented by the presence of the product variable `arcdage`), we must specify an age group to perform this calculation. For the younger age group, `dage = 0` and `arcdage = 0`, so the log odds of CHD for individuals with arcus in this group is given by  $\beta_0 + \beta_1$ . The corresponding log odds for individuals in this group without arcus is then  $\beta_0$ . The difference between these log odds gives the log odds ratio associated with arcus in the younger individuals, i.e.  $\beta_1$ . The odds ratio and confidence interval for this are given directly in the logistic output (if we do not request coefficients to be printed). A similar calculation shows that the corresponding log odds ratio for arcus among older individuals is  $\beta_1 + \beta_3$ . Although we can calculate this from the coefficients of the logistic model, the standard error and 95% confidence interval are not provided. This is where `lincom` comes in. The command

```
lincom arcus + arcdage
```

requests that the coefficients for these two variables be summed and an associated 95% confidence interval estimated. Stata exponentiates `lincom` results when the preceding model is fitted by logistic regression. Note that even though the output is labeled "Odds Ratio", `lincom` can also be used to estimate odds. For example, the odds for CHD among older individuals with arcus is given by

```
lincom arcus + dage + arcdage
```

The odds ratios and 95% confidence intervals produced using `lincom` in this lab correspond very closely to the results from the frequency table approach applied earlier. The latter approach would clearly suffice for this particular analysis. However, if we want to consider age as a continuous variable, the logistic model would clearly be preferred.

The tables used to aid in identifying the requested odds ratios are given below.

arcus	dage	_cons $\beta_0$	arcus $\beta_1$	dage $\beta_2$	arcuage $\beta_3$
Yes	Yes	1	1	1	1
Yes	No	1	1	0	0
	Difference	0	0	1	1

arcus	dage	_cons	arcus	dage	arcuage
Yes	No	1	1	0	0
No	No	1	0	0	0
	Difference	0	1	1	0

arcus	dage	_cons	arcus	dage	arcuage
No	Yes	1	0	1	0
No	No	1	0	0	0
	Difference	0	0	1	0

**Question 6.** *What do the results imply about the presence of an arcus-age interaction as measured by the coefficient for the product variable **arcuage** and its 95% confidence interval?*

The results imply that a significant interaction is operating. The P-value for the Wald test associated with the product term corresponds almost exactly to the result of the test for homogeneity obtained above.

**Question 7.** *Compute the above odds ratio for a 40 year old. Given what we know about the age range of subjects in the WCGS, does it make any sense to use the above model to estimate this odds ratio for an individual of age 20?*

We can compute the odds ratio for a 40-year old individual by using the model coefficients (stored in internal variables after the last model fitted), or by using `lincom` again. The odds ratio arcus among older individuals is given by

```
lincom arcus + arcuage*40
```

Recall that the age range of participants in this study is 35-60. We would not expect regression models to apply to ages well outside this range, especially considering the disease outcome under consideration.

**Question 8.** *What do the plotted lines reveal about the relationship between CHD risk and age in the two arcus groups?*

Although both groups exhibit increased CHD risk with increasing age, individuals with arcus appear to be at greater risk overall except among individuals over age 55. (This observation is consistent with the results obtained above for the categorical representation of age.) The slope of increase in risk with age is greater among individuals without arcus.

**Question 9.** *What do you conclude from the graph about the adequacy of the linear representation of age in the two groups?*

Since the smoothed estimates follow the observed data quite closely, differences with the corresponding linear estimates reveal possible non-linearity in the relationship between the log-odds of CHD risk and age in the two groups. The lines appear to represent the data quite well with the exception of the younger age range (35-45), where some increased risk is evident. Without further investigation of this phenomenon (possibly including formal comparison of a model allowing the log odds to increase with younger ages with the linear alternative), we do not have enough information to conclude that the linear representation is adequate. However, the linear models appear to capture the differences between the two arcus groups well and in this sense provide a reasonable summary of the interaction effect.