

Biostatistics 208, Lab #8 Discussion 02/25/10

Question 1. What do you conclude about the association between chd69 and arcus based on the odds ratio and its 95% confidence interval? How do these conclusions correspond to the result of the chi-squared test provided?

```
. cs chd69 arcus, or
```

	arcus senilis		
	Exposed	Unexposed	Total
Cases	102	153	255
Noncases	839	2058	2897
Total	941	2211	3152
Risk	.1083953	.0691995	.080901
	Point estimate	[95% Conf. Interval]	
Risk difference	.0391959	.0166915	.0617003
Risk ratio	1.566419	1.233865	1.988603
Attr. frac. ex.	.3616011	.1895387	.4971343
Attr. frac. pop	.1446404		
Odds ratio	1.63528	1.257732	2.126197 (Cornfield)
	chi2(1) = 13.64 Pr>chi2 = 0.0002		

The odds ratio for CHD comparing subjects with arcus to subjects without arcus is 1.64 with 95% confidence interval from 1.26 to 2.13. The confidence interval excludes 1, which is the odds ratio if there is no association. Hence, we can conclude there is a positive association between arcus and CHD. We also see that the estimated association is probably not extremely strong (as evidenced by the upper bound of the 95% CI, 2.1). The CI agrees with the chi-squared test results: the test rejects the null hypothesis of no association between arcus and CHD at the 5% significance level.

A couple of points: first, the odds ratio agrees with the chi-squared test but adds much more information. An analysis limited to the chi-squared test will allow us to reject the hypothesis of no association, but it omits 3 pieces of important information: the direction of the association, the magnitude of the association and the plausible magnitude of the association. The odds ratio and the confidence interval give this information. Second, in general, there is a close correspondence between hypothesis tests and confidence intervals in statistics. If a 95% confidence interval excludes a particular value (e.g., 1) then this is equivalent to rejecting a 0.05 level two-sided test of the null hypothesis that the parameter equals 1. We saw an example in class where the Fisher exact test and the CI didn't agree (the CI excluded 1 but the exact test had $p=0.06$). Different tests may use different approximations and will sometimes disagree slightly.

Question 2. Does fitting the logistic regression model add anything to our understanding of the association between these two variables (considering the results from the prior analysis based on the cs command)?

```

Logistic regression
Log likelihood = -879.10783
Number of obs   =    3152
LR chi2(1)      =    12.98
Prob > chi2     =    0.0003
Pseudo R2      =    0.0073
-----+-----
      chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      arcus |    1.63528   .2195035    3.66   0.000     1.257     2.127399
-----+-----

```

Not really. We get the odds ratio, confidence interval and two tests (Wald and Likelihood Ratio). They correspond closely to the results we got from the `epitab` analysis above. In general, a logistic regression model of the association between a binary outcome and a single binary predictor will agree with the corresponding two-by-two table analysis based on a chi-squared test. Recall that the same agreement holds between linear regression involving a single binary predictor (and continuous outcome) and the corresponding the t-test for comparing outcome averages in the two groups.

As discussed in lecture, the frequency table approach breaks down a bit when we consider a continuous variable like age.

Question 3. What do the proportions of participants with CHD reveal about the association between CHD risk and age?

CHD event	35-40	41-45	46-50	51-55	56-60	Total
no	512 94.29	1,036 94.96	680 90.67	463 87.69	206 85.12	2,897 91.85
yes	31 5.71	55 5.04	70 9.33	65 12.31	36 14.88	257 8.15
Total	543 100.00	1,091 100.00	750 100.00	528 100.00	242 100.00	3,154 100.00

Overall, we see the proportion developing CHD seems to increase with age. However, the proportion is similar between the 35-40 year-old age category and the 41-45 year-old age category. Hence, there appears to be an association between age and CHD. The actual form of this relationship is unclear, partially due to our arbitrary selection of five-year age groups.

Question 4. What do the results from the `tabodds` analyses allow you to conclude about the dependence of CHD risk on age?

```
tabodds chd69 agec, or
```

agec	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
35-40	1.000000
41-45	0.876822	0.32	0.5692	0.557454	1.379156
46-50	1.700190	5.74	0.0166	1.095789	2.637958
51-55	2.318679	14.28	0.0002	1.479779	3.633160
56-60	2.886314	18.00	0.0000	1.728069	4.820876

Test of homogeneity (equal odds): chi2(4) = 46.64
Pr>chi2 = 0.0000
Score test for trend of odds: chi2(1) = 40.76
Pr>chi2 = 0.0000

We see that the odds ratio generally increases with the age categories with the exception of the category of age 41-45. We see that there is a trend between the odds of CHD and the age categories and this is highly significant by the trend test.

Question 5. What does the resulting plot reveal about the relationship between (the log odds of) CHD risk and age?

This plot is a non-parametric estimate of the shape of the log odds and age. We are trying to assess whether or not a linear relationship is reasonable. It appears reasonable from about 42 on, but it appears flat to decreasing from 39ish to 42ish.

Question 6: What can you conclude about the adequacy of the linear logistic model for the relationship between CHD risk and age? Do you have an explanation for the apparently increased CHD risk for the youngest (aged 39) participants (relative to the other "younger" age groups)?

We get a plot that overlaps the log odds as modeled by logistic regression and estimated non-parametrically by the LOWESS smoother. We see pretty close agreement except at the low end of the age range. At those ages, logistic regression predicts a risk (measured on the log odds scale) that is lower than is observed in the data. Perhaps the very young subjects joined the cohort because they were at higher risk due to family history, blood pressure so some other factor we have not taken account of.

The violation is slightly troubling. If we model age as linear, we get a nice simple interpretation. However, it may not be the most accurate model for younger ages. If a major goal of the study is to describe the relationship between disease risk and age, lack of fit would be a primary concern; in that case the cut-point approach might or another nonlinear representation of age might be preferred to the linear model. However, if age is being considered primarily as a possible confounder, then assuming linearity may be sufficient. We could verify this with further analyses comparing models controlling for age as linear with others including nonlinear representations. If the effect of these two approaches on coefficients/odds ratios for other variables of interest is the same, the linear representation would suffice. Since there are more than 3000 observations in the WCGS data, we can fit a more complex model for age easily without worrying about too many predictors.