Biost 536: Categorical Data Analysis in Epidemiology

Emerson, Fall 2013

Homework #2

October 10, 2013

Written problems: To be submitted as an email attachment in by 5pm on Thursday, October 17, 2013. See the instructions for peer grading of the homework that are posted on the web pages.

> *On this (as all homeworks) unedited Stata output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*
>
> *Keys to past homeworks from quarters that I taught Biost 517 (e.g. HW #8) or Biost 518 (e.g., HW #3) might be consulted for the presentation of inferential results.*

The following problems make use of a dataset exploring the prognostic value of certain biomarkers of inflammation on all cause mortality. The documentation file inflamm.doc and the data file inflamm.txt can be found on the class web pages.

In all problems, we are interested in any associations between estrogen use and mortality from cardiovascular disease (CVD) within four years of enrolment in the study. Note that no subject was censored prior to four years of follow-up, however some subjects were deemed to die from non CVD causes. For the purposes of this homework, we will treat the patients who die of other causes as if they would definitely not died of CVD within 4 years. Hence, you can create a binary variable indicating CVD death within 4 years. The following Stata code will create this variable:

```
g cvddeath4 = 0
replace cvddeath4 = 1 if ttodth <= 4*365.25 & cvddth==1
```

All references to “CVD mortality” mean CVD death within 4 years.

Some subjects are missing data for estrogen, but for the purposes of this homework we will presume that such data is missing completely at random (MCAR).

Note that only women are expected to have used estrogen therapy, and thus all analyses should be restricted to women.

Problems 1-3 each ask the same questions, but ask for different measures of association. Where such would be appropriate, it is permissible to give answers to parts of problems 2 and 3 as “same answer as in problem 1”.

1. Suppose we are interested in measuring any association between estrogen use at any time prior to study enrollment (*estrogen==1)* and CVD death within 4 years using the **risk difference (RD)**.

    a. Provide complete statistical inference regarding such an association. (Include point estimates, confidence intervals, and a p value, along with a full interpretation of those quantities.)

The CVD mortality rate is 4.925% among subjects with no estrogen use, compared to 1.16% among those with. The risk difference of 3.76% (95% CI: 1.45 - 6.07) was statistically significant (p = 0.0014).

b. Is there evidence in the dataset that any such effect is modified by a history of prior CVD (as measured by variable *prevdis)*? Provide results of a statistical analysis in support of your answer.

Subjects with estrogen use had lower risk of CVD mortality on average than subjects without estrogen use, regardless of prior history of CVD. The risk difference was 1.87% among subjects with no prior history of CVD, and 8.42% among subjects with prior history. This suggests some evidence of effect modification by CVD history; however, the interaction was not statistically significant (p=0.0946).

c. Suppose we just want to ignore any such effect modification. Is there evidence in the dataset that any estrogen-CVD mortality association is confounded by a history of prior CVD? Provide results of a statistical analysis in support of your answer.

		Previous Diagnosis of CVD	
		No	Yes
Estrogen Use	No	3533 (91.9%)	1117 (97.2%)
	Yes	312 (8.1%)	32 (2.8%)
CVD Death in 4 yrs	No	3748 (97.3%)	1019 (88.7%)
	Yes	103 (2.7%)	130 (11.3%)

The above table suggests possible confounding by prior CVD. The proportion of subjects with estrogen use is lower in subjects with previous diagnosis of CVD (2.8% vs. 8.1%), while the incidence of death within 4 years from CVD is higher (11.3% vs. 2.7%). Additionally, the adjusted model (part d, below) gives a different estimate of the risk difference of CVD mortality between estrogen users and non-estrogen users than the unadjusted model (3.76% before adjusting vs. 2.51% after adjusting). This is an indication of confounding, and suggests that we should adjust for previous CVD in the final model.

d. Provide complete statistical inference regarding an association between estrogen and CVD mortality after adjustment for a prior history of CVD.

Estrogen users had lower risk of CVD mortality on average compared to non-users with the same CVD history. The adjusted risk difference of 2.51% was statistically significant (p=0.0312; 95% CI: 0.23 – 4.79).

e. Is there evidence in the dataset that the prior disease adjusted analysis of an association between estrogen-CVD mortality is further confounded by age? Provide results of a statistical analysis in support of your answer.

The model adjusting for both age and previous CVD has a smaller effect estimate for estrogen use compared to the model adjusting only for previous CVD (RD = 1.60% in the fully adjusted model vs. 2.51% in the model only adjusting for previous CVD). This is an indication of further confounding of the effect of estrogen by age, and suggests that age should be adjusted for in the final model.

f. Provide complete statistical inference regarding an association between estrogen and CVD mortality after adjustment for age and any prior history of CVD.

Estrogen users had lower risk of CVD mortality on average compared to non-users with the same age and CVD history. The adjusted risk difference of 1.60% was not statistically significant (p=0.168; 95% CI: -0.68 to 3.88).

2. Answer all parts of problem 1 using the **odds ratio (OR)** as the measure of association.

a. The odds of CVD mortality were 0.0518 among subjects without estrogen use, compared to 0.0117 among subjects with estrogen use. The odds ratio of 0.227 (95% CI: 0.084 – 0.614) was statistically significant (p=0.0035)

b. Subjects with estrogen use had lower odds of CVD mortality on average than subjects without, regardless of previous CVD history. The odds ratio was 0.333 among subjects with no prior history, compared to 0.247 among subjects with prior history. This suggests some evidence of effect modification by CVD history; however, the interaction was not statistically significant (p=0.0799).

c. The adjusted model (part d, below) gives a different estimate of the odds ratio than the unadjusted model (part a). This is to be expected in the case of logistic regression even with a covariate that is unrelated to the predictor of interest (i.e. precision variable). However, since the adjusted estimate of 0.306 is closer to the null than the unadjusted estimate of 0.227 (instead of farther away from the null as might be expected with a precision variable), there is strong evidence that previous CVD is indeed a confounder of the association between estrogen use and mortality.

d. Estrogen users had lower odds of CVD mortality on average compared to non-users with the same CVD history. The adjusted odds ratio of 0.306 was statistically significant (p=0.021; 95% CI: 0.113 – 0.834).

e. The fully adjusted model (part f, below) has an estimated odds ratio closer to the null than the model without adjustment for age (part d, above). Following the reasoning in part c, this is evidence that age further confounds the association between estrogen use and mortality.

f. Estrogen users had lower odds of CVD mortality on average, compared to non-users with the same age and CVD history. The adjusted odds ratio of 0.361 was statistically significant (p=0.048; 95% CI: 0.132 – 0.988).

3. Answer all parts of problem 1 using the **risk ratio (RR)** as the measure of association. (Note that the Stata `glm` command can be used to effect such analyses.)

a. The CVD mortality rate is 4.925% among subjects with no estrogen use, compared to 1.16% among those with. The risk ratio of 0.236 (95% CI: 0.088 – 0.631) was statistically significant (p = 0.0040).

b. Subjects with estrogen use had lower risk of CVD mortality on average than subjects without, regardless of previous CVD history. The risk ratio was 0.340 among subjects with no prior history, compared to 0.271 among subjects with prior history. This suggests some evidence of effect modification by CVD history; however, the interaction was not statistically significant (p=0.0843).

c. The adjusted model (part d, below) gives a different estimate of the risk ratio than the unadjusted model (part a). This is strong evidence that previous CVD is a confounder of the association between estrogen use and mortality.

d. Estrogen users had lower risk of CVD mortality on average compared to non-users with the same CVD history. The adjusted risk ratio of 0.319 was statistically significant (p=0.023; 95% CI: 0.119 – 0.851).

e. The fully adjusted model (part f, below) gives a different estimated risk ratio than the model without adjustment for age (part d, above). This is strong evidence that age further confounds the association between estrogen use and mortality.

f. Estrogen users had lower odds of CVD mortality on average, compared to non-users with the same age and CVD history. The adjusted risk ratio of 0.368 was statistically significant (p=0.046; 95% CI: 0.138 – 0.983).

4. Of the three measures of association used above, how similar were the conclusions? What are the relative advantages and disadvantages of the three?

	Risk Differences (Linear regression)	Odds Ratios (Logistic regression)	Risk Ratios (Log-linear regression)
Unadjusted effect (estrogen use vs. non-use)	3.76% (1.45, 6.07)	0.227 (0.084, 0.614)	0.236 (0.08,0.631)
Effect adjusted for previous CVD	2.51% (0.23, 4.79)	0.306 (0.113, 0.834)	0.319 (0.119, 0.851)
Effect adjusted for previous CVD & age	1.60% (-0.68, 3.88)	0.361(0.132, 0.988)	0.368 (0.138, 0.983)

All numbers in the above table represent point estimates (95% CIs) of the effect of estrogen use on CVD mortality (adjusted for the variables indicated in the row headings).

The risk differences in the first column must be interpreted with relation to the fact that the risk of CVD mortality in the reference group (no estrogen use) is 4.925%. In this range of binomial probabilities, using risk differences is less advantageous in that they are harder to interpret and compare, and may result in estimates outside the possible range.

As seen from the above table, the inferences using odds ratios and risk ratios are very similar, which is to be expected when the binomial probabilities are close to zero as they are here. Also, the 4.925% risk of mortality in the reference group corresponds to odds of 0.0518, which is pretty close. For ease of interpretation and to ensure stable estimates when possibly adjusting for further covariates, I would prefer the risk ratio to the odds ratio.