Biost 536: Categorical Data Analysis in Epidemiology

Emerson, Fall 2013

Homework #3

November 21, 2013

Written problems: To be submitted as an email attachment in by 5pm on Wednesday, November 27, 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.*

All questions relate to the question of whether the nadir PSA level following hormonal treatment for prostate cancer is prognostic of time in remission independently of any information from other commonly used covariates. The data is posted on the class web pages (psa.txt), with documentation in the file psa.doc. Note that the variable *inrem* is text (“yes” or “no”). You will need to tell Stata that this variable should be stored as a “string” rather than as a number. The following code would do the trick:

```
infile ptid nadir pretx ps bss grade age obstime str8 inrem using psa.txt
```

Note that all patients were followed for a minimum of 24 months. In all problems we will be considering the probability (or odds) of a patient surviving relapse-free for 24 months following therapy. You can create a variable indicating relapse within 24 months using the following Stata code:

```
g relap24 = 0
replace relap24 = 1 if obstime <= 24 & inrem==”no”
```

1. Provide suitable descriptive statistics for this dataset as might be presented in Table 1 of a manuscript appearing in the medical literature. (Because the primary question is comparing 24 month relapse free survival across groups defined by nadir PSA, you might consider presenting descriptive statistics in groups according to some dichotomization of nadir PSA levels. Alternatively, you could provide descriptive statistics within groups defined by whether the subjects relapse within 24 months or not.)

	No relapse within 24 mo. (n=28)	Relapse within 24 mo. (n=22)
Pre-treatment PSA (ng/mL)	617.2 [1252.1] (5 NA)	732.4 [1357.3] (2 NA)
Nadir PSA (ng/mL)	4.1 [17.3] (0 NA)	31.9 [52.5] (0 NA)
Perf. Status (score out of 100)	83.9 [9.6] (0 NA)	76.5 [11.8] (2 NA)

Bone scan score:	(0 NA)	(2 NA)
1 (Least disease)	5 [17.9%]	0 [0.0%]
2	9 [32.1%]	4 [18.2%]
3 (Most disease)	14 [50.0%]	16 [72.7%]
Tumor Grade:	(4 NA)	(5 NA)
1 (Least aggressive)	7 [25.0%]	3 [13.6%]
2	8 [28.6%]	7 [31.8%]
3 (Most aggressive)	9 [32.1%]	7 [31.8%]
Age [yrs]	66.7 [5.8] (0 NA)	68.4 [5.7] (0 NA)

2. Perform logistic regression analyses to determine whether the distribution of relapse within 24 months differs across groups defined by nadir PSA level after adjustment for bone scan score and performance status. For each of the following models, provide full statistical inference for your measure of association.

    a. Perform an adjusted logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as a continuous, untransformed variable.

       Between groups with similar bone scan score and performance status, and differing in nadir PSA level by 10 ng/mL, the odds of relapse within 24 months were, on average, 39.0% percent higher in subjects with higher values (95% CI: -10.8 to 116.7). This difference was not statistically significant at the 0.05 level (p = 0.146).

    b. Perform an adjusted logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as a continuous, log transformed variable.

       Between groups with similar bone scan score and performance status, and differing in nadir PSA level by a factor of 2, the odds of relapse within 24 months were, on average, 79.5% percent higher in subjects with higher values (95% CI: 26.0 to 155.7). This difference was statistically significant at the 0.05 level (p = 0.0012).

    c. Perform an adjusted logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as linear splines with knots at 1, 4, and 16 ng/ml.

       The Wald test for significance of nadir PSA level, modeled as linear splines as specified above, indicates a statistically significant effect on odds of relapse within 24 months, after adjustment for bone scan score and performance status (p = 0.0458).

d. For each of the above regression models, provide an interpretation of the intercept.

In models 2(a) and 2(c), the intercept term can be interpreted as the average log-odds of relapse within 24 months among subjects with bone scan score of 1 (least amount of disease), performance status of 0 (worst possible score), and nadir PSA of 0.0 ng/mL.

The interpretation is the same in model 2(b) except that it is among subjects with nadir PSA of 1.0 ng/mL instead of 0.0 ng/mL.

Note: *For all parts of this problem, I adjusted for performance status as a continuous covariate. Examination of the dataset suggests it may be more appropriate as a categorical variable, as it only takes on values between 50 and 100 in increments of ten in this dataset. However, adjusting for it may create problems with the stability of the estimates since the sample size is so small. One would have to consider whether modeling PS linearly is a good enough approximation considering it's not the POI. Also, some transformation of it, or dichotomization, may be appropriate if we are truly worried about nonlinearity affecting the model.*

3. In this longitudinal study, we could instead have considered the "reverse" analyses in which nadir PSA is used as the response and the predictor is the indicator of relapse within 24 months.

a. Perform linear regression analyses to determine whether there is an association between mean nadir PSA level and relapse within 24 months after adjustment for bone scan score and performance status. Make clear the statistical analysis you perform. Provide full statistical inference for your measure of association.

Among subjects with similar bone scan score and performance status, the mean nadir PSA was 23.36 ng/mL higher in subjects who relapsed within 24 months compared to those who didn't (95% CI: -1.31 to 48.03). This difference was not statistically significant (p = 0.063).

b. Perform linear regression analyses to determine whether there is an association between geometric mean nadir PSA level and relapse within 24 months after adjustment for bone scan score and performance status. Make clear the statistical analysis you perform. Provide full statistical inference for your measure of association. (Recall that inference on the geometric mean is obtained by performing linear regression on log transformed response variables.)

Among subjects with similar bone scan score and performance status, the geometric mean nadir PSA was 13.73 times higher in subjects who relapsed within 24 months compared to those who didn't (95% CI: 4.60 to 40.97). This difference was statistically significant (p < 0.001).

4. Consider the analyses performed in problems 2 and 3 above.

a. What are the relative merits of the five analyses. Which might you prefer *a priori*? Why?

A priori, given some knowledge of the typical behavior of variables measured as blood concentration, we would prefer an analysis in which PSA was log-transformed. A linear spline could work as well if we examined the distribution of PSA and chose sensible knots, but the coefficients are harder to interpret.

We might also prefer the analyses in question 3, since there are potentially fewer confounders, as they would have to be causally related to the outcome (nadir PSA).

b. All of these analyses suffer from a serious definitional problem inherent in this study. Can you deduce this problem? (Hint: There is no analysis that you can do to address this problem. It is a problem with the study design.)

One problem is that since nadir PSA is defined as the minimum value observed after treatment, it is necessarily dependent on how many observations are taken and for how long. Since how many observations are available is plausibly associated with whether the patient was in remission for 24 months, this could introduce bias and confounding to the results. This could potentially addressed by using some other summary measure of post-treatment PSA. However, there is still the additional problem that the question of interest seems to be a prognostic one, and post-treatment PSA is by definition not available for some amount of observation time directly after treatment, during which the patient could potentially relapse. It's hard to see the usefulness of nadir PSA as a prognostic variable since it depends on values potentially as far out as 24 months after treatment.