**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 Relapsed with 24 months N= 28N(%) | Relapsed within 24 months N=22N(%) |
| Nadir PSA\* | 4.11 (17.3) | 31.9 (52.5) |
| Pre-TX PSA\* | 617.2 (1252.1) | 732.4 (1357.4) |
| PS\* | 83.9 (9.6) | 76.5 (11.8) |
| BSS |  |  |
| 1\* | 5 (17.9) | 0 (0) |
| 2 | 9 (32.1) | 4 (20.0) |
| 3 | 14 (50.0) | 16 (80.0) |
| Tumor Grade |  |  |
| 1 | 7 (25.0) | 3 (13.64) |
| 2 | 8 (28.6) | 7 (31.8) |
| 3 | 9 (32.1) | 7 (31.8) |
| Age\* | 66.7 (5.8) | 68.4 (5.7) |

\*denotes a mean and standard deviation

\*\*because we have no cases with a BSS of 1, I re-categorized BSS as BSS </=2 and BSS=3 for the models.

1. 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.
	1. 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.

**After controlling for bone scan score (BSS) and performance status, those with the same performance status and BSS have, on average an odds ratio for relapsing within 24 months for every one unit increase in nadir PSA is 1.04 which would not be unusual if the true odds ratio was between 0.95-1.14.**

* 1. 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.

**After controlling for bone scan score (BSS) and performance status, those with the same performance status and BSS have, on average, an odds ratio for relapsing within 24 months for every one unit increase in log transformed nadir PSA is 2.40 which is not unusual if the true odds ratio was between 1.29 and 4.46.**

* 1. 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.

**From the linear spline model we can do a test of non-linearity which gives us a p value of 0.0024. This demonstrates that the data is non-linear. From the graph below we can see the different slopes for each of the knots, and that we clearly do not have a linear relationship between Nadir PSA and the probability of relapse in 24 months.**

**Interpretation of the point estimates:**

**After controlling for bone scan score (BSS) and performance status, those with the same performance status and BSS, among those with a nadir PSA of </=1 for every one unit increase in nadir PSA the odds ratio 19.93 (95%CI 1.38-287.6) for relapse within 24 months.**

**After controlling for BSS and performance status, among those with a nadir PSA between 1-4, on average for every unit increase in nadir PSA the OR for relapse is 1.86 (95%CI 0.89-3.77).**

**After controlling for BSS and performance status, among those with a nadir PSA between 4-16, on average for every unit increase in nadir PSA the OR for relapse is 0.99 (95%CI 0.97-1.02).**

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

**For 2a- The interpretation of the intercept is the Odds Ratio for relapse among men with PS of zero, BSS of </=2 which is OR 17.97 for a one unit increase in nadir PSA. The lowest value of PS in the dataset is 50, so this is extrapolating well outside of our data.**

**For 2b- The interpretation of the intercept is the Odds Ratio for relapse among men with PS of zero, BSS of </=2 which is OR 18.56 for a one unit increase in log nadir PSA. The lowest value of PS in the dataset is 50, so this is extrapolating well outside of our data.**

**2c –with splines, the intercept has no real interpretation.**

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.

* 1. 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.

**Controlling for PS and BSS, men with the same PS and BSS score, the difference in the mean nadir PSA between those who did and did not relapse within 24 months is 22.8 which is not unusual if the true difference is between 0.11 and 45.5.**

* 1. 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.)

**Controlling for PS and BSS, men with the same PS and BSS score, the difference in the geometric mean of nadir PSA between those who did and did not relapse within 24 months is 11.04 which is not unusual if the true difference is between 3.69 and 86.5.**

1. Consider the analyses performed in problems 2 and 3 above.
	1. What are the relative merits of the five analyses. Which might you prefer *a priori*? Why?

2a –The benefit of this analysis is that it is easy to interpret because it is the OR for the outcome based on a one unit increase in your predictor. This drawbacks are that I would not expect nadir PSA to have a linear relationship with relapse, therefore it difficult to fit a linear model using nadir PSA.

2b –A priori I would pick this analysis because I know PSA has increases at an exponential rate therefore modeling the log transformed PSA will fit a linear model better. It also is easier to interpret an OR for disease given exposure.

2c- This model is great if I think that the risk of relapse given different levels of PSA has a completely non-linear relationship and does not fit other standard curves because it oscillates. This model allows us to fit different curves at different levels of PSA. The weakness of splines is that you are no longer borrowing information across PSA, so you can have limited information to inform your model in specific splines.

3a – By flipping my model, given my outcome of relapse, is there a difference in my exposure, I am able to compare the difference in mean PSA between those who did and did not relapse. This analysis is great for interpretation because it is easy to say that one group has a higher mean PSA than the other.

3b. Again, this model gives us the flip side of saying is there is a difference in geometric mean PSA between those who did and did not relapse. This is easy to interpret and the log transformation of nadir PSA allows for a better fit in the linear model.

* 1. 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.)

We don’t know the timing of the lowest PSA value in regards to the timing of relapse. We can’t ensure that the lowest PSA recorded preceded our outcome of cancer relapse. If a person was in the study for 5 years their lowest PSA values could have been recorded after the 24 month cut off for relapse.