**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.

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.)

|  |  |  |
| --- | --- | --- |
| **Table 1. Characteristics of Patients with Prostate Cancer by 24-Month Remission Status** | | |
|  | In Remission (n=28) | Not in Remission(n=22) |
| **Age, mean (SD)** | 66.7 (5.84) | 68.4 (5.68) |
| **Time observed in remission (months), mean (SD)** | 42.1 (12.05) | 11.1 (6.40) |
| **Tumor grade (1= least aggressive, 3= most), n (%)** |  |  |
| **Grade 1** | 7 (25) | 3 (13.64) |
| **Grade 2** | 8 (28.57) | 7 (31.82) |
| **Grade 3** | 9 (32.14) | 7 (31.82) |
| **Unknown** | 4 (14.29) | 5 (22.73) |
| **PSA value prior to therapy (ng/ml), mean (SD)a** | 617.2 (1252.1) | 732.4 (1357.3) |
| **Lowest PSA value post therapy (ng/ml), mean (SD)** | 4.1 (17.3) | 31.9 (52.5) |
| **Bone scan score (1= least disease, 3= most), n (%)** |  |  |
| **Score 1** | 5 (17.86) | 0 |
| **Score 2** | 9 (32.14) | 4 (18.18) |
| **Score 3** | 14 (50) | 16 (72.73) |
| **Unknown** | 0 | 2 (9.09) |
| **Performance status (0=worst, 100= best), mean (SD)** | 83.9 (9.6) | 76.5 (11.8) |

a. Pre-treatment PSA values available for 23 of the patient in remission, 20 of the patients not in remission.

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 adjusting for performance status and bone scan score, the odds ratio for relapse among patients with a nadir PSA level of 0 ng/mL is estimated to be 1.03 (95% CI 0.94, 1.13). No one in this dataset had a nadir PSA of 0 ng/mL, thus we should be reluctant to extrapolate these results to a group of patients with such a nadir PSA value.

* 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 adjusting for performance status and bone scan score, the odds ratio for relapse among patients with a log nadir PSA level of 0 ng/mL is estimated to be 3.06 (95% CI 1.27, 4.40).

* 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.

After adjusting for performance status and bone scan score, the odds ratio for relapse among patients with a nadir PSA level of 0 ng/mL is estimated to be 0.51 (95% CI 0.00, 2081). Again, no one in this dataset had a nadir PSA of 0 ng/mL, thus we should be reluctant to extrapolate these results to a group of patients with such a nadir PSA value.

1. 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.

Using multivariable robust linear regression, after adjusting for performance status and bone scan score, the average nadir PSA value is estimated to differ between those who do and do not relapse at 24 months by 23.5 ng/mL, with higher values among those who relapse. This result is significantly different from 0 (p=0.046), with a 95% CI suggesting that such observed results would not be unusual if the true difference in mean nadir PSA value between remission groups were anywhere between 0.48 and 46.56 ng/mL. We thus reject the null hypothesis that mean nadir PSA does not differ across remission groups, in favor of a hypothesis that mean nadir PSA tends to be higher among those who relapse.

* 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.)

Using multivariable robust linear regression, after adjusting for performance status and bone scan score, the geometric mean nadir PSA value is estimated to differ between those who do and do not relapse at 24 months by 13.66 ng/mL, with higher values among those who relapse. This result is significantly different from 0 (p<0.001), with a 95% CI suggesting that such observed results would not be unusual if the true difference in geometric mean nadir PSA value between remission groups were anywhere between 4.13 and 45.16 ng/mL. We thus reject the null hypothesis that geometric mean nadir PSA does not differ across remission groups, in favor of a hypothesis that geometric mean nadir PSA tends to be higher among those who relapse.

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?

Five analyses (all adjusted for performance status and bone scan score): 1) logistic regression with relapse in 24 months as the outcome, nadir PSA as the predictor modeled as a continuous, untransformed variable; 2) logistic regression with relapse in 24 months as the outcome, nadir PSA as the predictor modeled as a continuous, log-transformed variable; 3) logistic regression with relapse in 24 months as the outcome, nadir PSA modeled as linear splines; 4) linear regression with nadir PSA as the outcome and relapse in 24 months as the predictor of interest; and 5) linear regression with log nadir PSA as the outcome and relapse in 24 months the predictor of interest.

Analysis 1 - if the truth is a straight-line relationship, this model would be an accurate and efficient way to detect associations.

Analysis 2 – log transformation of nadir PSA may allow a more normal distribution of PSA values.

Analysis 3 – the utilization of linear splines may enhance our ability to detect a non-linear trend, if present.

For 4 & 5, tests using linear regression may be more robust than logistic regression with a sample size < 50.

Analysis 4 – if the truth is a straight-line relationship, this model would be an accurate and efficient way to detect associations.

Analysis 5 – log transformation of nadir PSA may allow a more normal distribution of PSA values and assist in detection of trends.

The analyses using nadir PSA do not make conceptual sense to me. How can 24-month remission status serve as a predictor when it is likely to follow the outcome of interest (nadir PSA)?

In terms of the models using remission status at 24 months as the outcome of interest, I prefer the model using the log transformed value for nadir PSA, as this smoothes the distribution of this variable. The model with the linear splines I find difficult to interpret and am unsure about the rationale behind choosing those particular values as the knots.

There are patients in this study with varying times to follow-up and some patients who relapse before and some after 24 months. There will be patients who reach 24 months “in remission” but who relapse months later. The events that will happen sometime in the future beyond the minimum follow-up time are thus censored. Time to event analysis using the censored time to remission and the log nadir PSA as the predictor of interest might provide a better assessment of these data.

* 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.)

How was “in remission” defined? Some of these patients had high bone scan scores (3) and were documented as in remission. If you don’t have a consistent way to define your outcome it would seem hard to offer much in the way of prediction.